Информатика и её применения

Том 11 Выпуск 2 Год 2017

СОДЕРЖАНИЕ

Технический редактор *Л. Кокушкина* Художественный редактор *М. Седакова*
Сдано в набор 14.06.17. Подписано в печать 28.06.17. Формат 60 x 84 / 8
Бумага офсетная. Печать цифровая. Усл.-печ. л. 16,5. Уч.-изд. л. 15. Тираж 100 экз.
Заказ № 989
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
Отпечатано в НИПКЦ «Восход-А» с готовых файлов
Москва 109052, ул. Смирновская, д. 25, стр. 3

INFORMATICS AND APPLICATIONS 2017. Vol. 11. Iss. 2. P. 2–15

DYNAMIC MODELS OF SYSTEMIC RISK AND CONTAGION

Kh. El Bitar[1], Yu. Kabanov[2], and R. Mokbel[3]

Abstract: Modern financial systems are complicated networks of interconnected financial institutions and default of any of them may have serious consequences for others. The recent crises have shown that complexity and interconnectedness are the major factors of systemic risk, which became the subject of intensive studies usually concentrated on static models. The authors develop a dynamic model based on the so-called structural approach, where defaults are triggered by the exit of some stochastic process from a domain. In the case considered, this is a process defined by the evolution of bank's portfolios values. At the exit time, a bank defaults and a cascade of defaults starts. The authors believe that the distribution of the exit time and the subsequent losses may serve as indicators allowing regulators to monitor the state of the system and take corrective actions in order to avoid contagion in a financial system. The authors model the development of a financial system as a random graph using the preferable attachment algorithm and provide results of numerical experiments on simulated data.

Keywords: systemic risk; contagion; scale free network; default

DOI: 10.14357/19922264170201

1 Introduction

In interbank market, systemic risk is a risk arising from a complexity of financial network and threatening the entire system by a potential financial crisis, resulting in high economical and social costs. Controlling financial stability and assessing systemic risk is a major concern of central banks and financial regulators. The rapid growth of financial innovation and integration as well as a complicated network of claims and obligations linking the balance sheets of banks raise the challenge of the systemic risk analysis. This kind of risk is highly dynamic, slowly building up during periods of stability and rapidly rising during crises and spreading through the network. On the other hand, the interconnections of banks have a positive side since they enhance liquidity and increase the risk sharing among the financial institutions.

One of the aim of theoretical studies is to provide regulators comprehensive indicators allowing to monitor the risk of contagion, understood as a cascade of defaults that may lead to a serious consequences and even to the collapse of the whole economy. To the moment, there is a substantial progress in understanding various phenomena causing the contagion on the basis of modeling using random graphs. Network models became the mainstream of current researches in the field (see the recent book by Hurd [1] and the references wherein).

Recent crisis revealed that the systemic risk might take various forms. One of them is an interbank contagion process when, due to the interconnectedness of banks through interbank loans, the default of one bank leads to losses and subsequent chain of defaults of other banks. This kind of risk is usually combined with a risk related to a correlation between banks' portfolios which consists in the phenomena that a common shock, due to common asset holdings, affects many banks at once.

De Bandt *et al.* provided a categorization of systemic risks, distinguishing between those understood in a broad and in a narrow sense [2]: contagion effects pose a systemic risk in the narrow sense while in the broad sense, it is a common shock that affects many nodes and once. A similar idea is developed by Gai and Kapadia [3], who model two channels of contagion in financial system that can trigger further rounds of defaults: contagion due to the direct interbank claims and obligations and contagion due to common shocks on the asset side of the balance sheet, especially when the market for key financial system assets is illiquid.

Deposits also could affect the financial system stability: a large sudden withdrawal of funds by depositors in panic could lead to a collapse of the system. However, in the present paper, this is not considered as one of the major sources of system risk. in fact, its impact can be minimized and controlled by the central

[1]Laboratoire de Mathématiques, Université de Franche-Comté, 16 Route de Gray, 25030 Besançon, CEDEX, France, khalilbitar_aw@hotmail.com

[2]Laboratoire de Mathématiques, Université de Franche-Comté, 16 Route de Gray, 25030 Besançon, CEDEX, France; Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; National Research University "MPEI," 14 Krasnokazarmennaya Str., Moscow 111250, Russian Federation, Youri.Kabanov@univ-fcomte.fr

[3]Laboratoire de Mathématiques, Université de Franche-Comté, 16 Route de Gray, 25030 Besançon, CEDEX, France, ritamokbel@hotmail.com

bank intervention imposing an appropriate withdrawal limit.

A large part of the literature has focused on the analysis of the contagion effect due to the interbank market while only few authors studied the impact of the correlated defaults which is of great importance, the magnitude of correlation between the banks balance sheets, the amount of external investments, and the appropriate assessment of the risk embedded in these external assets. Acharya and Yorulmazer proved that banks are motivated to increase the correlation between their investments amplifying by such actions the risk of a common shock [4]. In their analysis, Elsinger *et al.* combine the two major sources of systemic risk and find that the correlation in investments is far more important than financial linkages [5, 6].

One can also consider a subordinate source of risk due to the fire sale of external assets of defaulting banks which will lead to other banks default because of the price depreciation. This is why some banks have an interest to bailout other peers in order to minimize the default cost of the system and to prevent fire sale and the writing down of their own external assets.

While the interbank risk is concerned, Gai and Kapadia show that the risk of systemic crises is reduced with increasing connectivity while at the same time, the amplitude of the systemic crises is increasing [7]. Higher connectivity simply creates more channels of contact through which default could spread, increasing the potential or probability for contagion. However, in the financial system setup, greater connectivity allows counterparties risk sharing as exposures are distributed over a wider set of banks, especially in periods of stability. In times of crisis, however, the same interconnections can amplify shocks that spread through the system.

Allen and Gale demonstrated that the spread of contagion depends on the network structure of the financial system and strongly interconnected banking systems are less affected by the systemic risk [8]. They also pointed out that the assumption that the agents have complete information on their environment is not realistic. Acharya and Bisin compared over-the-counter (OTC) and centralized clearing markets in a general equilibrium model [9]. They showed that the intransparency of OTC markets is ex-ante inefficient and will lead to underpricing of counterparty risk.

The counterparty risk makes it clear that the network structure of financial system plays an important role when assessing systemic risk.

Empirical analyses of the interbank network structure exist for a number of countries. It shows that the interbank network has a scale free topology. This means that there are a few large banks with many interconnections and many small banks with a few connections. In contrast, other authors argue that the intransparency of real data makes the random network more valid to capture the hidden links. More formally, the terminology "scale free network" means that at least, when the number of nodes increases to infinity, the number k of connections ("in" or "out") attributed to each node decays as $k^{-\gamma}$, $\gamma > 1$.

Georg proposed a dynamic model of cascading banking defaults [10]: at each stage of the cascade, each bank collects all his exposures, pays all his liabilities, adjust the price of its external assets, and, when remains solvent, it optimizes a portfolio of risky and risk-free assets and initiates other interconnections within the banking system.

On the other hand, Gai and Kapadia highlighted that in normal times, developed country banks are robust and minor variations in their default probabilities do not affect lending decisions on the interbank market [7]. But in crises, as illustrated by the sudden failures of Lehman Brothers, contagion may spread rapidly with banks having little time to alter their behavior before they are affected. Thus, the almost static behavior of the system during crisis is best captured by the static model as also applied in the present paper.

It seems that the majority of existing literature deals with "homogeneous" models, like Erdos–Renyi model where the network graph is generated by a matrix whose nondiagonal entries are identically distributed independent Bernoulli random variables (see [3]), or even models where all nodes have the same number of connections [11]. Though such models are convenient for theoretical studies, they look to be too far from the reality and in the present paper, the behavior of the systemic risk indicator is investigated using networks with a structure obtained by a preference attachment algorithm leading to a scale free network.

Under the Basel II accord, improving the quality of default models is the key risk-management priority. Many researchers have studied the loss or impact of the systemic risk once a crisis or shock is in place. However, there is a need to predict and prevent the defaults of banks before it happens. To the date, the major part of research papers concentrates on studies of static or stationary models. In this note, the authors suggest an approach influenced by the structural model of defaultable securities (see [12]). Namely, it is supposed that the cascade of default is triggered in a natural way when the value of a portfolio process of some bank falls below a certain level. Financial market reacts negatively to such an event. Prices of the external assets drop down and contagion propagates not only to interconnected banks but also via correlation. Assuming that the matrix of exposures as well as the vector of the investments into external assets is known, the regulators, having a model for the dynamic of the "reference portfolio," can compute, with moving time horizons, two "alert indi-

cators:" the probability that the default happens during the planning period and the total losses incurred when the default happens. The total losses are the aggregation of the losses due to the external asset price depreciation (correlation) and the losses due to the interbank linkages (contagion). To simplify the calculation, it is assumed that there is a single external risky asset common to all banks in the system and the difference is only in the size of portfolios. A model where each bank has its own portfolio structure can be treated in a similar way. The present approach is rather flexible and can be combined with existing methods of reconstructing of the exposure matrix.

Thus, the main novelty of this approach, in contrast to the majority of existing studies concentrated on static or stationary models, is in developing a dynamic model of financial system before the crisis in combination with a static contagion model for the crisis. The model is described by a graph whose nodes are banks (or other financial institutions). The directed graph structure arises from the matrix of liabilities/exposures. Each bank is characterized by a stylized balance sheet. On the asset side, there are exposures (due to the interbank lending) and liquid assets, risky (stocks) and nonrisky (cash). The liability side is composed by the received interbank loans and the net worth, the quantity, equating both sides of the balance sheet. The dynamic is introduced via random fluctuations of the value of the risky asset. Decreasing of its price means that the net worth is decreasing. The authors suppose that the risky asset is unique for all banks. One may think of this asset as a "benchmark (or reference) portfolio." Taking into account that banks try to mimic behavior of each other ("herding effect"), we believe that this assumption may suit to our highly stylized model but for practical applications, it can be relaxed. Of course, there is a need to introduce dynamics in other parts of the balance sheet but we prefer avoid this in the paper.

The paper contains some numerical experiments. Unfortunately, the liability matrix of a financial system is not publicly available (with rare exceptions). By this reason, the applicability of the model was tested on simulated data. In numerical experiments, a construction of the scale-free network is used with the help of a preferential attachment algorithm (see [13]). The present authors populate the model by balance sheets and compute the alert indicators. The carried out experiments show that the alert indicators can be used as a tool to support regulator's decision to increase the stability of the financial system by withdrawal of the license of the bank having low reliability.

The structure of the article is as follows. In Section 2, the general network approach to contagion is described. Section 3 gives the model description and the definition of the alert indicators. Section 4 is devoted to simulation results.

2 Network Approach

2.1 General principles

The basic ideas are very simple and can be described as follows. The set $G = \{1, \dots, N\}$ stands for the banking system involving N financial institutions described by an $N \times N$ matrix $L = (L^{ij})$ with nonnegative entries vanishing on the diagonal ($L^{ii} = 0$) and a vector $C \in \mathbf{R}^N$ with nonnegative components.

The entry L^{ij} represents the *liability* of the ith bank to the jth bank, i. e., the debts of i to j or, in other words, the total amount of credit provided by j to i. By the reciprocity, for the ith bank, the value L^{ji} is its *exposure* to the bank j. By this reason, in the literature, the model is described quite often by the matrix of the liabilities $X = (X^{ij})$, $X = L'$ where $'$ is used to denote the transpose. Let $B^{ij} = I_{\{L^{ij}>0\}}$. The matrix B (whose entries are zeros and units) defines the directed graph structure on the set of N points in a usual way (as is done in the theory of Markov chains): there is a flesh $i \to j$ if $B^{ij} > 0$, showing that the ith bank is indebted to the jth bank (attention: in some papers, the direction of fleshes can be opposite). With this observation, one can use the standard terminology of the network theory and identify banks with the nodes of the (weighted) oriented graph.

The component C^i of the vector C represents the proper capital reserve the ith bank; it is *solvent* if the net worth

$$\mathrm{NW}^i := \sum_{j \in G} L^{ji} - \sum_{j \in G} L^{ij} + C^i \geq 0\,.$$

If the above solvency condition does not hold, the bank *defaults*.

It is important to note that the definitions "exposure," "liability," and "default" appeal to a common sense rather having a precise meaning. Their understanding varies from paper to paper. In practice, the balance sheet of a bank has a much more complicated structure. For example, the exposure may include overnight credits as well as long term loans, the debts are of different seniority, and so on. The "standard" highly stylized balance sheet, i. e., the equality $\mathrm{Assets} = \mathrm{Liabilities}$ presented as a table, containing the interbank exposures (loans) and external assets (that can be split in liquid and illiquid, risky and nonrisky) on the assets sides while on the liability side, there are interbank borrowings, deposits, and, to equate the both side, the net worth (called also capital reserve, or equity) — in the case that the bank is solvent.

2.2 Defaults

In the literature, the typical descriptions of the contagion process and defaults "in cascade" can be found (e. g., in [1]). Here, they are presented in a succinct way as follows. Let us denote by $I_{\rm out}(i)$ (respectively, by $I_{\rm in}(i)$) the set of banks to which the bank i has a liability (respectively, an exposure). That is, $I_{\rm out}(i)$ is the set of nodes terminal for the fleshes (arcs) outgoing from the node i while $I_{\rm in}(i)$ is the set of nodes with fleshes ending at this node. Let us denote by $n_{\rm out}(i)$ and $n_{\rm in}(i)$ the cardinality of the corresponding sets, i. e., the numbers of outgoing and ingoing fleshes. Clearly, $n_{\rm out}(i) = \sum_j B^{ij}$ and $n_{\rm in}(i) = \sum_j B^{ji}$.

The default of the bank i triggers the following procedure. The bank is excluded from the network. Debts are collected from debtors at liquidation. Creditors loose a fraction $1 - R$ of their exposures to i, where the parameter R is referred to as *recovery rate*. Formally, one can think that the matrix L is replaced by the matrix $\bar{L}$ obtained by replacing the elements of the ith row and ith column by zeros. The transformed vector of capital reserves $\bar{C}$ has the components $\bar{C}^j = C^j + RL^{ij} - L^{ji}$, $j \neq i$, $\bar{C}^i = 0$. Put $D_0(i) := \{i\}$ and skip further the argument i here and in further definitions (depending also on R). For some j (different from i), the solvency condition

$$\sum_{k\in G\setminus D_0} \bar{L}^{kj} - \sum_{k\in G\setminus D_0} \bar{L}^{jk} + \bar{C}^j \geq 0\,,$$

which can be written also as

$$\sum_{k\in G} L^{kj} - \sum_{k\in G} L^{jk} + C^j - (1-R)L^{ij} \geq 0\,,$$

may fail. Let us denote by $D_1 := D_1(i)$ the set of such indices, corresponding to the first-order defaults in the cascade of the defaults triggered by the default of i. In the same way, the contagion is propagated further, to the set of banks $D_2 = D_2(i)$ which is a subset of indices j outside of the union D_0^1 of D_0 and D_1 and such that the solvency condition becomes negative:

$$\sum_{k\in G} L^{kj} - \sum_{k\in G} L^{jk} + C^j - (1-R)\sum_{k\in D_0^1} L^{kj} < 0\,.$$

Continue in the same way, for the set D_0^n, put $D_0^{n+1} := D_0^n \cup D_{n+1}$ where D_{n+1} is the set of indices j in the complement of D_0^n such that

$$\sum_{k\in G}^{N} L^{kj} - \sum_{j\in G} L^{jk} + C^j - (1-R)\sum_{k\in D_0^n} L^{jk} < 0\,.$$

The process stops if $D_{n+1} = \emptyset$. One can consider the value

$$L(i) := (1-R)\sum_{n=0}^{N}\sum_{j\in D_{n+1}}\sum_{k\in D_0^n} L^{jk}$$

as the total losses incurred by the cascade of defaults triggered by the default of the ith bank.

It is not difficult to extend the above definitions to obtain expressions for losses triggered by simultaneous defaults of a group of banks.

2.3 Practical aspects and difficulties

On the first sight, the above formulae are of great help for the researchers in financial systemic risk providing them N functions of the recovery rate which allows to classify banks according to their systemic importance. The described procedure can be also used to find the most vulnerable banks, sensitive to defaults of others. However, the practical implementation is not so straightforward. Indeed, in the majority of cases, the exposure matrix X (having one million entries for a system with $N = 1000$) is not publicly available though a certain subset of its entries may be known. Usually, only the sums of elements along each line and each column can be extracted from the balance sheets. If only this information is available, one cannot recover the matrix L in a unique way: one needs to solve the system of $2N$ equations

$$\sum_{j\in G} L^{ji} = a^i\,;\ \sum_{j\in G} L^{ij} = b^i\,,\ \ 1 \leq i,j \leq N\,, \qquad (1)$$

with $N^2 - N$ unknown $L^{ij} \geq 0$, $i \neq j$, and all $L^{ii} = 0$.

Obviously, system (1) has the nonnegative solution $x^{ij} = a^j b^i / \sum_i b^i$ (note that $\sum_i b^i = \sum_j a^j$). But this is not the needed solution since not all $x^{ii} = 0$. In the literature (see, for example, [14]), it is recommended to take as the matrix X the solution of the entropy minimization problem:

$$\sum_{ij} \ln \frac{L^{ji}}{x^{ji}} \to \min\,,$$

under constraints (1), $L^{ij} \geq 0$ and $L^{ii} = 0$ for all i, j.

This approach is criticized since it leads to a matrix generating a complete graph and the overestimation of stability of financial system. On the other hand, in some cases, a part of the matrix L is known, e. g., the absence of connections between some nodes can be a plausible hypothesis. The entropy minimization method can be easily adapted to such cases leading to a rather realistic recovery of the exposure matrix.

2.4 Probabilistic modeling

Due to the lack of the information on the real structure of the financial system, there is an interest to generate

numerically models which have, at least, basic features of such models.

Apparently, the liability matrix L and the reserve vector R are random and evolve as stochastic processes. Due to the high dimensionality of the problem, their modeling is extremely complicated and simplifying assumptions are unavoidable. The majority of available studies consider static models or stationary models and start modeling with the description of the incidence matrix B.

The simplest model is based on the hypothesis that the nondiagonal elements of the incidence matrix B are independent identically distributed Bernoulli random variables (see, e.g., [15] where low-parameter models are suggested to evaluate the impact of various factors on the financial stability). In addition to N and $p = P(B^{ij} = 1)$, there are three more parameters: the total value of assets A, the value of external assets C, and the net worth as the percentage of the total value of assets γ. These parameters are used to generate the balance sheets. In our notations, the interbank exposures and liabilities for the ith bank are defined as follows:

$$a^i = (A - C)\frac{n_{\rm in}(i)}{|B|}\,; \quad b^i = (A - C)\frac{n_{\rm out}(i)}{|B|}$$

where $|B| := \sum_{ij} B^{ij}$. The value of external assets of the bank are defined by the formula:

$$C^i = \left(b^i - a^i\right) I_{\{a^i < b^i\}} + \frac{1}{N}\left(C - \sum_j \left(b^j - a^j\right) I_{\{a^j < b^j\}}\right).$$

If the second term is positive, then all banks in the system are solvent. Since a^j and b^j are random, one should have a sufficiently high ratio C/A (in [15], it was always taken greater than 0.3). The quantity $\gamma(a^i - b^i + C^i)$ models the net worth while $(1-\gamma)(a^i - b^i + C^i)$ stands for the consumer deposits.

3 Dynamic Models and Alert Indicators

3.1 Structural model

The aim of the model is to provide regulators two functions on the current state of the system which can be used to calculate the alert indicators. The first one is the probability that the system will suffer a cascade of defaults before a specified time horizon. The second indicator is the total losses incurred by the cascade of defaults, if it happens.

Suppose that at time zero, the regulators dispose the liability matrix L or its transpose the exposure matrix $X = L'$ (in reality, this information is public only in rare countries, like Brasil, but can be available to central banks) and the vector of capital reserves C which is decomposed into nonrisky reserve c (say, Treasury bonds) and investments y into a single risky asset which can be interpreted as a market index, or a market portfolio. In the present very stylized model, all these values are fixed up to the time horizon T. Of course, in reality, the banks trade and portfolios are composed in many assets. Nevertheless, quite often banks mimic the behavior of each other and one may guess that a typical portfolio value has the same evolution as a certain reference portfolio. Let us describe its dynamics by a geometric Brownian motion:

$$\frac{dS_t}{S_t} = \mu\, dt + \sigma\, dW_t\,, \tag{2}$$

that is,

$$S_t = S_0 e^{\sigma W_t + (\mu - \sigma^2/2)t}\,.$$

At time zero, all banks are supposed to be solvent.

The default cascade will be triggered at the instant when one of the solvency conditions is violated.

The solvency condition for the ith bank has the form:

$$V_t + y_0^i S_0 e^{\sigma W_t + (\mu - \sigma^2/2)t} \geq 0\,.$$

Here,

$$V_t{:=}b_t^i - a_t^i + c_t^i$$

where

$$b_t^i{:=}\sum_{j\in G} L_t^{ji}\,; \qquad a_t^i := \sum_{j\in G} L_t^{ij}\,.$$

Hypothesis: $V_t = V$ for all $t \in [0, T]$.

The above assumption allows one to provide the regulators some easily calculated indicators of the system stability. Without any doubts, in the present oversimplified form, they can be criticized. For example, assume that the interbank operations to a large extend are balanced by liquid assets. In favor of this are evidences that interbank lending is not the main activity of banks. Let us also assume a rigidity of the investment portfolio. Again, econometric studies confirm that banks have a tendency to follow similar behavior. The benchmark portfolio process may have various dynamics and various theoretical and statistical models can be used for its description.

Put

$$\lambda^i := \frac{1}{\sigma}\ln\frac{V^i}{y^i S_0}$$

with a convention that $\lambda^i := -\infty$, if $V^i \leq 0$. Let i_0 be the index corresponding to the largest of values of λ^i. One may assume, with very minor loss of generality, that all finite values of λ^i are different (the coincidence is not

expected in the present context) and that λ^{i_0} is finite (otherwise, there will be no defaults).

Let us introduce the stopping time

$$\tau := \inf\left\{t \geq 0 :\ w_t + \beta t \leq \lambda^{i_0}\right\}$$

where $\beta := \mu/\sigma - \sigma/2$. If $\tau \leq T$, the system will have a default and it happens with the node i_0; the price of the market portfolio at this date will be $S_0 e^{\sigma\lambda^{i_0}}$. The distribution of τ is the well-known inverse Gaussian distribution (see, e. g., [16]) and one has that

$$P(\tau \leq T) = \Phi(h_1(T)) + e^{2\beta\lambda^{i_0}}\Phi(h_2(T))$$

where Φ is the standard Gaussian distribution function;

$$h_1(T) := \frac{\lambda^{i_0} - \beta T}{\sqrt{T}}; \qquad h_2(T) := \frac{\lambda^{i_0} + \beta T}{\sqrt{T}}.$$

The default of the bank i_0 generates a cascade of the defaults. It seems reasonable to suppose that the market reacts to such an event and the risky asset may loss a certain percentage of its value. With this assumption the set $D_1 = D_1(i_0)$ of the first-order defaults of the banks correspond to the indices j such that

$$\sum_{k\in G} L^{kj} - \sum_{j\in G} L^{jk} + c^j + \alpha y^j S_0 e^{\lambda^{i_0}} - (1-R)L^{ij_0} < 0\,,$$

$D_0^1(i_0) := D_0 \cup D_1$, etc. The parameter $\alpha \in]0,1]$ represents the default impact on the price of the reference portfolio.

The second alert indicator is the amount of total losses:

$$L(i_0) := (1-R)\sum_{n=0}^{N}\sum_{j\in D_{n+1}}\sum_{k\in D_0^n} L^{jk}\,.$$

In the considered setting, it can be augmented, e. g., by the losses of nondefaulted banks due to a depreciation of their portfolios:

$$\tilde{L}(i_0) := (1-\alpha)\sum_{j\in G\setminus D_0^N} y^j S_0 e^{\lambda^{i_0}}\,.$$

3.2 Discussion

The model introduced above has an advantage of its simplicity. It combines structural approach to defaultable securities with ideas of modern theory of financial networks. The alert indicators have a simple and comprehensive meaning. They can be easily computed at the monitoring dates t_m (when the new balance sheets are communicated) for the moving time horizons $t_m + T$. This allows regulator to see dangerous trends in the evolution of the system. It is worth noting that the model combines two channels of contagions: via the network as well as via the correlation due to common source of randomness.

Surely, the model is highly stylized. How serious are the weak points and how the model can be improved?

It is assumed that the investment in the single risky asset are static though in reality, there is an intensive trading. For a fixed input, there is only the bank triggering the default is uniquely determined.

To our mind, these objections should be examined carefully. Due to extreme complexity of financial systems (recall that they may contain hundreds of banks) and complexity of individual balance sheets, for more sophisticated models, one can have an accumulation of various factors: misspecification errors, calibration errors, data aggregation errors, etc. That is why, simplifying hypotheses seems to be inevitable. It seems that one can accept that banks investment portfolios are close to the most performant one.

Of course, the predetermined bank triggering of the default cascade is not intuitive. However, as it is known from the literature, the matrix L is rarely known and should be reconstructed from the aggregated exposure of the banks. It is not difficult to implement a random reconstruction procedure; for each realized reconstruction, one can compute conditional alert indicators and take the average.

4 Numerical Experiments

4.1 Network construction

Here, the numerical simulations are presented to offer further insight into the role of the external assets in contributing to a systemic risk in the financial system and to show an impact of parameters range to financial stability. Table 1 summarizes the baseline simulations parameters. The system comprises N banks. As in [17], the banking system is considered as a network of nodes

Table 1 Summary of the parameters considered in the network construction of the model

Parameter	Description	Value	Range of variation
n	Number of banks in the initial random graph	10	fixed
N	Total number of banks in the network	250	fixed
m	Maximum number of connections in the random network	5	fixed
m'	Maximum number of connections in the scale free network	3	fixed

where each node represents a bank (or any other financial institution) and each link represents a directional lending relationship between two nodes (two banks). We believe that network reflects its "genetic" structure. The development of the system starts from relatively small kernel composed from a few banks. A newly created bank establishes relationships preferably with more "important" nodes of the network, namely, those that already have more connections than others. Also, usually, well connecting bank has better chances not to be eliminated from the system ("too connected to fail").

As an example, let us make a look on the development of the banking system in Russia. In the USSR, the number of banks were about a dozen. The first commercial bank was registered in August 1988 (Cooperative bank "Patent," Leningrad). Already to the end of 1989, the country had 43 commercial banks. Afterwards, the number of banks in Russia evolved (approximately) as follows: 1991 — 1400; 1994 — 2300; 1997 — 2000; 2003 — 1300; and 2011 — 1000.

In 2016, the total number was less than 700. To compare: in 2014, the USA had more than 6800 banks.

Of course, the evolution of the banking system is a rather complicated process but statistically, it leads to a network having common features with other types of networks like Internet connections. In particular, interbank networks have a scale-free topology with a few large banks having many interconnections and many small banks with a few connections. By this reason, we generate our simulating financial system using the methodology introduced in [13]. We are based on the idea that a larger and more connected bank is usually more trusted and, as a consequence, other system banks tend to deal with it rather than with the less connected one.

The algorithm starts with creating the initial network (the "seed") with a small amount of nodes n. The connections between them are taken randomly, and the maximal number of connections ("in" and "out") is limited by m as well as the total number M of connections (in the present experiments, $n = 10$, $m = 5$, and $M = 20$). When drawing the network, we allow for the possibility that two banks can be linked to each other via both lending and borrowing links but at most one in the same direction is possible between the 2 banks (Fig. 1).

Remark. In general, we expect that the structure of the initial network has a relatively small impact on the resulting network which may be in dozens or even hundred times larger than the "seed." So, the initial network involving a few nodes can be created in various way, e. g., as Erdos–Rényi network with the matrix $B = (b^{ij})$ where entries $b^{i,j}$, $i \neq j$, form a set of independent identically distributed Bernoulli random variables with $P(b^{i,j} = 1) = p$.

To generate a scale free network, assuming that a seed network of n banks is randomly generated as mentioned above, let us proceed recursively. At each step, add $m' < m$ new nodes choosing each time a partner i between the existing nodes and selecting according to the Connection Probability $P(i)$ defined as follows:

$$P(i) = \frac{\text{total number of connections of } i}{\text{total number of connections of the network}}$$

where i is the existing node in the network. The nodes will be added each time until a network size limit equal to N is reached. As shown in Fig. 2, the distribution of a number of nodes in the considered model is in a conformity with that one can expect from a typical scale-free network topology where a few nodes have a high number of connections while the majority of nodes have a small

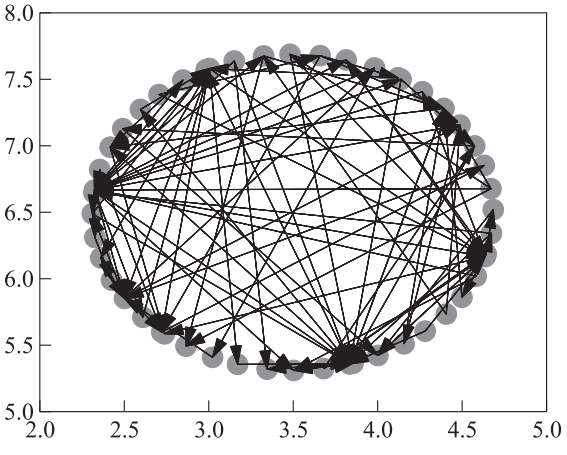

Figure 1 Illustration of a financial network generated by the algorithm and limited to 40 banks for clarity of the figure

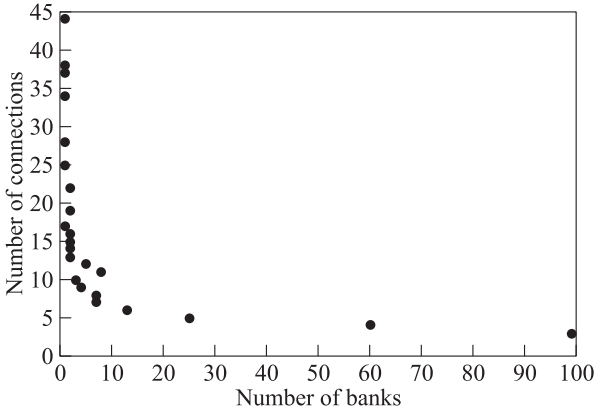

Figure 2 Representation of the scale-free topology

Table 2 Summary of the system benchmark parameters

Parameter	Description	Value	Range of variation
A	Value of the total assets of the network	10,000,000,000	fixed
β	Proportion of external assets from the total assets of a bank	0.5	fixed
S_0	Initial price of risky asset	10,000	fixed
T	Time horizon	40	fixed

Table 3 Summary of the balance sheet parameters

Parameter	Description	Value	Range of variation
α	Percentage of price degradation following the panic in the market during the crisis	1	0 to 1
R	Recovery rate	0.5	0 to 1
θ	Percentage of riskless asset from total external asset	0.4	0 to 1
σ	Volatility of the risky asset price	0.4	0 to 1
μ	Drift of the risky asset price	0.2	0 to 1

number of connections. In particular, on the realization of the algorithm depicted on the scatterplot, only 6 banks from 250 have each at least 25 connections while 185 have at most 5 connections.

For any realization of the random graph, we populate the individual banks' balance sheets in a manner consistent with bank level and aggregate balance sheet identities (Table 3).

Amongst assets, let us distinguish external assets (investors' borrowing), denoted by C^i, and interbank assets (other banks borrowing), denoted by a^i. Thus, for the bank i, the asset part of the balance sheet can be decomposed as

$$\text{Total assets} = C^i + a^i\,, \qquad i = 1, \ldots, N\,.$$

Moreover, the external assets can be of two types: risky r^i and riskless (cash) l^i, so that $C^i = r^i + l^i$. Introducing the parameter θ as a proportion of cash holdings, the volume of the risky assets is $r^i = (1-\theta)C^i$. The liabilities of each bank are composed of the net worth of a bank, denoted by NW^i, and the interbank borrowing, denoted by b^i. Hence, for the bank i, one has:

$$\text{Total liabilities} = NW^i + b^i\,, \qquad i = 1, \ldots, N\,.$$

An example of a bank balance sheet as generated by the simulator is also shown in Fig. 3.

The balance sheets have been constructed for individual banks in a sequence of steps. The entry parameter is the total value of all assets in the system denoted by A and the parameter β which defines the proportion of the external asset C representing the total loans made to ultimate investors and thus relating to the total size

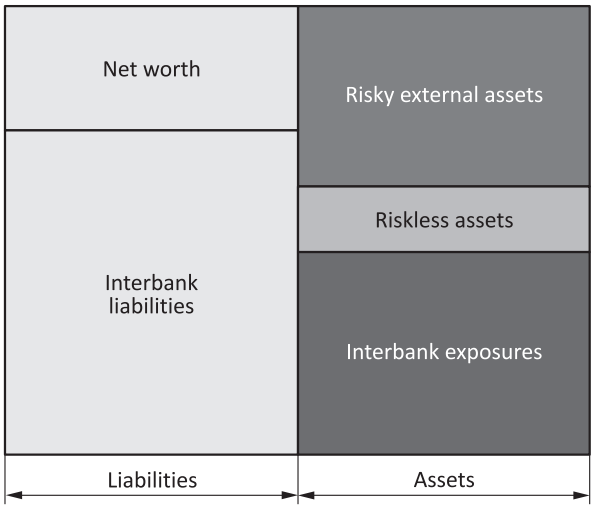

Figure 3 Representation of the balance sheet

of the flow of funds from savers to investors through the banking system. That is, $\beta = C/A$. The aggregate assets of the whole banking industry can be written as $A = C + I$ where $I = (1-\beta)A$ represents the aggregate volume of interbank exposures, i. e., $I := \sum_{i,j} L^{ij}$.

Dividing the total interbank assets by the total number of nodes in the network, one arrives at the level of each bank. So, weights of all links are equal banks borrow and lend by equal portions $w = I/|B|$. Though this looks not very realistic, such a hypothesis is accepted to reduce the number of parameters. Hence, using w and the structure of the network, one can calculate for each bank the volume of its liabilities $a^i = n_{\text{in}}(i)w$ and exposures $b^i = n_{\text{out}}(i)w$. For any bank to be able to operate, the value of its external assets is required to be

not less than its net interbank borrowing, that is, one has: $C^i \geq b^i - a^i$. Let us fulfill this constraint by applying the following two-step algorithm.

First, for each bank, fill up the bank external assets part of the balance sheet in such a way that its external assets plus interbank lending will equalize its interbank borrowing. That is, provide first the bank i the volume $\tilde{C}^i = (b_i - a_i)^-$ where $\tilde{C}^i$ is the fraction of the total volume C reserved for the external assets. At the second stage, what is left in aggregated external assets is equally distributed among all banks. Note that the total of external assets is equal to C. Hence, in the second step, distribute $\bar{C} = C - \sum_{i=1}^{N} \tilde{C}^i$ equally among all N banks. Hence, one has $C^i = \tilde{C}^i + \bar{C}/N$. The constraint can become difficult to meet if the percentage of external asset is too low. Since the distribution of links is stochastic, some banks may be assigned interbank borrowing much larger than interbank lending. When the total amount of external assets is low, there may be not enough assets to go round to close all balance sheet gaps opened up in this way. To avoid this difficulty, make sure that the total volume of external assets is at least 30% of the total volume of all assets.

Although the model is applied to fully heterogeneous banks, for the purpose of illustration and simplicity, let us consider one common risky asset for all banks (full correlation among the banks). Further studies can be conducted for portfolios composed of many different risky assets. Furthermore, let us choose the risky asset evolving according to value of a reference portfolio whose dynamics follows a geometric Brownian motion (2).

Since $S_t = S_0$ at time $t = 0$, one can define y^i as the amount of the risky assets in the portfolio of each bank with $y^i = r^i/S_0$.

Hence, the asset side of the bank balance sheet is completed as well as interbank borrowing b on the liability side. The determination of the remaining component, the net worth (equity) NW on the liability side, is relatively straightforward. The net worth is set as $\mathrm{NW} = a + C - b$. This completes the construction of the banking system and of each constituent bank balance sheet.

Now, let us calculate the probability of the first default. Then, let us specify which bank will first default to check the price of the risky asset at the time of default S_τ where τ is the time of the first default (1st stopping time). This generates a loss in the asset price from S_0 to S_τ on each bank i which will suffer a loss of $y^i(S_0 - S_\tau)$. The sum of these losses is equal to $\sum_{i=1}^{N} y^i(S_0 - S_\tau)$ and is denoted as Corr_loss (correlation loss). In what follows, it is assumed that the first default will affect the neighborhood and will trigger a cascade of default due to the loss transmission through the interbank connections. The sum of the total losses generated by the cascade of default is denoted as Con_loss and is equal to $(1 - R) \sum_{n=0}^{N} \sum_{i \in D_{n+1}} \sum_{\in D_n}$. Having the probability as well as

$$\text{Network total loss} = \text{Correlation loss} + \text{Contagion loss},$$

one can calculate

$$\text{Probable loss Indicator} = P \cdot \text{Total loss}\,.$$

In what follows, the parameters will be varied in the experiments.

4.2 Experiment 1: The influence of the external assets volume and its riskless proportion on the probability and losses

Given the balance sheet above we want to compute the probability of default when the proportion of the risky asset varies. For the sake of more profit and wealth, banks have the choice to invest in either risky assets or riskless assets. It is known that the revenue in risky assets can be much larger than the riskless ones. Thus, banks have more tendency to invest there while keeping an eye on the risks and their liabilities towards other banks making sure they can settle when needed. Therefore, let us check the probability of the first default that can trigger a cascade of defaults in the system based on the percentage of the risky assets. Regulators may impose a minimum threshold on the level of riskless assets (θ) held by banks as well as define a certain proportion that a bank can invest in external assets. That is why, both parameters will be considered in this experience to check how both of them can affect the vulnerability of the system and what is the optimal requirement on the riskless assets thresholds for each volume of external investments.

As for the following figures, this experiment illustrates that with an increase of the riskless assets level, the probability of the first default decreases to become zero after a certain threshold of θ. Also, note that this threshold is smaller with an increasing volume of assets β (Fig. 4).

For example, for $\beta = 0.55$, the probability of default is always less than 0.5 decreasing with θ and becomes null after $\theta = 0.6$. If the system is engaged with high level of external investments $\beta > 0.75$, the probability of default is very low with a very low rate of riskless assets $\theta = 0.55$. This means that the system is rather stable even if the level of risky assets is high. On the other hand, for a high level of external assets investments as shown in Fig. 5, the probability of the first default remains low even with high level of risky assets.

In the following experiments, $\beta = 0.5$ will be fixed as it is assumed that banks have the same probability to invest in external and internal assets.

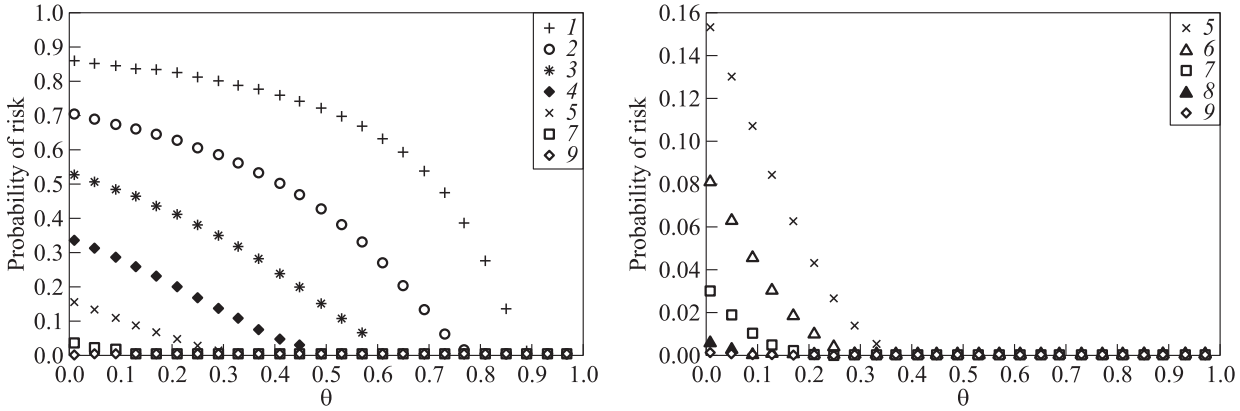

Figure 4 Probability of default with β percentage of external assets and θ proportion of riskless assets: *1* — $\beta = 0.35$; *2* — 0.45; *3* — 0.55; *4* —0.65; *5* — 0.75; *6* — 0.80; *7* — 0.85; *8* — 0.90; and *9* — $\beta = 0.95$

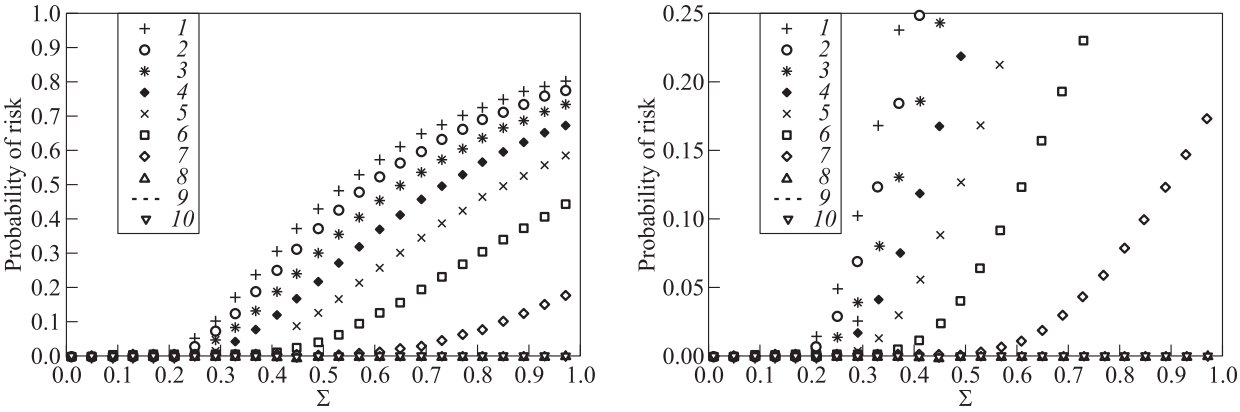

Figure 5 Probability of default with θ and volatility of the asset price: *1* — $\theta = 0.1$; *2* — 0.2; *3* — 0.3; *4* — 0.4; *5* — 0.5; *6* — 0.6; *7* — 0.7; *8* — 0.8; *9* — 0.9; and *10* — $\theta = 1.0$

4.3 Experiment 2: The influence of the volatility and drift of the risky asset price

Now, let us study how the volatility and drift of the asset price can also affect the probability taking into consideration also the level of the risky assets in banks balance sheet. The volatility reflects the risk of changing the portfolio price due to external and internal factors.

Assuming that the portfolio of risky asset price has a volatility that varies from 0.1 to 1, one can conclude that for the volatility less than 0.25, the probability of default is low for any θ level. When volatility increases above 0.25, the probability of default increases with θ decreasing and volatility increasing. The pattern of every plot is changing with θ: having the plot concave for high θ and convex for low θ shows that the behavior of the volatility influences more the probability of default since for higher θ, the increase in probability is faster than the lower one. So, summarizing, the volatility should be limited to a certain extent in order to save the network from a high probable default; otherwise, a high impact will be affecting the asset price which, in its turn, will trigger a higher indicator of default.

Moreover, we also evaluate the effect of the asset price drift on the probability considering at the same time the level of the risky assets in banks' balance sheet. We conclude that a portfolio price with high drift or high average of return will, for sure, lead to a more stable financial system. The optimal portfolio would be with high drift and low volatility.

Figure 6*a* shows for each level θ the variation of the probability relatively to the drift. In particular, one can see that for $\theta = 0.5$, the level of drift required to assure a probability less than 0.2 is also 0.5 but if $\theta = 0.6$, one can see that in this case, the required level of drift is 0.05.

Figure 6*b* highlights the relation between the volatility and the drift affecting the probability of default. It is normal that the probability of default increases with increasing volatility and decreases with increasing drift. But from Fig. 6*a*, one can also see that when volatility is very high, the influence of the increasing drift is reduced.

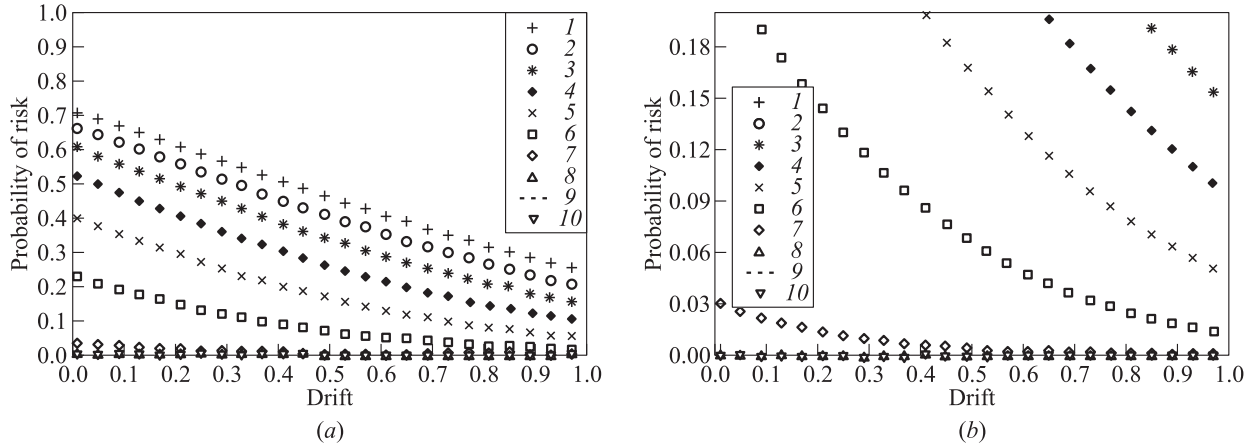

Figure 6 Probability of default with θ and the drift of the asset price: *1* — $\theta = 0.1$; *2* — 0.2; *3* — 0.3; *4* — 0.4; *5* — 0.5; *6* — 0.6; *7* — 0.7; *8* — 0.8; *9* — 0.9; and *10* — $\theta = 1.0$

4.4 Experiment 3: The impact of recovery rate and the level of the riskless asset in the banks' balance sheet on the losses

In the present model, the system loss is another parameter that influences the indicator as probability of the first default can be low; however, when it happens, the loss in the system can be huge. Therefore, we analyze the behavior of the network when recovery rate is changing knowing that recovery rate is one of the main influencer of the losses due to the default cascade. Figure 7 shows that the total loss in the system decreases when recovery rate is increasing due to the fact that defaulted banks need to settle their due payment. One can clearly observe that there is a linear relation between the recovery rate and the contagion losses.

4.5 Experiment 4: The impact of the fire sale on the losses

Also, let us consider a subordinate source of risk due to the fire sale of external assets of defaulting banks which will lead to other banks default because of the price depreciation. In bad times, in order to compensate certain losses, distressed banks tend to sell assets in a depressed price, a situation called asset "fire sale." Because of the correlation between the banks' balance sheets having common assets, the decrease in the asset price affects all banks holding these assets and, as a consequence, a cascade of losses created by others banks, too. In this experiment, we want to compute what will be the losses resulting from the "fire sale" mechanism that will respectively add to the total loss of the system (Fig. 8). It is expected that the fire sales loss will increase with the drop rate of the price but it is worth to note that the increase is very sharp and fast.

4.6 Experiment 5: Removing the weakest bank to make the system more resilient

The financial system is a number of financial institutions, in the present paper considered as banks. Every bank has its investment strategy, priorities, relationship, and connections as a consequence different influence,

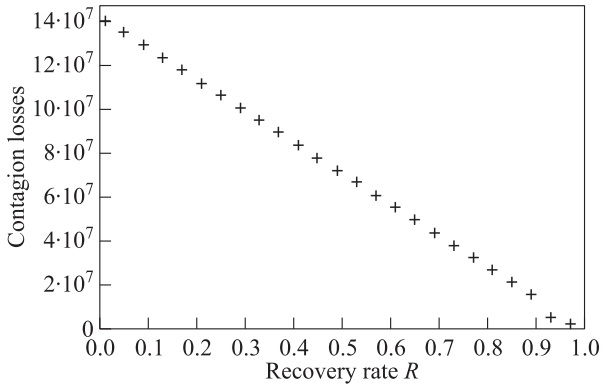

Figure 7 Contagion losses

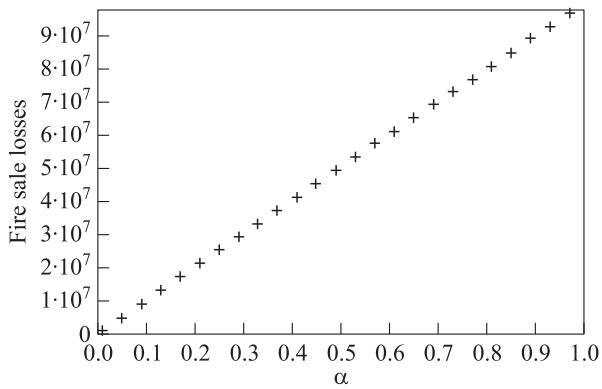

Figure 8 Fire sale losses

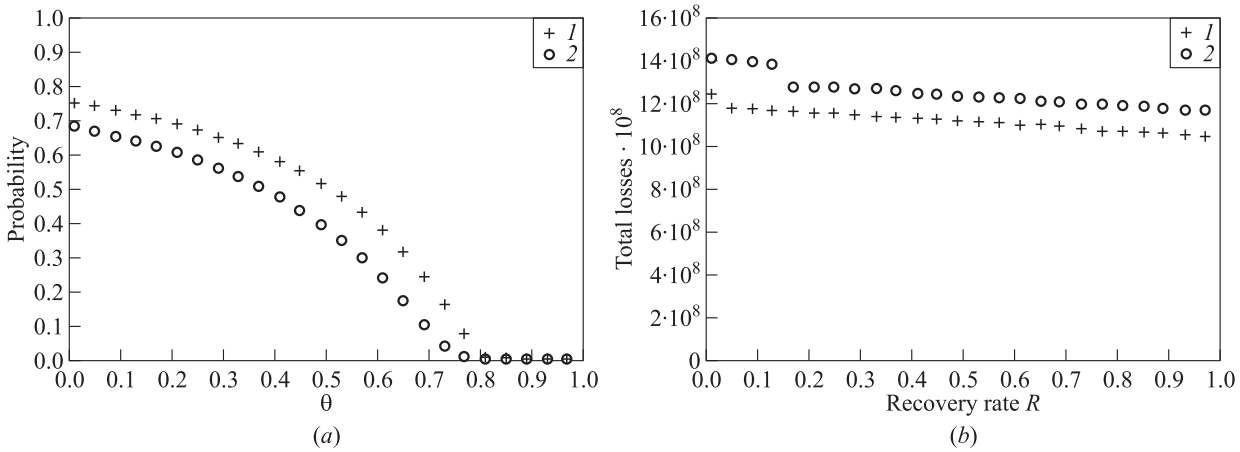

Figure 9 Comparison between before (*1*) and after (*2*) removing the weak bank(s): (*a*) probability of default; and (*b*) total system loss

power, and risk level in the financial system. Banks balance sheets are populated on the quarterly basis and can be available to regulators at any time. Thus, it is possible to understand the risk level of each bank to be defaulted. Since banks default can generate a default cascade, it is worth to verify if it is better to exclude the risky bank from the financial network. It is important to determine the weakest bank that can default anytime either due to a market price drop or liquidity shortage.

In this experiment, we create a scale-free network of 250 nodes using the same methodology as above and determine the bank i, having the highest probability to default or, in other words, the weakest bank to default. Also, we calculate, on the basis of the above, the probable loss Indicator of this network. In some cases, we may have more than one bank defaulting simultaneously.

Once the bank is identified, let us check if removing the bank from the network is healthier for the financial system. Of course, such an action has important consequences to the owner, employees as well as debtors and creditors in this bank. The modification on the balance sheet of this bank and all banks connected and in relation to this bank will be described as follows:

- removed bank pays all its liabilities to all its creditors and, as a consequence, updates its creditors' balance sheets a_j by adding the amount of money bank i has landed from bank j to the riskless assets;
- removed bank collects all its exposures from all its debtors and, as a consequence, updates its debtors balance sheets b_j by removing the amount of money bank i has credited to bank j. Bank j is supposed to pay to bank i from their external assets (riskless assets and risky assets; so, either bank j has to pay from its liquid reserve or sell some external assets to pay its due).

We conduct this experiment and compute the probability, total loss, and indicator having the vulnerable bank (banks) in the system and after removing it (them) from the network.

Figure 9 confirms that when removing the risky bank, the probability of default will decrease. This means that the system becomes more resilient to default.

Now, let us check the total losses that may occur in the system due to a default in the above two mentioned cases and note that, contrary to the probability, the total losses will be higher in the case where it has been decided to remove the vulnerable banks. The reason could be that the external assets prices that have decreased to S_τ causing the first default in the first round has to decrease more to trigger the default after removing the set of banks that could defaulted on S_τ. In this case, though the probability is lower, the increase in the total losses from before to after removing the banks is due to the fact that the risky assets price should drop more and, accordingly, the correlation losses increase. On the other hand, since the shock on the net worth of every bank becomes higher, we expect that the contagion losses are larger now. Having both correlations and contagion losses higher, this will lead to a high total loss as shown in Fig. 9. It is worthy to note that for small recovery rate, the increase of the total loss after removing the weak banks is even higher.

5 Concluding Remarks

In this paper, a financial network model of interbank interactions has been developed which incorporates a dynamical behavior of banks portfolios and combines it with cascade defaults. It is assumed that the portfolios contain, together with riskless asset, a unique risky asset,

which can be interpreted as a market index or a benchmark portfolio and whose price evolution is described by a geometric Brownian motion. This part of modeling follows the ideas of the so-called structural approach well known in the context of pricing defaultable securities. A crisis starts when the net worth of a bank hits zero triggering a cascade of defaults. The time to default and the total losses calculated for the "frozen" parameters of the balance sheets can constitute indicators of the "health" of financial system. They can be easily monitored, on a regular basis, by the regulators. Usually, the detailed structure of the system is available only to regulators but it is not public. By this reason, the present study is completed with numerical experiments with simulated data. Due to complexity of financial systems this is a nontrivial problem. The network graph is built by a version of preferable attachment algorithm augmented by a procedure of simulations of balance sheets. The results of the experiments are presented by plots showing dependences of the indicators as functions of parameters. In particular, an experiment with removing the weakest bank from the system is presented. We believe that developing this approach on the basis of practical data can provide regulator additional tools of monitoring vulnerability of banking system and measuring its stability.

Acknowledgments

This work was done under partial financial support of the grant of the Russian Science Foundation No. 14-49-00079.

References

1. Hurd, T. 2016. *Contagion! The spread of systemic risk in financial networks*. Springer. 139 p.
2. De Bandt, O., P. Hartmann, and J. Peydró. 2009. Systemic risk in banking: An update. *Oxford handbook of banking*. Eds. A. Berger, P. Molyneux, and J. Wilson. Oxford University Press. 994 p.
3. Gai, P., and S. Kapadia. 2010. Contagion in financial networks. Bank of England. Working Paper 383. 36 p.
4. Acharya, V., and T. Yorulmazer. 2008. Information contagion and bank herding. *J. Money Credit Bank*. 40(1):215–231.
5. Elsinger, H., A. Lehar, and M. Summer. 2006. Risk assessment for banking systems. *Manage. Sci.* 52(9):1301–1314.
6. Elsinger, H., A. Lehar, and M. Summer. 2006. Systemically important banks: An analysis for the European banking system. *Int. Econ. Econ. Policy* 1:73–89.3.
7. Gai, P., and S. Kapadia. 2010. Contagion in financial networks. *Proc. R. Soc. A* 466:2401–2423.
8. Allen, F., and D. Gale. 2000. Financial contagion. *J. Polit. Econ.* 108(1):1–33.
9. Acharya V., and A. Bisin. 2014. Counterparty risk externality: Centralized versus over-the-counter markets. *J. Econ. Theory* 149(C):153–182.
10. Georg, C. P. 2013. The effect of the interbank network structure on contagion and common shocks. *J. Bank. Financ.* 37(7).
11. May, R. M., and N. Arinaminpathy. 2010. Systemic risk: The dynamics of model banking systems. *J. R. Soc. Interface* 7:823–838.
12. Bielecki, T. R., and M. Rutkowski. 2002. *Credit risk: Modelling, valuation and hedging*. Berlin: Springer. 501 p.
13. Barabási, A., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286:509–512.
14. Mistrulli, P. E. 2007. Assessing financial contagion in the interbank market: Maximum entropy versus observed interbank lending patterns. Rome, Italy: Bank of Italy. Working Paper 641.
15. Nier, E., J. Yang, T. Yorulmazer, and A. Alentorn. 2008. Network models and financial stability. Bank of England. Working Paper 346.
16. Borodin, A., and P. Salminen. 2002. *Handbook of Brownian motion: Facts and formulae*. 2nd ed. Birkhäuser. 465 p.
17. Eboli, M. 2004. Systemic risk in financial networks: A graph-theoretic approach. Italy: Universitа di Chieti Pescara. Preprint. 19 p.

Received November 14, 2016

Contributors

El Bitar Khalil (b. 1981) — PhD student, Laboratoire de Mathematiques, Universite de Franche-Comte, 16 Route de Gray, 25030, Besançon, CEDEX, France; khalilbitar_aw@hotmail.com

Kabanov Yuri M. (b. 1948) — professor, Laboratoire de Mathematiques, Universite de Franche-Comte, 16 Route de Gray, 25030, Besancon, CEDEX, France; leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; National Research University "MPEI," 14 Krasnokazarmennaya Str., Moscow 111250, Russian Federation; Youri.Kabanov@univ-fcomte.fr

Mokbel Rita (b. 1981) — PhD student, Laboratoire de Mathematiques, Universite de Franche-Comte, 16 Route de Gray, 25030, Besancon, CEDEX, France; ritamokbel@hotmail.com

ДИНАМИЧЕСКИЕ МОДЕЛИ СИСТЕМНОГО РИСКА И ЗАРАЖЕНИЯ*

Х. Эль Битар[1], Ю. Кабанов[1,2,3], Р. Мокбель[1]

[1]Лаборатория математики Университета Франш-Конте, г. Безансон, Франция
[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, Российский университет дружбы народов
[3]Национальный исследовательский университет «МЭИ»

Аннотация: Современные финансовые системы являются сложными сетями взаимосвязанных финансовых институтов (банков, хедж-фондов, страховых компаний, и т. д.), и дефолт одного из них может вызвать цепную реакцию дефолтов других институтов системы. После недавних финансовых кризисов важность системного риска вышла на первый план, и теоретические исследования в этой области интенсифицировались. Бóльшая часть известных результатов относится к статическим моделям, которые посвящены процессам, происходящим в системе, когда каскад дефолтов уже начался. Авторы предлагают динамическую модель так называемого структурного типа, когда дефолт начинается в момент выхода некоторого стохастического процесса из области. Каскад инициируется в момент достижения критического уровня процессом, описывающим портфели банков. Вероятность выхода и суммарные издержки в результате каскада дефолтов могут служить индикаторами, позволяющими регуляторам осуществлять мониторинг системы и предпринимать упреждающие коррекции для понижения системного риска. Проводится численное моделирование системы, которая строится на основе случайного графа, полученного при помощи алгоритма предпочтительного присоединения. Приводятся результаты численных экспериментов при различных значениях параметров.

Ключевые слова: системный риск; финансовые сети; заражение; дефолт

DOI: 10.14357/19922264170201

Литература

1. *Hurd T.* Contagion! The spread of systemic risk in financial networks. — Springer, 2016. 139 p.
2. *De Bandt O., Hartmann P., Peydro J.* Systemic risk in banking: An update // Oxford handbook of banking / Eds. A. Berger, P. Molyneux, J. Wilson. — Oxford University Press, 2009. 994 p.
3. *Gai P., Kapadia S.* Contagion in financial networks. — Bank of England, 2010. Working Paper 383. 36 p.
4. *Acharya, V., Yorulmazer T.* Information contagion and bank herding // J. Money Credit Bank., 2008. Vol. 40. No. 1. P. 215–231.
5. *Elsinger H., Lehar A., Summer M.* Risk assessment for banking systems // Manage. Sci., 2006. Vol. 52. No. 9. P. 1301–1314.
6. *Elsinger H., Lehar A., Summer M.* Systemically important banks: An analysis for the European banking system // Int. Econ. Econ. Policy, 2006. Vol. 3. No. 1. P. 73–89.
7. *Gai P., Kapadia S.* Contagion in financial networks // Proc. R. Soc. A, 2010. Vol. 466. P. 2401–2423.
8. *Allen F., Gale D.* Financial contagion // J. Polit. Econ., 2000. Vol. 108. No. 1. P. 1–33.
9. *Acharya, V., Bisin A.* Counterparty risk externality: Centralized versus over-the-counter markets // J. Econ. Theory, 2014. Vol. 149(C). P. 153–182.
10. *Georg C. P.* The effect of the interbank network structure on contagion and common shocks // J. Bank. Financ., 2013. Vol. 37. No. 7. P. 2216–2228.
11. *May R. M., Arinaminpathy n.* Systemic risk: The dynamics of model banking systems // J. R. Soc. Interface, 2010. Vol. 7. P. 823–838.
12. *Bielecki, T. R., Rutkowski M.* Credit risk: Modelling, valuation and hedging. — Berlin: Springer, 2002. 501 p.
13. *Barabási A., Albert R.* Emergence of scaling in random networks // Science, 1999. Vol. 286. P. 509–512.
14. *Mistrulli P. E.* Assessing financial contagion in the interbank market: Maximum entropy versus observed interbank lending patterns. — Rome, Italy: Bank of Italy, 2007. Working Paper 641.
15. *Nier E., Yang J., Yorulmazer T., Alentorn A.* Network models and financial stability. — Bank of England, 2008. Working Paper 346.
16. *Borodin A., Salminen P.* Handbook of Brownian motion: Facts and formulae. — 2nd ed. — Birkhäuser, 2002. 465 p.
17. *Eboli M.* Systemic risk in financial networks: A graph-theoretic approach. — Italy: Universit à di Chieti Pescara, 2007. Preprint. 19 p.

Поступила в редакцию 14.11.2016

*Работа выполнена при частичной поддержке Российского научного фонда (грант № 14-49-00079).

INFORMATICS AND APPLICATIONS 2017. Vol. 11. Iss. 2. P. 16–24

ON THE EFFICIENCY OF BRIDGE MONTE-CARLO ESTIMATOR

O. V. Lukashenko[1], E. V. Morozov[2], and M. Pagano[3]

Abstract: Long-term correlation is a key feature of traffic flows and has a deep impact on network performance. Indeed, the arrival rate can persist on relatively high values for a considerable amount of time, provoking long busy periods and possibly bursts of lost packets. The authors focus on Gaussian processes, well-recognized and flexible traffic models, and consider the probability that the normalized cumulative workload grows at least as the length T of the considered interval. As T increases, such event becomes rare and ad-hoc techniques should be used to estimate its probability. To this aim, the authors present a variant of the well-known conditional Monte-Carlo (MC) method, in which the target probability is expressed as a function of the corresponding bridge process. In more detail, they derive the analytical expression of the estimator, verify its effectiveness through simulations (for different sets of parameters), and investigate the effects of the discretization step.

Keywords: Gaussian processes; conditional Monte Carlo; bridge process; rare events; variance reduction

DOI: 10.14357/19922264170202

1 Introduction

Key features of traffic patterns in modern computer networks are the high level of statistical multiplexing and, at the same time, strong correlations over several timescales [1]. In this framework Gaussian processes have emerged as well-recognized and flexible models able to describe the traffic dynamics of a wide class of networks [2, 3]. Indeed, they permit to capture, in a simple and parsimonious way, the properties of self-similarity and long-range dependence, which have a deep impact on network dimensioning and QoS (Quality of Service) issues [4]. In a nutshell, self-similarity means that the distribution of the process remains unchanged under suitable scaling of time and space, while long-range dependence (also known as Joseph effect) [5] implies a slow decay of the autocorrelation function.

Network performance are, in general, deeply influenced by packet losses and many works have been devoted to the estimation of the overflow probability in presence of long-range dependent traffic (see, for instance, [3] and references therein). However, not only the loss rate is relevant, but also the way in which losses are distributed over time. Indeed, bursts of losses can significantly degrade the QoE (Quality of Experience) in case of real-time multimedia applications and also affect the throughput of elastic applications, since TCP congestion control [6] poorly reacts in presence of multiple losses during the same congestion window, which often lead to the expire of timeouts (instead of the AIMD (Additive Increase Multiplicative Decrease) behavior that happens when losses are detected via triple duplicate acknowledgements). Such bursts of losses are often determined by high-activity periods that last for relative long intervals of time, a typical consequence of the above-mentioned Joseph effect.

Moreover, the properties of long-memory and self-similarity make difficult the theoretical analysis even for simple queuing systems and, as a consequence, simulation is often the only available tool to investigate network performance.

Simulation permits to study the performance of complex systems with an arbitrary level of detail, but the traditional approach, known in the literature as crude MC, becomes highly inefficient when the event of interest gets rarer and rarer. Indeed, given a required level of accuracy (typically expressed in terms of relative error), the length of the simulation is inversely proportional to the target probability, which can assume (for instance, in the case of high-quality video flows) values of the order of 10^{-9} [7]. Moreover, every estimate may be related to the simulation of complex networks and so includes the generation of a huge amount (of the order of millions or more, depending on the time horizon and the complexity of the system) of random numbers, with additional concerns related not only to the length of the simulation, but also to the goodness of the random generator itself.

Variance reduction techniques aim at achieving the desired accuracy with a lower number of samples [8],

[1]Institute of Applied Mathematical Research of Karelian Research Centre of the Russian Academy of Sciences, 11 Pushkinskaya Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; Petrozavodsk State University, 33 Lenin Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation, lukashenko@krc.karelia.ru

[2]Institute of Applied Mathematical Research of Karelian Research Centre of the Russian Academy of Sciences, 11 Pushkinskaya Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; Petrozavodsk State University, 33 Lenin Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; emorozov@karelia.ru

[3]University of Pisa, 43 Lungarno Pacinotti, Pisa 56126, Italy; m.pagano@iet.unipi.it

but require some additional information about the behavior of the system, typically provided (although in an asymptotic and eventually approximate form) by the Large Deviation Theory (LDT) [9]. Among them, Importance Sampling (IS) is probably the most popular approach [10] due to its links with LDT[1], providing a solid theoretical background for its applicability. However, under an improper choice of the change of measure (the optimal choice is known just for simple queuing systems), the variance may even grow infinitely [11].

In this paper, based on its predecessors [12, 13], the focus is on an alternative approach, known as Conditional MC, in which the target probability is expressed as a conditional expectation. Although conditional MC always leads to variance reduction, it is often impossible, or at least very difficult, to find a suitable conditioning quantity. However, in case of Gaussian processes, the target probability can be easily expressed as a function of the corresponding bridge process, as originally proposed in [14] under the name of *Bridge Monte Carlo* (BMC), for the estimation of the overflow probability. As mentioned above, in this work, the authors investigate the applicability of BMC to the tail distribution of the duration of high activity periods, which indeed become rare events when the duration of the considered interval goes to infinity. In comparison with the preliminary work [12], the experimental results have been widely extended: indeed, the authors investigated the asymptotic properties of the estimator, compared its performance with a basic version of IS, and analyzed the effect of the discretization step on the estimated probability.

The rest of the paper is organized as follows. In Section 2, the authors formally define the problem addressed in this work and recall the few available asymptotic results. Then, Section 3 addresses the general issues related to rare event simulation, including the basic definitions about simulation efficiency and a short description of (single-twist) IS, the most widely used variance reduction technique that will be considered later for performance comparison. The use of the bridge process is investigated in Section 4, while its performance is analyzed in Section 5, taking into account the effect of different parameters (such as the length of the activity period, the conditioning point, and the discretization step). Finally, Section 6 ends the paper with some final remarks.

2 High Activity Periods for Gaussian Processes

In traffic modeling, Gaussian processes have emerged as a flexible and powerful tool, able to take into account the long memory properties of real traffic, while keeping a relatively simple and elegant description.

In this work, a centered Gaussian process with stationary increments $\{X_t,\, t \in \mathbb{R}_+\}$ is considered. Let us denote by $v_t := \mathrm{Var} X_t$ the variance of X_t; then, the covariance function has the following expression:

$$\Gamma_{s,t} = \frac{1}{2}\left(v_t + v_s - v_{|t-s|}\right) .$$

It is interesting to estimate the following probability:

$$\pi(T) := \mathbb{P}\left(\forall t \in \mathbb{T} :\, X_t > t\right) \tag{1}$$

where $\mathbb{T} = (0, T] \subseteq \mathbb{R}_+$. Such probability is closely related to the duration of busy periods and plays an important role in the study of QoS indexes since it takes into account bursts of losses (for more details, see [15, 16]).

It is worth mentioning that the present approach only requires that v_t is an increasing function. Such condition is quite general and holds, for instance, for the following processes, well-known in the literature and widely-used in traffic modeling:

(1) fractional Brownian Motion (FBM), one of the most studied self-similar long-range dependent Gaussian processes, originally proposed in the traffic modeling framework by Norros [2]. It has been shown in [17] that FBM arises as the scaled limit process when the cumulative workload is a superposition of on-off sources with mutually independent heavy-tailed on and/or off periods. In this case,

$$v_t = t^{2H} , \qquad H \in (0, 1) ,$$

and in the teletraffic framework, usually, $H \in (1/2, 1)$, corresponding to processes with long-range dependence;

(2) sum of independent FBMs:

$$v_t = \sum_i t^{2H_i} .$$

The use of this model is also motivated by the fundamental result in [17] in case of heterogeneous on-off sources; and

(3) integrated Ornstein–Uhlenbeck process (IOU):

$$v_t = t + e^{-t} - 1$$

is the Gaussian counterpart of the well-known Anick–Mitra–Sondi fluid model [18], and its relevance in the framework of teletraffic is also discussed in [19].

[1] On one side, several proofs in LDT are based on IS arguments and, on the other side, efficient changes of measure are often related to sample-path LDT results

Note that the analytical expression of the target probability is not known in explicit form for a general Gaussian input (including the considered examples). Indeed, there are only a few asymptotic results available, based on LDT. For instance, in the case of FBM input,

$$\lim_{T\to\infty}\frac{1}{T^{2-2H}}\log\mathbb{P}\left(\forall t\in(0,T]:\ X_t>t\right)$$
$$=\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\forall t\in(0,1]:\ \frac{X_t}{\sqrt{n}}>t\right)$$
$$=-\inf_{f\in\mathscr{B}}\Lambda(f):=-\nu \quad (2)$$

where

$$\mathscr{B}:=\{f\in\mathcal{R}:f(r)>r,\ \forall r\in(0,1]\};$$

Λ denotes the rate function; and $\mathcal{R}$ is the reproducing kernel Hilbert space (RKHS) associated with the distribution of FBM (for more details, see [20]). Moreover, it is known [15] that the constant $\nu\in[1/2,\ c_H^2/2]$, where

$$c_H:=\left[H(2H-1)(2-2H)\times\mathsf{B}\left(H-\frac{1}{2},\ 2-2H\right)\right]^{-1/2}$$

and B is the Beta function. Note that the upper bound for this constant is close to $1/2$ in case when $H>1/2$ (Fig. 1). A further characterization of the most likely path in the set $\mathscr{B}$ has been found in [21], and since an explicit expression for ν is not available, numerical methods to calculate ν have been proposed. The above-mentioned asymptotic result has been generalized in [22] to the case of Gaussian processes with regularly varying at infinity variance, which includes the sum of independent FBMs and IOU as well.

Due to the lack of exact analytical results, simulation is the only available tool for estimating the target probability (1). On the other hand, when $T\to\infty$, the event $\{\forall t\in\mathbb{T}:\ X_t>t\}$ becomes rare and, hence, standard MC requires an unacceptable large number of generated sample paths. Indeed, the key contribution of this work is the application of a variant of the conditional MC method for variance reduction.

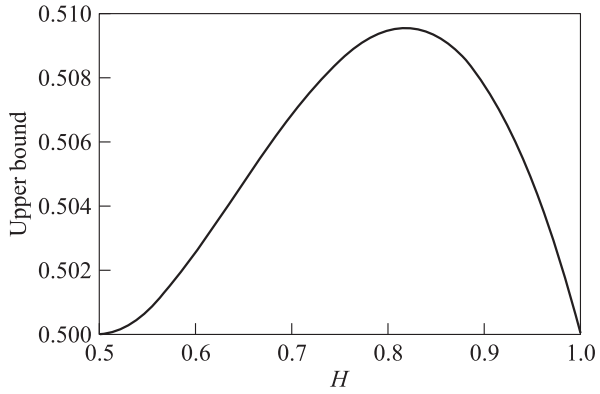

Figure 1 Upper bound for ν

3 Preliminaries on Rare Event Simulation

Let X be a random process. Consider estimating the probability

$$\pi(T):=\mathbb{P}\left(X\in A_T\right)=\mathbb{E}I\left(X\in A_T\right)$$

for some Borel set A_T of the paths of the process X where I denotes the indicator function and T is the so-called parameter of rarity: $\pi_T\to 0$ as $T\to\infty$. To estimate π_T by standard MC simulation, one should generate N replications $X^{(1)},\ldots,X^{(N)}$ of the process X. In the following, there will be considered the estimators of the form

$$\widetilde{\pi}_N(T)=\frac{1}{N}\sum_{n=1}^{N}F_T\left(X^{(n)}\right) \quad (3)$$

for a measurable function F_T. If $F_T(X)=I(X\in A_T)$, one has the crude MC estimator. The relative error (RE) of the estimate $\widetilde{\pi}_N(T)$ is defined as

$$\mathrm{RE}\left(\widetilde{\pi}(T)\right):=\frac{\sqrt{\mathrm{Var}[\widetilde{\pi}_N(T)]}}{\mathbb{E}[\widetilde{\pi}_N(T)]},$$

and for the crude estimate, it behaves as

$$\mathrm{RE}\left(\widetilde{\pi}(T)\right)\sim\frac{1}{\sqrt{\pi(T)N}}\ \text{as}\ \pi(T)\to 0\,.$$

Therefore, the RE of the standard MC estimation is unbounded when the event becomes rare. That is why for the rare event simulation, it is crucially important to define modified estimators in order to reduce variance (and, as a result, RE).

Let us consider the number of sample paths required to obtain some given maximal RE:

$$N_T=\inf\left\{N\in\mathbb{N}:\ \mathrm{RE}\left(\widetilde{\pi}(T)\right)\le\mathrm{RE}_{\max}\right\}$$

and let us introduce the concept of relative efficiency

$$R_T:=\frac{\log\mathbb{E}\left[F_T(X)^2\right]}{\log\mathbb{E}\left[F_T(X)\right]}\,.$$

An estimate (3) is said to be asymptotically optimal [23, 24] with respect to the parameter T if

$$\limsup_{T\to\infty}\frac{1}{T}\log N_T=0\,,$$

i. e., if the corresponding RE increases slower than any exponential function. The latter condition is equivalent to

$$\lim_{T\to\infty} R_T = 2\,.$$

Remark that the limit above is always less or equal 2 since $\mathbb{E}[F_T(X)^2] \geq (\mathbb{E}[F_T(X)])^2$. The proof of asymptotic optimality usually relies on LDT, namely, on Varadhan's lemma [23].

Let us briefly describe the method of simulation by conditioning, known in the literature as *conditional MC* [8]. Denote

$$Z = I\left(X \in A_T\right)$$

and assume that one has an auxiliary random variable (r.v.) Y correlated with Z such that $\mathbb{E}[Z|Y]$ is available in explicit form. Let $Y^{(1)}, \ldots, Y^{(N)}$ be the sample of Y; then, the corresponding unbiased estimator of $\mathbb{E}[Z|Y]$ is defined as

$$\widehat{\pi}_N(T) := \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[Z|Y^{(n)}\right]\,.$$

Note that the variance of this estimator is always less than the variance of the standard MC one since

$$\mathrm{Var}Z = \mathbb{E}\left[\mathrm{Var}[Z|Y]\right] + \mathrm{Var}\left[\mathbb{E}[Z|Y]\right]\,. \qquad (4)$$

Another popular method for variance reduction is IS. The basic idea of IS is the change of the probability measure, so that the target rare event becomes more likely to occur [10]:

$$\pi(T) = \mathbb{E}I(X \in A_T) = \int I\left(x \in A_T\right)\, d\mathbb{P}(x)$$
$$= \int I(x \in A_T)\frac{d\mathbb{P}(x)}{d\tilde{\mathbb{P}}(x)}\, d\tilde{\mathbb{P}}(x) = \tilde{\mathbb{E}}\left[I(X \in A_T)L(X)\right]$$

where $L := d\mathbb{P}(x)/d\tilde{\mathbb{P}}(x)$ is the likelihood ratio and $\tilde{\mathbb{E}}$ means expectation associated with the probability measure $\tilde{\mathbb{P}}$. Hence, the IS estimator is defined as

$$\widehat{\pi}^{\mathrm{IS}}_N(T) := \frac{1}{N}\sum_{n=1}^{N} I\left(X^{(n)} \in A_T\right) L\left(X^{(n)}\right)$$

where $(X^{(1)}, \ldots, X^{(N)})$ are independent and identically-distributed replications generated according to $\tilde{\mathbb{P}}$.

It is well known that the optimal change of measure (zero-variance) requires the knowledge of the probability of interest and, therefore, cannot be practically adopted.

A class of IS estimators (known in the literature as single-twist estimators) can be constructed by shifting the process X with a deterministic path η_t ($\tilde{\mathbb{P}}$ is the law of $\{X_t + \eta_t\}$) in order to make the rare event more likely to occur. In the finite-dimensional case, when X is a centered Gaussian random vector with nondegenerate covariance matrix $\mathbf{\Gamma}$, it is easy to show (see, for example, [25]) that the likelihood ratio is given by

$$L(x) = \exp\left\{-\eta'\mathbf{\Gamma}^{-1}x + \frac{1}{2}\eta'\mathbf{\Gamma}^{-1}\eta\right\}.$$

4 Bridge Monte-Carlo Estimator

The BMC is a special case of the conditional MC method, particularly suitable for the estimation of the rare event probabilities in a queueing system with Gaussian input.

Originally proposed by some of the authors in [14, 26, 27], BMC is based on the idea of expressing the overflow probability as the expectation of a function of the Bridge $Y := \{Y_t\}$ of the Gaussian input process X, i. e., the process obtained by conditioning X to reach a certain level at some prefixed (deterministic) time instant τ:

$$Y_t = X_t - \psi_t X_\tau$$

where ψ_t is expressed via the covariance function as

$$\psi_t := \frac{\Gamma_{t,\tau}}{\Gamma_{\tau,\tau}}\,.$$

Since the variance of the input is an increasing function of t in all models considered, it is easy to see that $\psi_t > 0$ for all $t \in T$. Moreover, note that for any $t \in \mathbb{T}$, Y_t is independent of X_τ since

$$\mathbb{E}\left[X_\tau Y_t\right] = \Gamma_{\tau,t} - \frac{\Gamma_{t,\tau}}{\Gamma_{\tau,\tau}}\,\Gamma_{\tau,\tau} = 0$$

and (X_τ, Y_t) has bivariate normal distribution.

The target probability can be expressed in the following form:

$$\pi(T) = \mathbb{P}\left(\forall\, t \in \mathbb{T}:\ X_t > t\right)$$
$$= \mathbb{P}\left(\forall\, t \in \mathbb{T}:\ X_\tau > \frac{t - Y_t}{\psi_t}\right)$$
$$= \mathbb{P}\left(X_\tau \geq \sup_{t\in\mathbb{T}} \frac{t - Y_t}{\psi_t}\right) = \mathbb{P}\left(X_\tau \geq \overline{Y}\right)$$

where

$$\overline{Y} := \sup_{t\in\mathbb{T}} \frac{t - Y_t}{\psi_t}\,.$$

Observe that random variable $\overline{Y}$ is independent of X_τ. For the sake of simplicity, let us prove this property in the case $\mathbb{T} = \{1, \ldots, T\}$ which is enough for simulation needs. Indeed, the random vector $(X_\tau, Y_1, \ldots, Y_T)$

has multivariate normal distribution and, moreover, as it was shown above, X_τ is independent of Y_i, $i = 1, \ldots, T$; hence, X_τ is independent of the vector $(Y_1, \ldots, Y_T)$ (due to the properties of the multivariate normal distribution, see [28] for more details) and, as a consequence, of any function of its components.

Having in mind that $X_t =_d \sqrt{\Gamma_{t,\,t}}\, N(0,\,1)$, the considered probability can be rewritten as follows:

$$\pi(T) = \mathbb{P}\left(X_\tau \geq \overline{Y}\right) = \int\limits_R \mathbb{P}(X_\tau \geq u)\mathbb{P}\left(\overline{Y} \in du\right) = \mathbb{E}\left[\Phi\left(\frac{\overline{Y}}{\sqrt{\Gamma_{\tau,\,\tau}}}\right)\right]$$

where independence between $\overline{Y}$ and X_τ is used and Φ denotes the tail distribution of standard normal variable, that is,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_x^\infty e^{-y^2/2} dy \ .$$

Hence, given a sample $\{\overline{Y}^{(n)},\ n = 1, \ldots, N\}$ of $\overline{Y}$, the estimator of $\pi(T)$ is defined as follows:

$$\widehat{\pi}_N^{\mathrm{BMC}} := \frac{1}{N} \sum_{n=1}^{N} \Phi\left(\frac{\overline{Y}^{(n)}}{\sqrt{\Gamma_{\tau,\tau}}}\right) .$$

Note that

$$\Phi\left(\frac{\overline{Y}}{\sqrt{\Gamma_{\tau,\,\tau}}}\right) = \mathbb{E}\left[I(X_\tau > \overline{Y})|\overline{Y}\right] \ ;$$

therefore, the BMC approach is actually a special case of the conditional MC method. By (4), $\mathrm{Var}\, Z \geq \mathrm{Var}[\mathbb{E}[Z|\overline{Y}]]$; so, one can expect that the BMC estimator implies variance reduction (with regard to crude MC simulation) in the estimation of the target probability $\pi(T)$.

5 Simulation Results

In this section, through simulation results, the accuracy of the BMC estimator and the dependence of its performance on different parameters will be pointed out. For sake of brevity, only the results for FBM input considering $N = 10000$ replications (unless otherwise stated) will be presented.

Figure 2 shows the dependence of the target probability on the interval duration T in case of $H = 0.8$, a typical value of the Hurst parameter for traffic data. The probability $\pi(T)$ exhibits an exponential decay in agreement with the known LDT asymptotic results (see formula (2)). To better understand the practical applicability of such limits, in Fig. 3, the ratio between BMC estimates and (2) is reported for different values of T.

In order to verify the goodness of the present estimator, the dependence of the RE on the parameters T has been considered. Figure 4 highlights that the RE grows slowly and for probabilities of the order of 10^{-11}, it is still less than 18%, as can be easily verified by comparing the values in Figs. 2 and 4.

The goodness of the present method is also confirmed by the table, where BMC is compared to single twist (with a constant linear drift chosen by minimizing the variance of the estimator) IS in terms of RE: for all

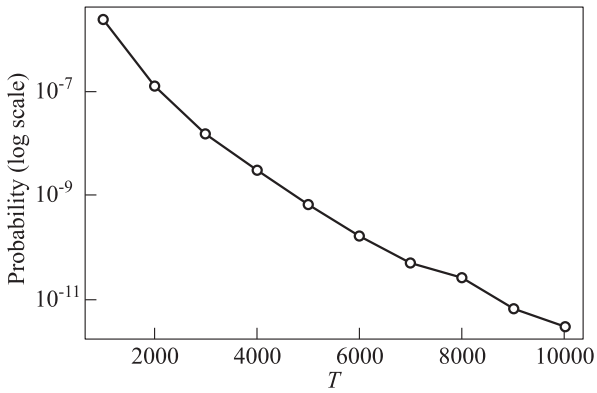

Figure 2 Dependence of π on parameter T

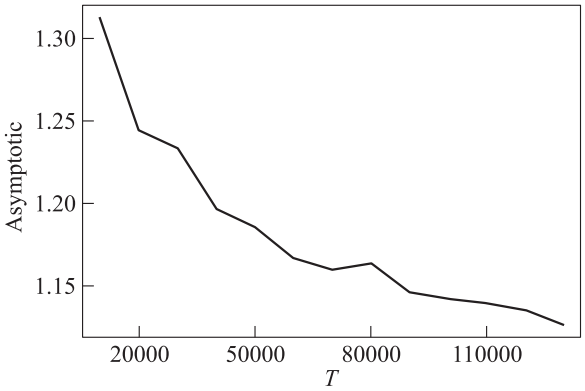

Figure 3 Comparison with LDT asymptotic results

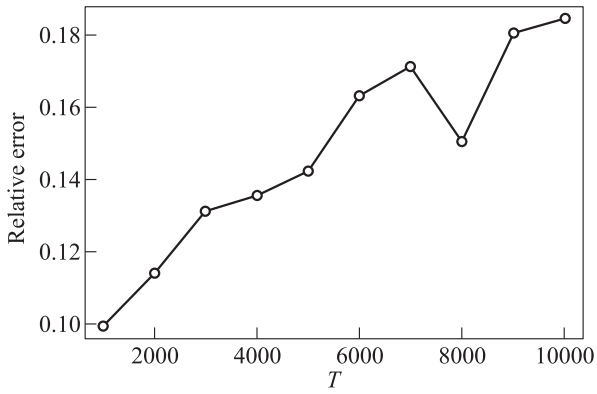

Figure 4 Dependence of the RE on T

Relative errors for BMC and IS

T	BMC	IS
200	0.0691	0.1483
400	0.0779	0.1850
600	0.0941	0.1755
800	0.0894	0.1772
1000	0.1039	0.2275
1200	0.1044	0.2084
1400	0.1101	0.1980
1600	0.1104	0.2462
1800	0.1150	0.2182
2000	0.1177	0.2696

the considered values of T, BMC reduces the RE by a factor around 2.

To better understand the asymptotic properties of the estimator (at least heuristically), in Figs. 5 and 6, the behavior of N_T and R_T is shown: the required number of sample paths (for a fixed value of the RE) grows very slowly (at least in logarithmic scale) and the relative efficiency is above 1.9 for $T > 20\,000$ (and gets closer to 2 for higher values of T).

In all previous simulation sets, the conditioning point τ in the BMC algorithm has been assumed equal

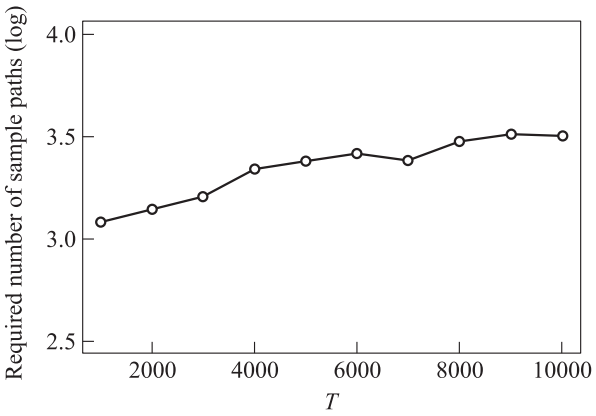

Figure 5 Dependence of N_T on T for $\mathrm{RE}_{\max} = 0.1$

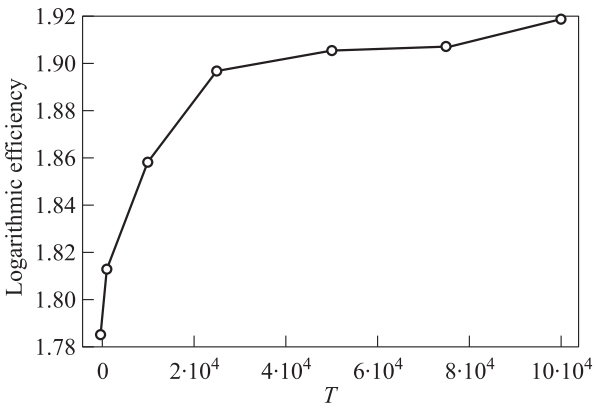

Figure 6 Dependence of R_T on T for $N = 10000$

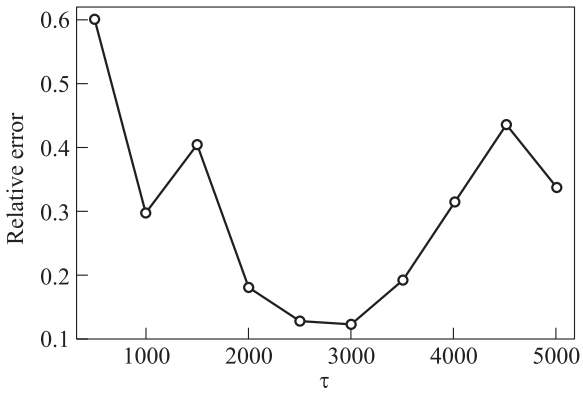

Figure 7 Dependence of RE on τ for $T = 3000$

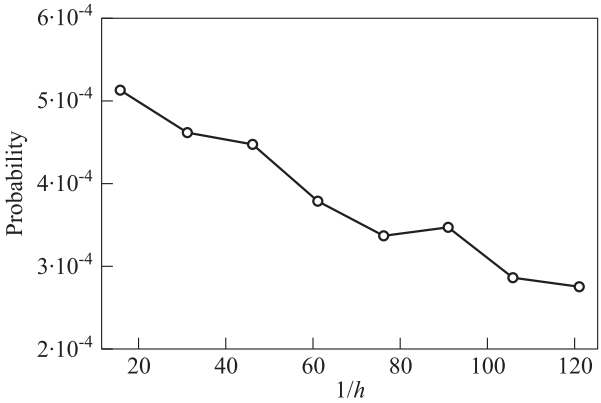

Figure 8 Effect of the discretization step h

to the duration of the interval. The correctness of such choice is confirmed for $T = 3000$ by Fig. 7 in which an absolute minimum of the RE can be identified in a neighborhood of T.

Finally, in Fig. 8, the effect of the discretization step (simulations always involve finite-size vectors and not continuous-time processes!) on the estimated probability is highlighted. In more detail, $T = 100$ with discretisation step h (in the previous simulations, $h = 1$) has been considered. This means that each FBM sample path consists of T/h points: apart some oscillations of the estimated value (the confidence intervals should also be taken into account!), as expected, the target probability decreases when sampling is more dense.

6 Concluding Remarks

In this paper, the estimation of the busy period duration in Gaussian queues was considered with focus on the upper tail of the distribution. To address the issues related to the simulation of such rare events, the authors considered a special case of conditional MC estimator based on bridge processes. In more detail, the BMC

approach exploits the Gaussian nature of the input process (independence is equivalent to uncorrelatedness) and relies on the properties of bridges to write down the target probability as the conditional one.

To study the properties of the proposed estimator, several simulation experiments have been carried out focusing on FBM sample paths, although the methods are applicable to any Gaussian process with increasing variance. In the experimental analysis, different values of the relevant parameters (duration of the interval, choice of the conditioning point, and discretization step) have been considered and the asymptotic properties of the estimator (in terms of relative efficiency and duration of the simulation for a given precision of the estimates) have been investigated. Finally, it is worth mentioning that the relative error is halved with respect to single twist IS, highlighting the efficiency of BMC over well-known rare event simulation techniques.

Acknowledgments

This work is supported by the Russian Foundation for Basic Research, projects 15–07–02341, 15–07–02354, and 15–07–02360 and also by the Program of Strategic Development of Petrozavodsk State University.

References

1. Leland, W. E., M. S. Taqqu, W. Willinger, and D. V. Wilson. 1994. On the self-similar nature of Ethernet traffic (extended version). *IEEE ACM Trans. Network*. 2(1):1–15.
2. Norros, I. 1995. On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE J. Sel. Area. Comm*. 13(6):953–962.
3. Mandjes, M. 2007. *Large deviations of Gaussian queues*. Chichester: Wiley. 340 p.
4. Erramilli, A., O. Narayan, and W. Willinger. 1996. Experimental queueing analysis with long-range dependent packet traffic. *IEEE ACM Trans. Network*. 4(2):209–223.
5. Samorodnitsky, G. 2007. Long range dependence. *Found. Trends® Stochastic Syst.* 1(3):163–257. doi: 10.1561/0900000004.
6. Allman, M., V. Paxson, and E. Blanton. 2009. TCP Congestion Control. RFC 5681 (Draft Standard).
7. Kouvatsos, D. D. 2000. *Performance evaluation and applications of ATM networks*. Kluwer Academic. 472 p.
8. Ross, S. M. 2006. *Simulation*. Elsevier. 314 p.
9. Ganesh, A., N. O'Connell, and D. Wischik. 2004. *Big queues*. Lecture notes in mathematics ser. Springer. 260 p.
10. Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Model. Comput. Simul.* 5(1):43–85.
11. Glasserman, P., and Y. Wang. 1997. Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab*. 7(3):731–746.
12. Lukashenko, O. V., E. V. Morozov, and M. Pagano. 2016. On Conditional Monte Carlo estimation of busy period probabilities in Gaussian queues. *Comm. Com. Inf. Sc*. 601:280–288. doi: 10.1007/978-3-319-30843-2_29.
13. Lukashenko, O. V., E. V. Morozov, and M. Pagano. 2016. On the use of a bridge process in a Conditional Monte Carlo simulation of Gaussian queues. *Comm. Com. Inf. Sc*. 638:207–220. doi: 10.1007/978-3-319-44615-8_18.
14. Giordano, S., M. Gubinelli, and M. Pagano. 2005. Bridge Monte-Carlo: A novel approach to rare events of Gaussian processes. *5th St. Petersburg Workshop on Simulation Proceedings*. 281–286.
15. Norros, I. 1999. Busy periods of fractional Brownian storage: A large deviations approach. *Adv. Perf. Anal*. 2:1–19.
16. Mandjes, M., I. Norros, and P. Glynn. 2009. On convergence to stationarity of fractional Brownian storage. *Ann. Appl. Probab*. 19:1385–1403.
17. Taqqu, M. S., W. Willinger, and R. Sherman. 1997. Proof of a fundamental result in self-similar traffic modeling. *Comput. Commun. Rev.* 27:5–23.
18. Addie, R., P. Mannersalo, and I. Norros. 2002. Most probable paths and performance formulae for buffers with Gaussian input traffic. *Eur. Trans. Telecommun*. 13:183–196.
19. Kulkarni, V., and T. Rolski. 1994. Fluid model driven by an Ornstein–Uhlenbeck process. *Probab. Eng. Inform. Sc*. 8:403–417.
20. *Deuschel, J. D., and D. W. Stroock*. 1989. *Large deviations*. Academic Press. 330 p.
21. Mandjes, M., P. Mannersalo, I. Norros, and M. van Uitert. 2006. Large deviations of infinite intersections of events in Gaussian processes. *Stoch. Proc. Appl*. 116:1269–1293.
22. Dieker, A. B. 2005. Conditional limit theorems for queues with Gaussian input: A weak convergence approach. *Stoch. Proc. Appl*. 115(5):849–873.
23. Dieker, A. B., and M. Mandjes. 2005. On asymptotically efficient simulation of large deviation probabilities. *Adv. Appl. Probab*. 37:539–552.
24. Dieker, A. B., and M. Mandjes. 2006. Fast simulation of overflow probabilities in a queue with Gaussian input. *ACM Trans. Model. Comput. Simul*. 16:119–151.
25. Asmussen, S., ans P. Glynn. 2007. *Stochastic simulation: Algorithms and analysis*. Springer. 476 p.
26. Giordano, S., M. Gubinelli, and M. Pagano. 2007. Rare events of Gaussian processes: A performance comparison between Bridge Monte-Carlo and Importance Sampling. *Next generation teletraffic and wired/wireless advanced networking*. Eds. Y. Koucheryavy, J. Harju, and A. Sayenko. Computer communication networks and telecommunications ser. Berlin–Heidelberg: Springer. 4712:269–280.
27. Lukashenko, O. V., E. V. Morozov, and M. Pagano. 2012. Performance analysis of Bridge Monte-Carlo estimator. *Transactions of KarRC RAS* 5:54–60.
28. Gut, A. 2009. *An intermediate course in probability*. Springer. 304 p.

Received February 16, 2017

Contributors

L ukashenko Oleg V. (b. 1986) — Candidate of Science (PhD) in physics and mathematics, scientist, Institute of Applied Mathematical Research of Karelian Research Centre of the Russian Academy of Sciences, 11 Pushkinskaya Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; lecturer, Petrozavodsk State University, 33 Lenin Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; lukashenko@krc.karelia.ru

Morozov Evsei V. (b. 1947) — Doctor of Science in physics and mathematics, professor, leading scientist, Institute of Applied Mathematical Research of Karelian Research Centre of the Russian Academy of Sciences, 11 Pushkinskaya Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; professor, Petrozavodsk State University, 33 Lenin Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; emorozov@karelia.ru

Pagano Michele (b. 1968) — PhD in Information Engineering, associate professor, University of Pisa, 43 Lungarno Pacinotti, Pisa 56126, Italy; m.pagano@iet.unipi.it

ОБ ЭФФЕКТИВНОСТИ ОЦЕНКИ МОНТЕ КАРЛО НА ОСНОВЕ ГАУССОВСКОГО МОСТА*

О. В. Лукашенко[1,2], Е. В. Морозов[1,2], М. Пагано[3]

[1]Институт прикладных математических исследований Карельского научного центра Российской академии наук
[2]Петрозаводский государственный университет
[3]Университет г. Пиза, Италия

Аннотация: Наличие долговременной зависимости в современных сетях передачи данных приводит к тому, что объем передаваемого трафика может быть большим на протяжении значительного периода времени. Это, в свою очередь, влечет перегрузку систем на протяжении длительного периода времени. В данной работе рассматривается задача оценки вероятности занятости системы обслуживания с гауссовским входным потоком в течение некоторого заданного периода T. При больших значениях T интересующее нас событие является редким, и для оценки его вероятности с приемлемой точностью необходимо использовать специальные методы понижения дисперсии оценки. В статье рассмотрен частный случай условного метода Монте Карло, который заключается в том, что искомая вероятность может быть выражена как математическое ожидание некоторой функции от так называемого гауссовского моста. Исследована эффективность предложенной процедуры, а также влияние шага дискретизации на свойство получаемой оценки.

Ключевые слова: гауссовские процессы; условный метод Монте Карло; процесс моста; редкие события; уменьшение дисперсии

DOI: 10.14357/19922264170202

Литература

1. *Leland W. E., Taqqu M. S., Willinger W., Wilson D. V.* On the self-similar nature of Ethernet traffic (extended version) // IEEE ACM Trans. Network., 1994. Vol. 2. No. 1. P. 1–15.
2. *Norros I.* On the use of fractional Brownian motion in the theory of connectionless networks // IEEE J. Sel. Area. Comm., 1995. Vol. 13. No. 6. P. 953–962.
3. *Mandjes M.* Large deviations of Gaussian queues. — Chichester: Wiley, 2007. 340 p.
4. *Erramilli A., Narayan O., Willinger W.* Experimental queueing analysis with long-range dependent packet traffic // IEEE ACM Trans. Network., 1996. Vol. 4. No. 2. P. 209–223.
5. *Samorodnitsky G.* Long range dependence // Found. Trends® Stochastic Syst., 2007. Vol. 1. No. 3. P. 163–257. doi: 10.1561/0900000004.
6. *Allman M., Paxson V., Blanton E.* TCP congestion control. RFC 5681 (Draft Standard), 2009.
7. *Kouvatsos D. D.* Performance evaluation and applications of ATM networks. — Kluwer Academic, 2000. 472 p.
8. *Ross S. M.* Simulation. — Elsevier, 2006. 314 p.
9. *Ganesh A., O'Connell N., Wischik D.* Big queues. — Lecture notes in mathematics ser. — Springer, 2004. 260 p.

*Работа поддержана грантами РФФИ №№ 15–07–02341, 15–07–02354 и 15–07–02360, а также программой стратегического развития Петрозаводского государственного университа.

10. *Heidelberger P.* Fast simulation of rare events in queueing and reliability models // ACM Trans. Model. Comput. Simul., 1995. Vol. 5. No. 1. P. 43–85.
11. *Glasserman P., Wang Y.* Counterexamples in importance sampling for large deviations probabilities // Ann. Appl. Probab., 1997. Vol. 7. No. 3. P. 731–746.
12. *Lukashenko O. V., Morozov E. V., Pagano M.* On Conditional Monte Carlo estimation of busy period probabilities in Gaussian queues // Comm. Com. Inf. Sc., 2016. Vol. 601. P. 280–288. doi: 10.1007/978-3-319-30843-2_29.
13. *Lukashenko O. V., Morozov E. V., Pagano M.* On the use of a bridge process in a Conditional Monte Carlo simulation of Gaussian queues // Comm. Com. Inf. Sc., 2016. Vol. 638. P. 207–220. doi: 10.1007/978-3-319-44615-8_18.
14. *Giordano S., Gubinelli M., Pagano M.* Bridge Monte-Carlo: A novel approach to rare events of Gaussian processes // 5th St. Petersburg Workshop on Simulation Proceedings, 2005. P. 281–286.
15. *Norros I.* Busy periods of fractional Brownian storage: A large deviations approach // Adv. Perf. Anal., 1999. Vol. 2. P. 1–19.
16. *Mandjes M., Norros I., Glynn P.* On convergence to stationarity of fractional Brownian storage // Ann. Appl. Probab., 2009. Vol. 19. P. 1385–1403.
17. *Taqqu M. S., Willinger W., Sherman R.* Proof of a fundamental result in self-similar traffic modeling // Comput. Commun. Rev., 1997. Vol. 27. P. 5–23.
18. *Addie R., Mannersalo P., Norros I.* Most probable paths and performance formulae for buffers with Gaussian input traffic // Eur. Trans. Telecommun., 2002. Vol. 13. P. 183–196.
19. *Kulkarni V., Rolski T.* Fluid model driven by an Ornstein–Uhlenbeck process // Probab. Eng. Inform. Sc., 1994. Vol. 8. P. 403–417.
20. *Deuschel J. D, Stroock D. W.* Large deviations. — Academic Press, 1989. 330 p.
21. *Mandjes M., Mannersalo P., Norros I., van Uitert M.* Large deviations of infinite intersections of events in Gaussian processes // Stoch. Proc. Appl., 2006. Vol. 116. P. 1269–1293.
22. *Dieker A. B.* Conditional limit theorems for queues with Gaussian input: A weak convergence approach // Stoch. Proc. Appl., 2005. Vol. 115. No. 5. P. 849–873.
23. *Dieker A. B., Mandjes M.* On asymptotically efficient simulation of large deviation probabilities // Adv. Appl. Probab., 2005. Vol. 37. P. 539–552.
24. *Dieker A. B., Mandjes M.* Fast simulation of overflow probabilities in a queue with Gaussian input // ACM Trans. Model. Comput. Simul., 2006. Vol. 16. P. 119–151.
25. *Asmussen S., Glynn P.* Stochastic simulation: Algorithms and analysis. — Springer, 2007. 476 p.
26. *Giordano S., Gubinelli M., Pagano M.* Rare events of Gaussian processes: A performance comparison between Bridge Monte-Carlo and importance sampling // Next generation teletraffic and wired/wireless advanced networking / Eds. Y. Koucheryavy, J. Harju, and A. Sayenko. — Computer communication networks and telecommunications ser. — Berlin–Heidelberg: Springer, 2007. Vol. 4712. P. 269–280.
27. *Lukashenko O. V., Morozov E. V., Pagano M.* Performance analysis of Bridge Monte-Carlo estimator // Труды Карельского НЦ РАН, 2012. Т. 5. С. 54–60.
28. *Gut A.* An intermediate course in probability. — Springer, 2009. 304 p.

Поступила в редакцию 16.02.2017

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 25–32

МАКСИМИЗАЦИЯ СРЕДНЕГО СТАЦИОНАРНОГО ДОХОДА СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ ТИПА $M/G/1$*

Я. М. Агаларов[1]

Аннотация: Рассматривается задача оптимизации порогового значения длины очереди системы $M/G/1$ в смысле максимизации предельного дохода, получаемого системой в единицу времени. Доход определяется платой, получаемой за обслуживание заявок, затратами на техническое обслуживание прибора и штрафами за задержку заявок в очереди, за отказ заявке в обслуживании, за простой системы. Сформулированы достаточные условия существования конечного порогового значения для системы $M/G/1$, доказаны утверждения о необходимых и достаточных условиях оптимальности порогового значения и о существовании конечного оптимального порога. Предложен алгоритм расчета оптимального порогового значения и соответствующего значения максимального дохода. Приведены результаты вычислительных экспериментов, иллюстрирующие работу предложенного алгоритма.

Ключевые слова: система массового обслуживания; пороговое управление; оптимизация

DOI: 10.14357/19922264170203

1 Введение

В данной работе проводится дальнейшее исследование проблемы оптимального управления очередью системы $M/G/1$, сформулированной в работе [1] в виде задачи максимизации предельного дохода системы$M/G/1$ в единицу времени на множестве пороговых стратегий. Доход, как и в работе [1], определяется платой, получаемой за обслуживание заявок, затратами на техническое обслуживание прибора и штрафами за задержку заявок в очереди, за отказ заявке в обслуживании, за простой системы. Аналогичная постановка задачи для одноканальных СМО типа $M/G/1$ и $G/M/1$ рассмотрена автором данной статьи также в работах [2, 3]. Из других работ отечественных авторов близки по постановке задачи работы [4, 5], в которых данная задача сформулирована в терминах теории управляемых цепей Маркова для системы массового обслуживания (СМО) типа $M/G/1$ и CBSMAP/$M/n/r$ и предлагается численный метод ее решения.

В работе [1] получены оценка снизу для оптимального порогового значения и алгоритм ее вычисления методом последовательного приближения. В качестве оценки предлагается решение вспомогательной оптимизационной задачи, сформулированной как задача поиска порогового значения длины очереди (оптимальной стратегии управления очередью) системы $M/G/1$, максимизирующего предельный доход системы, усредненный по числу обслуженных заявок. В работе [1] сформулированы необходимые и достаточные условия оптимальности порогового значения для вспомогательной задачи и предложен алгоритм гарантированного поиска решения при условии выполнения достаточных условий. В данной работе аналогичные результаты получены для задачи максимизации предельного дохода системы $M/G/1$ в единицу времени и доказано, что для этой системы достаточные условия существования оптимального порога выполняются всегда. Предложен алгоритм последовательного приближения, который для гарантированного поиска оптимального порога требует объема вычислений порядка двоичного логарифма значения этого порога.

2 Постановка задачи

Рассматривается СМО типа $M/G/1$ с накопителем бесконечной емкости и одним прибором обслуживания, на которую поступает пуассоновский поток заявок с интенсивностью $\lambda > 0$, а время обслуживания каждой заявки распределено по произвольному закону $H(t)$. Поступившая заявка допускается в накопитель системы (занимает любое свободное место в накопителе), если в момент ее поступления число занятых мест в накопителе меньше k, где $k > 0$ — некоторое заданное значение (тривиальный случай $k = 0$ не рассматривается), а в противном случае она отклоняется (не допус-

*Работа выполнена при частичной финансовой поддержке РФФИ (проект 15-07-03406).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, agglar@yandex.ru

кается в систему). В дальнейшем обозначение k будем называть порогом, а его значение пороговым значением. Если заявка допущена в накопитель, она занимает любое свободное место в накопителе и обслуживается на приборе в порядке поступления. Заявка покидает систему только при завершении обслуживания, освободив одновременно прибор и накопитель, а на освободившийся прибор поступает очередная заявка из накопителя (если таковая есть). Система получает доход, который определяется следующими составляющими:

- $C_0 v \geq 0$ — плата, получаемая системой, если поступившая заявка обслужена системой, где v — время занятия прибора заявкой;
- $C_1 \geq 0$ — величина штрафа, если поступившая заявка отклонена;
- $C_2 \geq 0$ — величина штрафа за единицу времени ожидания заявки в очереди к прибору;
- $C_3 \geq 0$ — величина штрафа за единицу времени простоя прибора;
- $C_4 \geq 0$ — затраты системы в единицу времени на техническое обслуживание системы.

Будем считать, что плату за обслуживание система получает в момент завершения обслуживания каждой заявки в зависимости от длины заявки (длительности занятия прибора).

Отметим, что процесс обслуживания заявок в данной системе описывается цепью Маркова, где переходы цепи определяются моментами окончания обслуживания и состояние системы есть число заявок, остающихся в ней в момент ухода с прибора обслуженной заявки (см., например, [6, 7]). Отметим также, что при пороге k указанная цепь Маркова имеет один положительный возвратный класс состояний $i = 0, \ldots, k-1$.

Введем обозначения:

- π_i^k, $0 \leq i \leq k-1$, — стационарное распределение вероятностей цепи при пороге k (π_i^k — вероятность того, что цепь находится в состоянии i);
- g^k — значение суммарного предельного дохода, усредненного по числу обслуженных заявок;
- q_i^k — средний доход, получаемый системой в состоянии i при пороге k, $i \geq 0$;
- $\overline{v} = \int_0^\infty t\, dH(t)$ — среднее время пребывания системы в состоянии i, $0 < \overline{v} < \infty$;
- $\rho = \lambda \overline{v}$ — входная нагрузка.

Значение предельного дохода, усредненного по числу обслуженных заявок, при пороге k равно [1]:

$$g^k = \sum_{i=0}^{k-1} \pi_i^k q_i^k \,. \tag{1}$$

Среднее значение предельного дохода системы в единицу времени при пороге k составляет [1]:

$$Q^k = \lambda \left(1 - \theta_k^k\right) g^k \,, \tag{2}$$

где θ_k^k — вероятность того, что поступившая заявка не будет допущена в систему.

Ставится задача: найти оптимальное пороговое значение $k^* > 0$, такое что

$$\max_{k>0} Q^k = Q^{k^*} \,. \tag{3}$$

Ниже всюду будем считать, что $k^0 > 0$ и $k^* > 0$ — оптимальные пороговые значения в случае функций (1) и (2) соответственно.

3 Метод решения

Приводимый ниже метод решения подробно описан в работе [1] применительно к задаче максимизации функции g^k. Для наглядности изложения приведем кратко основные формулы и утверждения из указанной работы, на которые ниже в тексте сделаны ссылки.

Для стационарного распределения вероятностей состояний системы при пороге k получены рекуррентные формулы (см. (4) и (6) в [1]):

$$\pi_j^k = \pi_0^k R_j \,, \quad j = 0, \ldots, k-1 \,, \tag{4}$$

где

$$\pi_0^k = \left(\sum_{i=0}^{k-1} R_i \right)^{-1} \,.$$

Здесь

$$R_0 = 1 \,; \quad R_1 = \frac{1 - r_0}{r_0} \,;$$

$$R_{i+1} = \frac{1}{r_0} \left(R_i - r_i - \sum_{j=1}^{i} R_j r_{i-j+1} \right) \,, \quad i = 1, \ldots, k-2 \,,$$

где

$$r_l = \int_0^\infty \frac{(\lambda v)^l}{l!}\, e^{-\lambda v}\, dH(v) \text{ при } l \geq 0 \,.$$

Для величины среднего дохода системы, получаемого за время пребывания в состоянии $0 \leq i \leq \leq k-1$, верна формула (см. (10) в [1]):

$$q_i^k = C_0\overline{v} - C_1 \sum_{m=k-i+1}^{\infty} (m-k+i)\, r_m - \\ -\frac{C_2}{\lambda}\left[\frac{1}{2}\sum_{m=2}^{k-i+1}(m-1)mr_m + (k-i)\sum_{m=k-i+2}^{\infty} mr_m - \right. \\ \left. -\frac{1}{2}(k-i)(k-i+1)\sum_{m=k-i+2}^{\infty} r_m\right] - \\ - C_2(i-1)\overline{v} - C_4\overline{v}, \;\; 1 \le i \le k-1; \\ q_0^k = q_1^k - \frac{C_3+C_4}{\lambda}. \quad (5)$$

Данная формула и формула, приведенная для q_i^k в [1], отличаются только слагаемыми с параметром C_0. Это является следствием того, что в рассматриваемой СМО плата за обслуживание заявки производится после завершения обслуживания и в размере $C_0 v$.

Справедливы следующие эквивалентные формулы (см. (16) в [1]):

$$R_k = \frac{1}{r_0}\left(\sum_{j=1}^{k-1} R_j \sum_{i=k-j+1}^{\infty} r_i + R_0\sum_{i=k}^{\infty} r_i\right); \\ \pi_k^{k+1} = \frac{1}{r_0}\left(\sum_{j=1}^{k-1}\pi_j^{k+1}\sum_{i=k-j+1}^{\infty} r_i + \pi_0^{k+1}\sum_{i=k}^{\infty} r_i\right). \quad (6)$$

С учетом изменений, внесенных в формулу (5), равенство (13) из работы [1] принимает вид:

$$g^k - g^{k+1} = \\ = \pi_k^{k+1}\left\{g^k - \frac{1-\pi_k^{k+1}}{\pi_k^{k+1}}\sum_{j=1}^{k-1}\pi_j^k\left\{C_1\sum_{m=k-j+1}^{\infty} r_m - \right.\right. \\ \left. - \frac{C_2}{\lambda}\sum_{m=k-j+1}^{\infty}[m-(k-j+1)]\, r_m\right\} - \\ - \frac{1-\pi_k^{k+1}}{\pi_k^{k+1}}\pi_0^k\left[C_1\sum_{m=k}^{\infty} r_m - \frac{C_2}{\lambda}\sum_{m=k}^{\infty}(m-k)r_m\right] - \\ \left. - q_k^{k+1}\right\}. \quad (7)$$

Далее, использовав обозначения (17) и (20) из работы [1], равенство (7) приведем к виду (см. (18) в [1]):

$$g^k - g^{k+1} = \pi_k^{k+1}\left[g^k - G(k)\right]. \quad (8)$$

Здесь

$$G(k) = C_0\overline{v} - C_1(\rho-1) - C_4\overline{v} - \frac{C_2}{\lambda}\left[r_0 F(k) - 1 + k\rho\right],$$

где

$$F(k) = \left(\sum_{j=1}^{k-1} R_j \sum_{m=k-j+1}^{\infty}[m-(k-j)]r_m + \right. \\ \left. + R_0\sum_{m=k}^{\infty}(m-k+1)r_m\right) \Bigg/ \left(\sum_{j=1}^{k-1} R_j \sum_{i=k-j+1}^{\infty} r_i + \right. \\ \left. + R_0\sum_{i=k}^{\infty} r_i\right) = \\ = \left(\sum_{j=1}^{k-1}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}[m-(k-j)]r_m + \right. \\ \left. + \pi_0^{k+1}\sum_{m=k}^{\infty}(m-k+1)r_m\right) \Bigg/ \left(\sum_{j=1}^{k-1}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty} r_m + \right. \\ \left. + \pi_0^{k+1}\sum_{m=k}^{\infty} r_m\right).$$

В [1] доказаны следующие утверждения.

Лемма 1 [1]. *Справедливы равенства*

$$\left.\begin{aligned} q_j^{k+1} &= q_j^k + C_1\sum_{m=k-j+1}^{\infty} r_m - \\ &- \frac{C_2}{\lambda}\left[\sum_{m=k-j+1}^{\infty} mr_m - (k-j+1)\sum_{m=k-j+1}^{\infty} r_m\right], \\ &\qquad j = 1,\ldots,k-1, \\ &\qquad q_0^{k+1} - q_0^k = q_1^{k+1} - q_1^k; \\ \pi_j^{k+1} &= \left(1-\pi_k^{k+1}\right)\pi_j^k, \;\; j = 0,\ldots,k-1. \end{aligned}\right\} \quad (9)$$

Теорема 1 [1]. *Пусть* $F(k) - F(k+1) < \lambda\overline{v}$, $0 < \lambda < \infty$, $0 < \overline{v} < \infty$, $k > 0$. *Тогда при любых значениях параметров* $0 \le C_i < \infty$, $i = 0,\ldots,4$, $0 < C_2 < \infty$ *существует стационарная стратегия* $0 < k^0 < \infty$. *При этом если* $g^1 \ge G(1)$, *то* $k^0 = 1$, *а если* $C_2 = 0$ *и* $g^1 < G(1)$, *то* $k^0 = \infty$.

Следствие 1 [1]. Пусть $F(k)$ удовлетворяет условию теоремы 1. Стратегия k^0 удовлетворяет условию $\max\limits_{k>0} g^k = g^{k^0}$ тогда и только тогда, когда k^0 удовлетворяет одному из трех условий:

(1) $G(1) \le g^1$, $k^0 = 1$;

(2) $g(1) > g^1$, $k^0 = \min\{k:\; G(k) \le g^k\}$;

(3) $g^{k^0-1} < g^{k^0}$, $g^{k^0+1} > g^{k^0}$, $1 < k^0 < \infty$.

Докажем ниже, что аналогичные утверждения справедливы и для задачи (3). Из (5), использовав соотношение $q_{i+1}^{k+1} = q_i^k - C_2\overline{v}$, находим:

$$q_k^{k+1} = q_1^2 - C_2(k-1)\overline{v} = \\ = C_0\overline{v} - C_1(\lambda\overline{v} - r_1 - 1 + r_0 + r_1) -$$

$$-\frac{C_2}{\lambda}\left[r_2+(\lambda\overline{v}-r_1-2r_2)-(1-r_0-r_1-r_2)\right]-$$
$$-c_4\overline{v}-C_2(k-1)\overline{v}=$$
$$=C_0\overline{v}-C_1\left(\rho-1+r_0\right)+\frac{C_2}{\lambda}\left(1-r_0\right)-C_4\overline{v}-C_2k\overline{v}\,.$$

Заменив в правой части равенства (7) π_k^{k+1} и π_j^k, $j=0,\ldots,k$, на их выражения в (6) и (4) и q_k^{k+1} на полученное выше для него выражение, получим после преобразований:

$$g_k-g^{k+1}=\pi_k^{k+1}\left\{g^k-\right.$$
$$-C_1\frac{1-\pi_k^{k+1}}{\pi_k^{k+1}}\left[\sum_{j=1}^{k-1}\pi_j^k\sum_{m=k-j+1}^{\infty}r_m+\pi_0^k\sum_{m=k}^{\infty}r_m\right]+$$
$$+\frac{C_2}{\lambda}\frac{1}{\pi_k^{k+1}}\left(\sum_{j=1}^{k}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}[m-(k-j+1)]r_m+\right.$$
$$\left.+\pi_0^{k+1}\sum_{m=k}^{\infty}(m-k)r_m\right)-$$
$$\left.-\sum_{m=1}^{\infty}(m-1)r_m-q_k^{k+1}\right\}=\pi_k^{k+1}\left\{g^k-C_1r_0+\right.$$
$$+\frac{C_2}{\lambda}\frac{1}{\pi_l^{k+1}}\left(\sum_{j=1}^{k}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}[m-(k-j+1)]r_m+\right.$$
$$\left.+\pi_0^{k+1}\sum_{m=k}^{\infty}(m-k)r_m\right)-$$
$$-\frac{C_2}{\lambda}\left(\rho-1+r_0\right)-\left[C_0\overline{v}-C_1\left(\rho-1+r_0\right)+\right.$$
$$\left.\left.+\frac{C_2}{\lambda}\left(1-r_0\right)-C_4\overline{v}-C_2k\overline{v}\right]\right\}=$$
$$=\pi_k^{k+1}\left\{g^k-C_0\overline{v}+C_1(\rho-1)+C_4\overline{v}+\right.$$
$$+\frac{C_2}{\lambda}\frac{1}{\pi_k^{k+1}}\left[\left(\sum_{j=1}^{k}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}[m-(k-j+1)]r_m+\right.\right.$$
$$\left.\left.\left.+\pi_0^{k+1}\sum_{m=k}^{\infty}(m-k)r_m\right)+(k-1)\rho\right]\right\}.$$

Обозначим

$$\tilde{F}(k)=\frac{1}{\pi_k^{k+1}}\left(\sum_{j=1}^{k}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}[m-(k-j+1)]r_m+\right.$$
$$\left.+\pi_0^{k+1}\sum_{m=k}^{\infty}(m-k)r_m\right)=$$
$$=r_0\left(\sum_{j=1}^{k-1}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}[m-(k-j)]r_m+\right.$$
$$\left.+\pi_0^{k+1}\sum_{m=k}^{\infty}(m-k+1)r_m\right)\Bigg/\left(\sum_{j=1}^{k-1}\pi_j^{k+1}\sum_{i=k-j+1}^{\infty}r_i+\right.$$
$$\left.+\pi_0^{k+1}\sum_{i=k}^{\infty}r_i\right)+$$
$$+\sum_{m=1}^{\infty}(m-1)r_m-r_0=r_0F(k)+\rho-1\,.$$

Тогда, как видно из последнего соотношения для разности g^k-g^{k+1}, выражение для $G(k)$ в (8), если заменить $F(k)$ на его выражение через $\tilde{F}(k)$, примет вид:

$$G(k)=C_0\overline{v}-C_1(\rho-1)-C_4\overline{v}-\frac{C_2}{\lambda}\left[\tilde{F}(k)+(k-1)\rho\right].$$

Обратим внимание на то, что $\tilde{F}(k)$ — среднее число заявок, отклоненных за время обслуживания одной заявки (на одном шаге соответствующей цепи Маркова) при стратегии $k+1$ в стационарном режиме работы системы, если число поступивших заявок не меньше, чем число свободных мест в накопителе.

Заметим, что приведенное в [1] доказательство теоремы 1 и следствия 1 полностью основывается на существовании для функции g^k соотношения $g^k-g^{k+1}=\alpha_k[g^k-G(k)]$, где $0<\alpha_k<1$ при $k>$ >0, $G(k)$ не возрастает по $k>0$. То, что $G(k)$ — невозрастающая по $k>0$ функция, следует из неравенства (условия теоремы 1):

$$F(k)-F(k+1)<\rho\,,\ \ 0<\rho<\infty\,,\ k>0\,.$$

Заметим также, что $G(k)$ является невозрастающей функцией по $k>0$ и при условии (см. (21) в [1])

$$\tilde{F}(k)-\tilde{F}(k+1)<\rho\,,\ \ 0<\rho<\infty\,,\ k>0\,, \qquad (10)$$

и теорема 1 остается справедливой, если неравенство для $F(k)$ в условии теоремы заменить на неравенство (10).

Так как в момент перехода системы в любое состояние в накопителе всегда есть хотя бы одно свободное место, длительность интервала времени с момента заполнения накопителя до момента перехода в другое состояние всегда меньше v (длительности нахождения в состоянии) и, следовательно, справедливо неравенство $|\tilde{F}(k)-\tilde{F}(k+1)|<\rho$. Отсюда следует

$$G(k)-G(k+1)=-\frac{C_2}{\lambda}\left[\tilde F(k)-\tilde F(k+1)\right]+C_2\overline{v}\geq$$

$$\geq-\frac{C_2}{\lambda}\left|\tilde F(k)-\tilde F(k+1)\right|+C_2\overline{v}>-\frac{C_2}{\lambda}\,\rho+C_2\overline{v}=0,$$

т. е. в рамках рассматриваемой задачи всегда $G(k)$ — убывающая по $k>0$ функция.

Положим

$$f^k=Q^k=\lambda(1-\theta_k^k)g^k\,,$$

где θ_k^k — вероятность того, что поступившая заявка будет допущена в систему (см. (2)), $\theta_k^k=1-1/(\pi_0^k+\rho)$ [8]. Использовав (9) и приведенное выше равенство $g^k-g^{k+1}=\pi_k^{k+1}[g^k-G(k)]$, находим:

$$\begin{aligned}
f^k-f^{k+1}&=\lambda\left[\left(1-\theta_k^k\right)g^k-\left(1-\theta_{k+1}^{k+1}\right)g^{k+1}\right]=\\
&=\lambda\left[\left(\theta_{k+1}^{k+1}-\theta_k^k\right)g^k+\left(1-\theta_{k+1}^{k+1}\right)\left(g^k-g^{k+1}\right)\right]=\\
&=\lambda\left[\left(\frac{1}{\pi_0^k+\rho}-\frac{1}{\pi_0^{k+1}+\rho}\right)g^k+\right.\\
&\qquad\left.+\,\pi_k^{k+1}\frac{1}{\pi_0^{k+1}+\rho}\left(g^k-G(k)\right)\right]=\\
&=\lambda\left[\pi_k^{k+1}\frac{1}{\pi_0^{k+1}+\rho}\left(g^k-G(k)\right)-\right.\\
&\qquad\left.-\,\frac{\pi_0^k\pi_k^{k+1}}{(\pi_0^k+\rho)(\pi_0^{k+1}+\rho)}\,g^k\right]=\\
&=\lambda\frac{\pi_k^{k+1}}{\pi_0^{k+1}+\rho}\left[g^k\left(1-\frac{\pi_0^k}{\pi_0^k+\rho}\right)-G(k)\right]=\\
&=\lambda\frac{\pi_k^{k+1}}{\pi_0^{k+1}+\rho}\left[g^k\frac{\rho}{\pi_0^k+\rho}-G(k)\right]=\\
&=\frac{\rho\pi_k^{k+1}}{\pi_0^{k+1}+\rho}\left[g^k\frac{\lambda}{\pi_0^k+\rho}-\frac{\lambda G(k)}{\rho}\right]=\\
&=\alpha_k\left[f^k-\tilde G(k)\right],\qquad(11)
\end{aligned}$$

где $\alpha_k=\rho\pi_k^{k+1}/(\pi_0^{k+1}+\rho)$; $\tilde G(k)=G(k)/\overline{v}$.

Так как функция $G(k)$ убывает по $k>0$, то $\tilde G(k)$ также убывает по $k>0$. Кроме того, $0<\alpha_k<1$, $k>0$. Как видим, функция $f^k=Q^k$ обладает всеми свойствами функции g^k использованными при доказательстве теоремы 1.

Следовательно, из доказательства теоремы 1 с учетом того, что $\tilde G(k)$ — строго убывающая функция по $k>0$, следует справедливость следующего утверждения.

Утверждение 1. При любых значениях параметров $0\leq C_i<\infty$, $i=0,\ldots,4$, $0<C_2<\infty$ существует оптимальный порог $0<k^*<\infty$, $k^*=\min\{k>0:\ \tilde G(k)\leq Q^k\}$ При этом если $Q^1\geq\tilde G(1)$, то $k^*=1$, а если $C_2=0$ и $Q^1<\tilde G(1)$, то $k^*=\infty$.

Справедливо также следующее утверждение.

Утверждение 2. Пороговое значение k^* удовлетворяет условию $\max\limits_{k>0}Q^k=Q^{k^*}$ тогда и только тогда, когда k^* удовлетворяет одному из трех условий:

(1) $\tilde G(1)\leq Q^1$, $k^*=1$;

(2) $\tilde G(1)>Q^1$, $k^*=\min\{k>0:\ \tilde Q(k)\leq Q^k\}$;

(3) $Q^{k^*-1}<Q^{k^*}$, $Q^{k^*+1}<Q^{k^*}$, $1<k^*<\infty$.

Доказательство. Необходимость указанных условий следует из утверждения 1, а их достаточность следует из (11) и из того, что $\tilde G(k)$ — убывающая функция: для всех $k>0$, $\tilde G(k)>Q^k$ тогда и только тогда, когда $Q^k<Q^{k+1}$, и если $k'>0$, $\tilde G(k')\leq Q^{k'}$, то $\tilde G(k)<Q^k$ для всех $k>k'$.

Далее приведем формулу, удобную для расчета функции $\tilde F(k)$ и алгоритм поиска оптимального порогового значения. Преобразуем $\overline{F}(k)$:

$$\begin{aligned}
\overline{F}(k)&=\frac{1}{\pi_k^{k+1}}\left(\sum_{j=1}^{k-1}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}[m-(k-j+1)]r_m+\right.\\
&\qquad\left.+\,\pi_0^{k+1}\sum_{m=k}^{\infty}(m-k)r_m\right)=\\
&=\frac{1}{\pi_k^{k+1}}\left(\sum_{j=1}^{k-1}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}[m-(k-j)]r_m-\right.\\
&\qquad\left.-\sum_{j=1}^{k-1}\pi_j^{k+1}\sum_{m=k-j+1}^{\infty}r_m\right)+\\
&+\frac{\pi_0^{k+1}\sum_{m=k}^{\infty}[m-(k-1)]r_m-\pi_0^{k+1}\sum_{m=k}^{\infty}r_m}{\pi_k^{k+1}}=\\
&=\frac{\pi_{k-1}^{k+1}\sum_{m=2}^{\infty}(m-1)r_m}{\pi_k^{k+1}}+\\
&+\frac{1-\pi_k^{k+1}}{\pi_k^{k+1}}\left(\sum_{j=1}^{k-2}\pi_j^{k}\sum_{m=k-j}^{\infty}[m-(k-j)]r_m+\right.\\
&\qquad\left.+\,\pi_0^{k}\sum_{m=k-1}^{\infty}[m-(k-1)]r_m\right)-\\
&-\frac{1-\pi_k^{k+1}}{\pi_k^{k+1}}\left(\sum_{j=1}^{k-1}\pi_j^{k}\sum_{m=k-j+1}^{\infty}r_m+\pi_0^k\sum_{m=k}^{\infty}r_m\right).
\end{aligned}$$

Использовав (6) и (9), перепишем последнее равенство в виде:

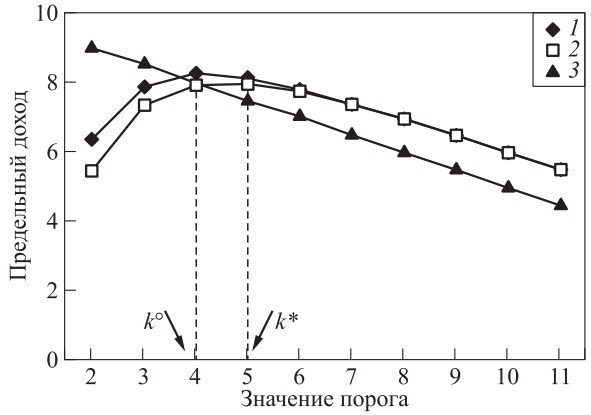

Рис. 1 Зависимости функций Q^k (*1*), g^k (*2*) и $\tilde{G}(k)$ ($G(k)$) (*3*) от порогового значения

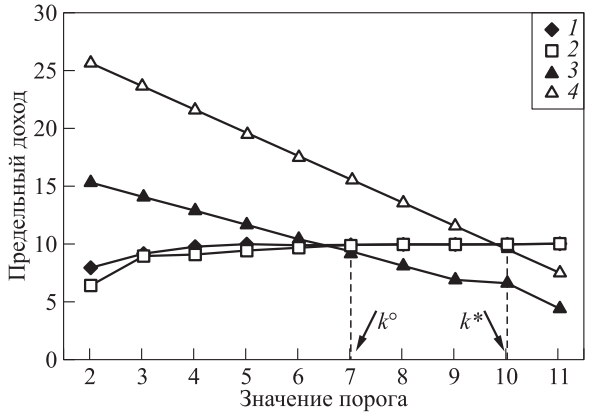

Рис. 2 Зависимости функций Q^k (*1*), g^k (*2*), $\tilde{G}(k)$ (*3*) и $G(k)$ (*4*) от порогового значения

$$\overline{F}(k) = \frac{\pi_{k-1}^{k+1}}{\pi_k^{k+1}} \sum_{m=2}^{\infty} (m-1) r_m + \\ + \frac{1-\pi_k^{k+1}}{\pi_k^{k+1}} \pi_{k-1}^k \overline{F}(k-1) - \frac{1-\pi_k^{k+1}}{\pi_k^{k+1}} \frac{r_0 \pi_k^{k+1}}{1-\pi_k^{k+1}} = \\ = \frac{\pi_{k-1}^{k+1}}{\pi_k^{k+1}} (\rho + r_0 - 1) + \frac{\pi_{k-1}^{k+1}}{\pi_k^{k+1}} \overline{F}(k-1) - r_0 = \\ = \frac{R_{k-1}}{R_k} \left[\overline{F}(k-1) + a\right] - r_0\,, \quad a = \rho + r_0 - 1\,, \\ k \geq 1\,, \quad \overline{F}(0) = 1 - r_0\,.$$

Применив последнюю формулу саму к себе, получим:

$$\tilde{F}(k) - a = \overline{F}(k) = \frac{R_0}{R_k} \overline{F}(0) + a \sum_{i=0}^{k-1} \frac{R_i}{R_k} - \\ - r_0 \left(1 + \sum_{i=1}^{k-1} \frac{R_i}{R_k}\right) = \frac{R_0}{R_k} + a \frac{R_0 + \cdots + R_{k-1}}{R_k} - \\ - r_0 \frac{R_0 + \cdots + R_k}{R_k} = \frac{1}{R_k} + \frac{\rho - 1}{\pi_k^{k+1}} - a\,. \quad (12)$$

Предлагается наиболее простой алгоритм, основанный на условии 2 утверждения 2.

1. Положить $k = 1$.
2. Вычислить $\tilde{G}(k)$ и Q^k.
3. Если $C_2 = 0$ и $Q^k < \tilde{G}(k)$, то положить $k^* = \infty$ и перейти к п. 8.
4. Если выполняется неравенство $\tilde{G}(k) \leq Q^k$, то перейти к п. 7.
5. Увеличить k на единицу.
6. Вычислить $\tilde{G}(k)$, Q^k и перейти к п. 4.
7. Положить $k^* = k$.
8. Конец алгоритма.

Как видим, предложенный алгоритм гарантирует поиск оптимального порогового значения и при этом число вычислений функций Q^k и $\tilde{G}(k)$ не превышает значения оптимального порога. При расчете на очередном шаге алгоритма значений Q^{k+1} и $\tilde{G}(k+1)$ используются рекуррентная формула (4), формулы (12) и (11).

Заметим, что для поиска оптимального порога можно применить метод бинарного поиска, который потребует порядка $\log k^*$ вычислений функций Q^{k+1} и $\tilde{G}(k+1)$.

4 Пример

В качестве примера рассмотрена СМО $M/H_n/1$ с функцией распределения времени обслуживания $H_n(t) = \sum_{i=1}^n f_i(1 - e^{-\mu_i t})$. На рис. 1 и 2 проиллюстрированы зависимости предельного максимального дохода СМО, усредненного по числу обслуженных заявок (g^k), и предельного максимального дохода СМО в единицу времени (Q^k) от порогового значения, а также взаимозависимость функций Q^k (g^k) и $\tilde{G}(k)$ ($G(k)$) при $C_0 = 20$, $C_1 = 10$, $C_2 = 0{,}5$, $C_3 = 0{,}01$, $C_4 = 0{,}01$, $n = 2$ и $\mu_1 = 1$. Графики рис. 1 построены при значениях параметров $\lambda = 2$, $f_1 = 0{,}3$, $f_2 = 0{,}7$ и $\mu_2 = 1$, а рис. 2 — при $\lambda = 1$, $f_1 = 0{,}2$, $f_2 = 0{,}8$ и $\mu_2 = 2$. При этом в случае рис. 1 $r_0 = 1/3$ и $\rho = 2$, в случае рис. 2 $r_0 = 0{,}633$ и $\rho = 0{,}6$.

5 Заключение

Данная работа является непосредственным продолжением работы [1], и повторное исследование этой задачи позволило получить следующие новые результаты:

– доказано, что при $C_2 > 0$ существует оптимальное пороговое значение $k < \infty$, гарантирующее максимальный предельный доход системы $M/G/1$ в единицу времени;
– сформулированы условия, при которых оптимальный порог равен бесконечности;
– доказано утверждение о необходимых и достаточных условиях оптимальности порогового значения;
– предложен алгоритм расчета оптимального порогового значения и значения максимального дохода.

Отметим, что, в отличие от постановки задачи, рассмотренной в [1], в данной работе плату за обслуживание заявки система получает в момент завершения обслуживания и величина платы прямо пропорциональна времени занятия прибора заявкой. Отметим также, что если константу C_0 заменить на любую положительную функцию от времени обслуживания заявки с конечным средним значением, то все приведенные выше утверждения останутся в силе.

Результаты работы могут быть использованы при исследовании и разработке эффективных пороговых стратегий управления в системах с очередями.

Литература

1. *Агаларов Я. М.* Пороговая стратегия ограничения доступа к ресурсам в системе массового обслуживания $M/D/1$ с функцией штрафов за несвоевременное обслуживание заявок // Информатика и её применения, 2015. Т. 9. Вып. 3. С. 56–65.
2. *Агаларов Я. М., Агаларов М. Я., Шоргин В. С.* Об оптимальном пороговом значении длины очереди в одной задаче максимизации дохода системы массового обслуживания типа $M/G/1$ // Информатика и её применения, 2016. Т. 10. Вып. 2. С. 70–79.
3. *Агаларов Я. М., Агаларов М. Я., Шоргин В. С.* Максимизация дохода системы массового обслуживания типа $G/M/1$ на множестве пороговых стратегий с двумя точками переключения // Системы и средства информатики, 2016. Т. 26. Вып. 4. С. 74–88.
4. *Каштанов В. А., Кондрашова Е. В.* Исследование полумарковских систем массового обслуживания при управляемом входящем потоке. BSMAP-поток // Управление большими системами, 2015. Вып. 57. С. 6–36.
5. *Гришунина Ю. Б.* Оптимальное управление очередью в системе $M/G/1/\infty$ с возможностью ограничения приема заявок // Автоматика и телемеханика, 2015. № 3. С. 79–93.
6. *Карлин С.* Основы теории случайных процессов / Пер. с англ. — М.: Мир, 1971. 536 с. (*Karlin S.* A first course in stochastic processes. — New York – London: Academic Press, 1968. 502 p.).
7. *Бочаров П. П., Печинкин А. В.* Теория массового обслуживания. — М.: РУДН, 1995. 529 с.
8. *Miyazawa M.* Complementary generating functions for the M^X/GI/1/k and GI/M^Y/1/k queues and their application to the comparison of loss probabilities // J. Appl. Probab., 1990. Vol. 27. P. 684–692.

Поступила в редакцию 09.02.17

MAXIMIZATION OF AVERAGE STATIONARY PROFIT IN $M/G/1$ QUEUING SYSTEM

Ya. M. Agalarov

Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The problem of optimization of the queue length threshold in a $M/G/1$ system is considered in terms of maximizing the marginal return received by the system per unit of time. The profit value consists of the following measures: service fee; hardware maintenance fee; cost of service delay; fine for unhandled requests; and fine for system idle. The author formulates the necessary conditions of existence of a finite threshold in an $M/G/1$ system and prove the statements of necessary and sufficient conditions for threshold optimality and existence of the finite optimal threshold. The author proposes an algorithm for calculating the optimal threshold value and the corresponding maximal profit. The author presents the results of computational experiments that illustrate the work of the proposed algorithm.

Keywords: queuing system; threshold management; optimization

DOI: 10.14357/19922264170203

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (project 15-07-03406).

References

1. Agalarov, Ya. M. 2015. Porogovaya strategiya ogranicheniya dostupa k resursam v sisteme massovogo obsluzhivaniya $M/D/1$ s funktsiey shtrafov za nesvoevremennoe obsluzhivanie zayavok [The threshold strategy for restricting access in the $M/D/1$ queuing system with penalty function for late service]. *Informatika i ee Primeneniya — Inform. Appl.* 9(3):56–65.
2. Agalarov, Ya. M., M. Ya. Agalarov, and V. S. Shorgin. 2016. Ob optimal'nom porogovom znachenii dliny ocheredi v odnoy zadache maksimizatsii dokhoda sistemy massovogo obsluzhivaniya tipa $M/G/1$ [About the optimal threshold of queue length in particular problem of profit maximization in the $M/G/1$ queuing system]. *Informatika i ee Primeneniya — Inform. Appl.* 10(2):70–79.
3. Agalarov, Ya. M., M. Ya. Agalarov, and V. S. Shorgin. 2016. Maksimizatsiya dokhoda sistemy massovogo obsluzhivaniya tipa $G/M/1$ na mnozhestve porogovykh strategiy s dvumya tochkami pereklyucheniya [Profit maximization in $G/M/1$ queuing system on a set of threshold strategies with two switch points]. *Sistemy i Sredstva Informatiki — System and Means of Informatics* 26(4):74–88.
4. Kashtanov, V. A., and E. V. Kondrashova. 2015. Issledovanie polumarkovskikh sistem massovogo obsluzhivaniya pri upravlyaemom vkhodyashchem potoke. BSMAP-potok [Research of semi-Markov queueing models using controlled input flow. BSMAP-flow]. *Upravlenie bol'shimi sistemami* [Large-Scale Systems Control] 57:6–36.
5. Grishunina, Yu. B. 2015. Optimal control of queue in the $M/G/1/\infty$ system with possibility of customer admission restriction. *Automat. Rem. Contr.* 76(3):433–445.
6. Karlin, S. 1968. *A first course in stochastic processes*. New York – London: Academic Press. 502 p.
7. Bocharov, P. P., and A. V. Pechinkin. 1995. *Teoriya massovogo obsluzhivaniya* [Queueing theory]. Moscow: RUDN. 529 p.
8. Miyazawa, M. 1990. Complementary generating functions for the M^X/GI/1/k and GI/M^Y/1/k queues and their application to the comparison of loss probabilities. *J. Appl. Probab.* 27:684–692.

Received February 9, 2017

Contributor

Agalarov Yaver M. (b. 1952) — Candidate of Science in technology, associate professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; agglar@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 33–41

КЛАССИФИКАЦИЯ ПО НЕПРЕРЫВНЫМ НАБЛЮДЕНИЯМ С МУЛЬТИПЛИКАТИВНЫМИ ШУМАМИ II: АЛГОРИТМ ЧИСЛЕННОЙ РЕАЛИЗАЦИИ ОЦЕНКИ*

А. В. Борисов[1]

Аннотация: Данная работа является второй частью статьи «Классификация по непрерывным наблюдениям с мультипликативными шумами I: формулы байесовской оценки», опубликованной в журнале «Информатика и её применения», 2017, том 11, выпуск 1. Исследования посвящены решению задачи оценивания случайного вектора с конечным множеством состояний по непрерывным зашумленным наблюдениям. Особенностью модели является то, что интенсивность шумов в наблюдениях зависит от оцениваемого вектора, что не позволяет применять классические результаты оптимальной нелинейной фильтрации. В первой части статьи искомая оценка получена как в явной интегральной форме, так и в виде решения некоторой стохастической дифференциальной системы со скачкообразными процессами в правой части. Во второй части представлен алгоритм приближенного вычисления оценки и характеристики его точности. Модельный пример иллюстрирует качество предлагаемой оценки и соответствующей численной процедуры.

Ключевые слова: оптимальная фильтрация; идентифицируемость; рекуррентная схема; порядок точности; дискретизация по времени

DOI: 10.14357/19922264170204

1 Введение

Статья является окончанием [1] и имеет следующую структуру.

Раздел 2 содержит постановку исследуемой задачи классификации, а также краткое изложение теоретических результатов, представленных в первой части работы.

В разд. 3 предлагается алгоритм приближенного вычисления оценки и характеристики точности соответствующих аппроксимаций. Раздел 4 содержит модельный пример, иллюстрирующий свойства предлагаемой оценки классификации и алгоритма ее вычисления.

Обсуждение результатов и заключительные комментарии представлены в разд. 5.

2 Постановка задачи и сводка имеющихся результатов

Рассматривается следующая система наблюдения:

$$\left.\begin{aligned} &dX_t = 0\,,\ \ X_0 = X\,;\\ &dY_t = \sum_{n=1}^{N} e_n^{\top} X_t f_t(n)\,dt + \\ &\qquad + \sum_{n=1}^{N} e_n^{\top} X_t g_t(n)\,dW_t\,,\ \ Y_0 = 0\,, \end{aligned}\right\} \tag{1}$$

где

- X_t — ненаблюдаемое состояние системы: начальное условие X принимает значения из множества $\mathbb{S}^N \triangleq \{e_1, \ldots, e_N\}$ единичных векторов евклидова пространства $\mathbb{R}^N$ с вероятностями $\{p_n\}_{n=\overline{1,N}}$, $p \triangleq \operatorname{col}(p_1, \ldots, p_N)$;
- Y_t — M-мерный процесс наблюдений;
- $W_t \in \mathbb{R}^M$ — независимый от X векторный стандартный винеровский процесс, характеризующий ошибки наблюдений;
- $f_t(n) : \mathbb{S}^N \times [0, +\infty) \to \mathbb{R}^{M\times 1}$ $(n = \overline{1,N})$ — набор неслучайных вектор-функций, характеризующий «план наблюдений»,

*Работа выполнена при финансовой поддержке РФФИ (проекты 16-07-00677 и 15-37-20611-мол_а_вед).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, aborisov@frccsc.ru

– $g_t(n):\ \mathbb{S}^N\times[0,+\infty)\to\mathbb{R}^{M\times M}$ $(n=\overline{1,N})$ — набор равномерно невырожденных неслучайных матричнозначных функций, характеризующий условную интенсивность шумов в наблюдениях в зависимости от значения состояния X_t.

Функции $\{f_t(n)\}_n$ и $\{g_t(n)\}_n$ непрерывны справа и имеют конечные пределы слева.

Через $\mathcal{Y}_t\triangleq\sigma\{Y_s:\ s\in[0,t]\}$ обозначен естественный поток σ-алгебр, порожденный наблюдениями Y до момента t включительно.

Задача байесовской классификации вектора X по наблюдениям Y, полученным на отрезке времени $[0,T]$, заключается в нахождении $\widehat{X}_T\triangleq\mathbf{E}\{X|\mathcal{Y}_T\}$.

Известно [2], что поток $\{\mathcal{Y}_t\}_{t\geqslant0}$ не является непрерывным справа, т. е. $\bigcap\limits_{s>t}\mathcal{Y}_s\neq\mathcal{Y}_t$, что не позволяет непосредственно применять эффективный аппарат стохастического анализа [3] для решения поставленной задачи оценивания, а требует некоторой ее модификации.

Задача локального сглаживания вектора X по наблюдениям Y заключается в нахождении $\widehat{X}_T^+\triangleq\triangleq\mathbf{E}\{X|\mathcal{Y}_{T+}\}$.

Обозначим

$$G_t(n)\triangleq g_t(n)g_t^\top(n)\,,\ \ \mathbf{G}_t(n)\triangleq\int\limits_0^t g_s(n)g_s^\top(n)\,ds\,.$$

Лемма 1 [1] определяет решение задачи байесовской классификации:

$$\widehat{X}_T(n)=\frac{\widetilde{X}_T(n)}{\sum\limits_{\ell=1}^N\widetilde{X}_T(\ell)}\,,\qquad(2)$$

где $\widetilde{X}_T(n)$ — ненормированная условная вероятность события $\{\omega:\ X(\omega)=e_n\}$ относительно $\mathcal{Y}_T$:

$$\widetilde{X}_T(n)=\begin{cases}p_n\exp\left\{\displaystyle\int\limits_0^{\mathrm{T}}\left[f_s^\top(n)G_s^{-1}(n)\,dY_s-\right.\right.\\ \left.\left.-\dfrac12\,\|f_s(n)\|^2_{G_s^{-1}(n)}\,ds\right]\right\},\ \text{если}\\ \qquad\langle Y,Y\rangle_t\equiv\mathbf{G}_t(n),\ \ t\in[0,T);\\ 0\ \text{в противном случае.}\end{cases}\qquad(3)$$

Здесь $\|x\|^2_A\triangleq x^\top Ax$ и $|A|\triangleq\det A$.

Для определения локально сглаженной оценки определим неслучайные моменты $u(\ell,n)$ $(\ell\neq n)$:

$$u(\ell,n)\triangleq\begin{cases}\inf\{t\geqslant0:\ \mathbf{G}_t(\ell)\neq\mathbf{G}_t(n)\}\ ;\\ +\infty,\ \text{если}\ \mathbf{G}_t(\ell)\equiv\mathbf{G}_t(n)\ \text{для}\ \forall\,t\geqslant0,\end{cases}$$

множества $\Xi(n)\triangleq\{u(\ell,n)\}_{\ell:\ell\neq n}$, $n=\overline{1,N}$, и множество $\Xi\triangleq\{u(\ell,n)\}_{(\ell,n):\ \ell\neq n}$.

Наблюдения $\{Y_t\}_{t\geqslant0}$ являются квадратично интегрируемым субмартингалом с $\mathcal{Y}_t$-предсказуемой квадратической характеристикой

$$\langle Y,Y\rangle_t=Y_tY_t^\top-\int\limits_0^t Y_s\,dY_s^\top-\int\limits_0^t dY_sY_s^\top=$$
$$=\int\limits_0^t\sum_{n=1}^N e_n^\top X_sG_s(n)\,ds\,.$$

Случайные моменты времени

$$\xi(n)\triangleq\begin{cases}\inf\{t\geqslant0:\ \mathbf{G}_t(n)\neq\langle Y,Y\rangle_t\}\ ;\\ +\infty,\ \text{если}\ \mathbf{G}_t(n)\equiv\langle Y,Y\rangle_t\ \text{для}\ \ t\geqslant0,\end{cases}$$

используют $\Xi(n)$ в качестве множества возможных значений и определяют следующие скачкообразные $\mathcal{Y}_{t+}$-согласованные процессы:

$$\mathcal{I}_t(n)\triangleq\begin{cases}1, & \text{если}\ t<\xi(n)\,;\\ 0, & \text{если}\ t\geqslant\xi(n)\,.\end{cases}$$

Теорема 1 [1] определяет локально сглаженную оценку $\widehat{X}_t^+=\mathcal{I}_t(n)\widehat{X}_t$ как решение стохастической дифференциальной системы с наблюдениями Y_t и скачкообразными процессами $\mathcal{I}_t(n)$ в правой части $(n=\overline{1,N})$:

$$\widehat{X}_t^+(n)=\frac{p_n\mathcal{I}_0(n)}{\sum\limits_{k=1}^N p_k\mathcal{I}_0(k)}+$$
$$+\int\limits_0^t\widehat{X}_s^+(n)\left(f_n^\top(s)-\sum_{\ell=1}^N\widehat{X}_{s-}^+(\ell)f_\ell^\top(s)\right)\times$$
$$\times\left(\frac{d\langle Y,Y\rangle_s}{ds}\right)^{-1/2}dZ_s+$$
$$+\sum_{s\leqslant t}\widehat{X}_{s-}^+(n)\left(\frac{1+\Delta\mathcal{I}_s(n)}{1+\sum\limits_{\ell=1}^N\widehat{X}_{s-}^+(\ell)\Delta\mathcal{I}_s(\ell)}-1\right),\qquad(4)$$

где Z_t — обновляющий процесс:

$$Z_t\triangleq\int\limits_0^t\left(\frac{d\langle Y,Y\rangle_s}{ds}\right)^{-1/2}\left(dY_s-\sum_{n=1}^N e_n^\top\widehat{X}_{s-}^+f_s(n)ds\right).$$

3 Алгоритм вычисления оценок классификации и его точность

На первый взгляд, для численного решения системы (4) применимы известные алгоритмы [4]. Однако в правую часть системы входят $\mathcal{Y}_{t+}$-согласованные процессы $d\langle Y,Y\rangle_t/dt$ и $\{\mathcal{I}_t(n)\}_{n=\overline{1,N}}$. Они

недоступны прямому наблюдению, и для их вычисления требуются нетривиальные преобразования наблюдений, по сложности эквивалентные решению исходной задачи оценивания.

В данной работе предлагается дискретизовать непрерывный процесс и строить оптимальные оценки уже по дискретизованным наблюдениям. Такой подход близок по смыслу к численным методам, предложенным в [5]. Введем в рассмотрение последовательность вложенных двоичных разбиений отрезка $[0,T]$, порожденных множествами точек $\mathcal{T}^K$, $K\in\mathbb{N}$:

$$\mathcal{T}^K \triangleq \{\tau_k^K\}_{k=\overline{0,2^K}}:\ \ \tau_k^K \triangleq kh_K\,,\ \ h_K \triangleq \frac{T}{2^K}\,,$$

соответствующие наборы дискретизованных наблюдений:

$$\Delta Y_k^K \triangleq Y_{\tau_k^K} - Y_{\tau_{k-1}^K} = \int\limits_{\tau_{k-1}^K}^{\tau_k^K} \sum_{n=1}^{N} e_n^\top X_s f_s(n)\,ds + \int\limits_{\tau_{k-1}^K}^{\tau_k^K} \sum_{n=1}^{N} e_n^\top X_s g_s(n)\,dW_s\,,\ \ k=\overline{1,2^K}\,,$$

и семейство вложенных σ-алгебр $\{\mathcal{Y}^K\}_{K\in\mathbb{N}}$: $\mathcal{Y}^K \triangleq \sigma\{\Delta Y_k^K,\ k=\overline{1,2^K}\}$. Так как процесс Y_t является сепарабельным, то $\mathcal{Y}^K \uparrow \mathcal{Y}_T$ при $K\to\infty$ и по теореме Леви [6]

$$\widehat{X}^K \triangleq \mathbf{E}\left\{X|\mathcal{Y}^K\right\} \to \mathbf{E}\left\{X|\mathcal{Y}_T\right\} = \widehat{X}_T\ \ \mathcal{P}\text{-п. н.}$$

Тогда компоненты $\widehat{X}^K(n) \triangleq \mathbf{P}\left\{X=e_n|\mathcal{Y}^K\right\}$ вектора $\widehat{X}^K$ определяются формулами:

$$\widehat{X}^K(n) = \frac{\widetilde{X}^K(n)}{\sum_{\ell=1}^{N}\widetilde{X}^K(\ell)} \tag{5}$$

и

$$\widetilde{X}^K(n) \triangleq p_n \exp\left\{-\frac{1}{2h_K}\sum_{k=1}^{2^K}\left[\ln\left|G_k^K(n)\right| h_K + \left\|\Delta Y_k^K - h_K F_k^K(n)\right\|^2_{\left(G_k^K(n)\right)^{-1}}\right]\right\}, \tag{6}$$

где

$$F_k^K(n) \triangleq \frac{1}{h_K}\int\limits_{\tau_{k-1}^K}^{\tau_k^K} f_s(n)\,ds\,,$$

$$G_k^K(n) \triangleq \frac{1}{h_K}\int\limits_{\tau_{k-1}^K}^{\tau_k^K} g_s(n)g_s^\top(n)\,ds\,.$$

Для упрощения выкладок будем дополнительно считать, что существует такое целое число K^*, что $\Xi\backslash\{+\infty\}\subset\mathcal{T}^K$ для всех $K>K^*$. Формулы (5) и (6) могут быть реализованы с помощью следующей рекуррентной схемы ($n=\overline{1,N}$):

$$\left.\begin{aligned}
&\widehat{\varkappa}_0^K(n) = p_n\,;\\
&\widetilde{\varkappa}_{\tau_k^K}^K(n) = \widehat{\varkappa}_{\tau_{k-1}^K}^K(n)\,|G_k(n)|^{-1/2}\times\\
&\times\exp\left\{-\frac{1}{2h_K}\|\Delta Y_k^K - h_K F_k^K(n)\|^2_{(G_k^K(n))^{-1}}\right\};\\
&\widehat{\varkappa}_{\tau_k^K}^K(n) = \frac{\widetilde{\varkappa}_{\tau_k^K}^K(n)}{\sum_{\ell=1}^{N}\widetilde{\varkappa}_{\tau_k^K}^K(\ell)}\,,\ \ \tau_k^K\in\mathcal{T}^K\,.
\end{aligned}\right\} \tag{7}$$

Теорема 1. *Набор векторов $\{\widehat{\varkappa}_t^K\}_{t\in\mathcal{T}^K}$ является аппроксимацией оценки $\widehat{X}_t$ на сетке $\mathcal{T}^K$ $\left(\widehat{X}_t\approx\widehat{\varkappa}_t^K,\ t\in\mathcal{T}^K\right)$. Аппроксимация обеспечивает точность порядка $1/2$, т. е.*

$$\sqrt{\mathbf{E}\left\{\|\widehat{\varkappa}_t^K - \widehat{X}_t\|^2\right\}} = O\left(h_K^{1/2}\right)\ \ \forall\,\tau_k^K\in\mathcal{T}^K. \tag{8}$$

Доказательство теоремы 1 приведено в приложении.

Порядок аппроксимации $1/2$ является обычным для традиционных схем численного решения стохастических дифференциальных систем с винеровскими и пуассоновскими процессами в правой части [4]. Однако «локальный порядок точности» аппроксимации $\widehat{\varkappa}_t^K(n)$ компоненты $\widehat{X}_t(n)$ сразу после момента скачка $\mathcal{I}_t(n)$ абсолютно другой.

Теорема 2. *Пусть множество $\Xi(n)$ возможных значений момента $\xi(n)$ и значение временно́го лага $H_K>0$ удовлетворяют следующим условиям*:

(а) $\max\{u:\ u\in\Xi(n)\}<+\infty$;

(б) $\Xi(n)\subseteq\mathcal{T}^{K^*}$, $\{t+H_k\}_{t\in\Xi(n)}\subseteq\mathcal{T}^{K^*}$ *для некоторого* $K^*\in\mathbb{N}$;

(в) *неравенство*

$$\min_{t\in\Xi(n)}\int\limits_{t}^{t+H_K}\left[\ln\frac{|G_s(n)|}{|G_s(\ell)|} + \operatorname{tr}\left(G_s(\ell)G_s^{-1}(n)\right)\right]ds \geqslant C_1$$

выполняется для некоторой константы $C_1>0$.

Тогда аппроксимация $\widehat{\varkappa}_{\xi(n)+H_K}^K$ оценки $\widehat{X}_{\xi(n)+H_k}$ имеет экспоненциальную точность, т. е.

$$\left(\mathbf{E}\left\{\left(\widehat{\varkappa}_{\xi(n)+H_K}^K(n) - \widehat{X}_{\xi(n)+H_K}(n)\right)^2\times\mathbf{I}_{\{\xi(n)<+\infty\}}(\omega)\right\}\right)^{1/2} \leqslant C_2\exp\left(-\frac{C_1}{2h_K}\right)$$

для некоторого $C_2>0$.

Если дополнительно

$$\min_{\substack{(s,t):\ t\in\Xi(n),\\ s\in(t,t+H_K]}} \left[\ln\frac{|G_s(n)|}{|G_s(\ell)|} + \operatorname{tr}\left(G_s(\ell)G_s^{-1}(n)\right)\right] \geqslant C_3$$

для некоторого $C_3 > 0$*, то*

$$\left(\mathbf{E}\left\{\left(\widehat{\varkappa}^K_{\xi(n)+H_K}(n) - \widehat{X}_{\xi(n)+H_K}(n)\right)^2 \times \right.\right.$$
$$\left.\left. \times\, \mathbf{I}_{\{\xi(n)<+\infty\}}(\omega)\right\}\right)^{1/2} \leqslant C_2 \exp\left(-\frac{C_3 H_K}{2h_K}\right).$$

Доказательство теоремы 2 дано в приложении.

Так как оценки $\{\widehat{X}_t\}_{t\geqslant 0}$ и $\{\widehat{X}^+_t\}_{t\geqslant 0}$ различаются не более чем на множестве Ξ потенциальных моментов скачков, численная схема (7) может быть использована также и для вычисления локально сглаженной оценки $\widehat{X}^+$. Действительно, ввиду теоремы 2, аппроксимации оценок классификации (7) могут быть легко преобразованы в аппроксимации оценок сглаживания с фиксированным лагом путем очевидного переопределения:

$$\widehat{X}^+_{\tau^K_k} \approx \widehat{\varkappa}^K_{\tau^K_k+H_K},$$

где H_K — величина лага. Процедура локального сглаживания позволит значительно увеличить точность аппроксимации в точках разрыва решения при малых абсолютных значениях временно́го лага H_K. Если выполнены условия второй части теоремы 2, то лаг следует выбирать из условия:

$$\lim_{K\to+\infty}\frac{H_K}{h_K} = +\infty.$$

4 Численный пример

Данный численный пример является абстрактным, однако он иллюстрирует свойства предложенной оценки и алгоритма ее аппроксимации.

Рассматривается система наблюдения (1) со следующими параметрами: $N = 4$; $M = 1$; $T = 1$; $p = \operatorname{col}(1/4, 1/4, 1/4, 1/4)$;

$$f_t(1) = 1\cdot\mathbf{I}_{[0,1/4)}(t) + 1\cdot\mathbf{I}_{[1/2,3/4)}(t);$$
$$f_t(2) = 2\cdot\mathbf{I}_{[0,1/4)}(t) + 2\cdot\mathbf{I}_{[1/2,3/4)}(t);$$
$$f_t(3) = 3\cdot\mathbf{I}_{[0,1/4)}(t) + 3\cdot\mathbf{I}_{[1/2,3/4)}(t);$$
$$f_t(4) = 4\cdot\mathbf{I}_{[0,1/4)}(t) + 4\cdot\mathbf{I}_{[\frac{1}{2},3/4)}(t);$$
$$g_t(1) = 1\cdot\mathbf{I}_{[0,3/4)}(t) + 1{,}1\cdot\mathbf{I}_{[3/4,1/2]}(t);$$
$$g_t(2) = 1\cdot\mathbf{I}_{[0,3/4)}(t) + 1{,}2\cdot\mathbf{I}_{[3/4,1/2]}(t);$$
$$g_t(3) = 1\cdot\mathbf{I}_{[0,3/4)}(t) + 1{,}3\cdot\mathbf{I}_{[3/4,1/2]}(t);$$

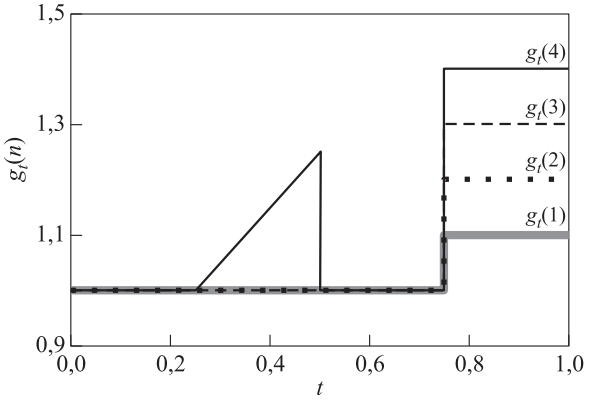

Рис. 1 Интенсивности шумов $g_t(n)$ $(n = \overline{1,4})$

$$g_t(4) = 1{,}0\cdot\mathbf{I}_{[0,1/2)}(t) + (t-1/4)\cdot\mathbf{I}_{[1/4,1/2)}(t) + 1{,}0\cdot\mathbf{I}_{[1/2,3/4)}(t) + 1{,}4\cdot\mathbf{I}_{[3/4,1]}(t).$$

Графики $\{g_n\}$ приведены на рис. 1.

Точки разрывности потока σ-алгебр $\mathcal{Y}_t$ образуют множество $\Xi = \{1/4, 1/2, 3/4\}$. Система наблюдения такова, что компонента $X(4)$ может быть точно восстановлена по $\mathcal{Y}_{(1/4)+}$, в то время как весь вектор X может быть точно восстановлен по $\mathcal{Y}_{(3/4)+}$. При этом в момент $t = 1/4$ для компоненты $X(4)$ выполняется первое условие идентифицируемости теоремы 2, а в момент $t = 3/4$ для оставшихся компонент — второе условие идентифицируемости.

Еще одной из целей данного примера является сравнение точности фильтрации при наличии аддитивных и мультипликативных шумов.

Наконец, в примере компоненты сноса $\{f_n\}$ значительно отличаются друг от друга (до 400%), в то время как компоненты диффузии $\{g_n\}$ варьируются значительно меньше (до 40%).

Результаты примера демонстрируют, что даже малые различия интенсивностей шумов в наблюдениях обеспечивают быстрое и точное восстановление вектора X.

Аппроксимации $\widehat{\varkappa}^K_t$ были вычислены с помощью алгоритма (7) при разных значениях шага дискретизации $h_K = 2^{-K}$ ($K = 10$, 12, 14, 16). Результаты оценивания представлены на рис. 2. В этом эксперименте $X = e_1$. Графики демонстрируют сходимость $\widehat{\varkappa}^K_t \to \widehat{X}_t$ при $K \to +\infty$.

Точность оценок фильтрации и их численной реализации представлены на рис. 3.

На нем приведены графики среднеквадратических отклонений (СКО) $\widehat{\sigma}^{16}_t$ ошибок оценивания компонент $X(n)$, вычисленных при шаге дискретизации $h_K = 2^{-16}$. Значения СКО рассчитываются с помощью метода Монте Карло путем осредне-

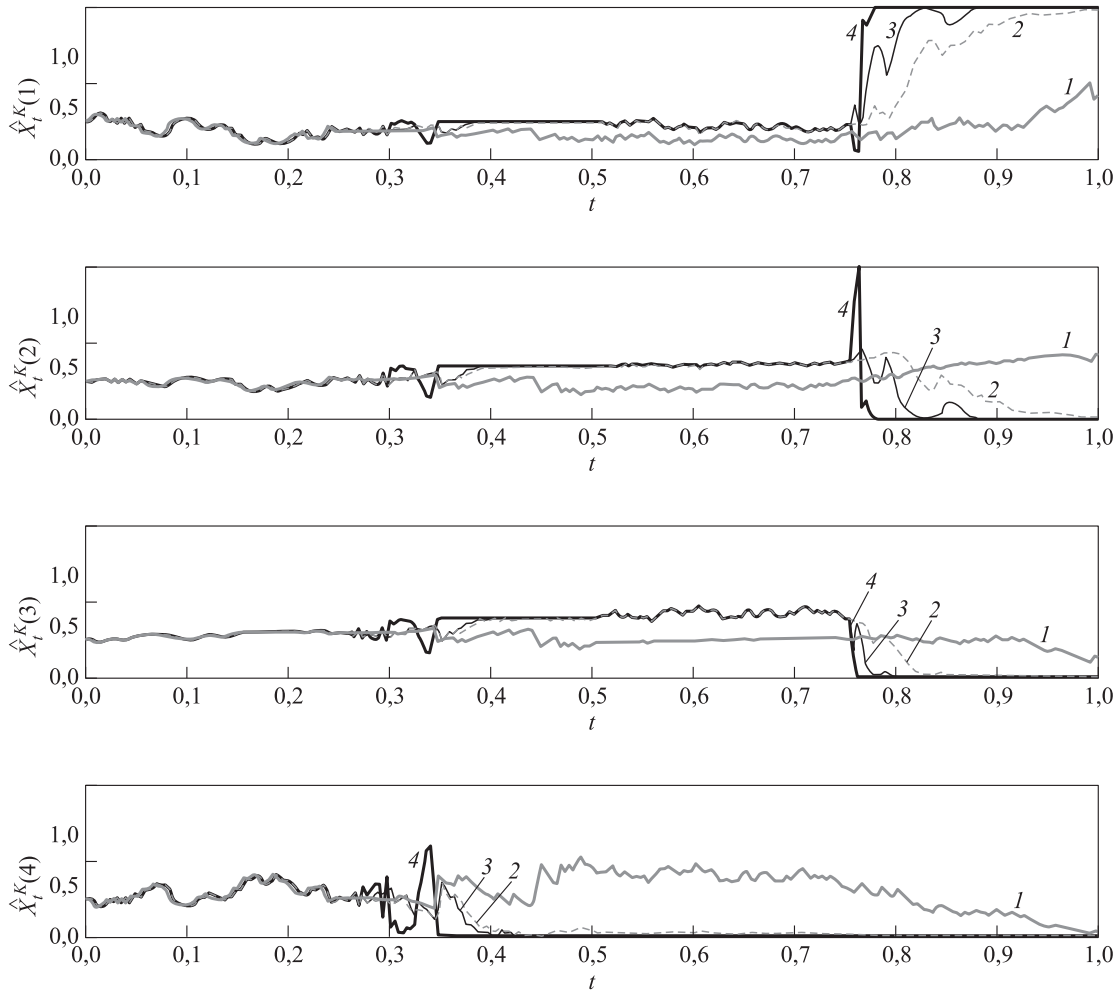

Рис. 2 Оценки $\widehat{\varkappa}_t^K(n)$, вычисленные при $h_K = 2^{-K}$: *1* — $K = 10$; *2* — 12; *3* — 14; *4* — $K = 16$

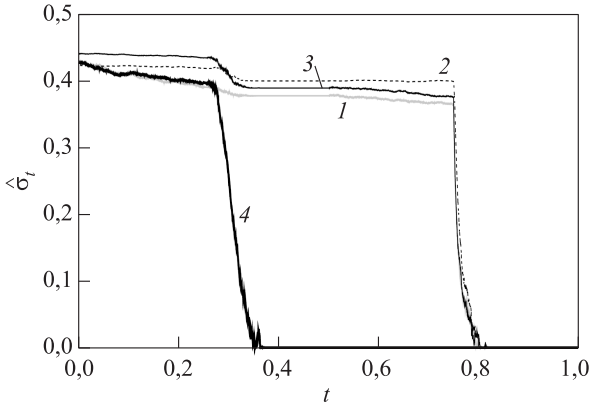

Рис. 3 Среднеквадратические отклонения $\widehat{\sigma}_t^{16}(n)$ ошибок оценок $\widehat{X}_t(n)$: *1* — $n = 1$; *2* — 2; *3* — 3; *4* — $n = 4$

ния 1000 траекторий. Графики зависимости СКО от времени подтверждают представленные теоретические результаты. Иследуемая система наблюдения сконструирована таким образом, что на интервалах времени $(0, 1/4)$ и $(1/2, 3/4)$ интенсивности шумов в наблюдениях одинаковы, а различаются только коэффициенты сноса. Это классический случай наблюдений с аддитивными шумами.

Из графиков можно заключить, что в данном примере на интервалах аддитивных шумов точность оценивания с ростом времени наблюдения меняется незначительно: не более чем на 10% за интервал наблюдения длиной $1/4$. В то же время после момента $t = 1/4$ компонента $X(4)$ идентифицируется точно по наблюдениям, полученным на интервале времени длиной менее 0,12. Оставшиеся компоненты точно идентифицируются после момента $t = 3/4$ за время не более 0,06. Причиной такой точности оценивания является наличие мультипликативных шумов и выполнение условий идентифицируемости. Компоненты X не идентифицируются точно в моменты $1/4$ и $3/4$ из-за того, что в примере используются аппроксимации оценок, а не их

точные значения. При этом время точной идентификации компоненты $X(4)$ больше, чем время идентификации остальных компонент. Это связано с тем, что в случае идентификации $X(4)$ выполняются мягкие условия идентифицируемости, определенные первой частью теоремы 2, а для остальных компонент — более ограничительные, но «эффективные» для идентификации условия второй части теоремы.

Одной из целей рассматриваемого примера являлась проверка порядка точности $1/2$ численной аппроксимации (7). Согласно следствию 1 [1], $\widehat{X}_t^+ = X$ $\mathcal{P}$-п. н. для любого $t > 3/4$. В качестве временно́го лага была выбрана величина $H = 2^{-7}$. Поэтому для исследования порядка точности был выбран момент времени $t^* = 3/4 + 1/2^7$. Среднеквадратическая ошибка аппроксимации (СКОА) $\Delta_{t^*}^K(n) \triangleq \mathbf{E}\left\{(\widehat{\varkappa}_{t^*}^K(n) - X(n))^2\right\}$ компонент вектора вычислялась методом Монте Карло путем осреднения пучка 1000 траекторий. Согласно теореме 1 $\Delta_{t^*}^K(n) \leqslant C_4 h_K$ для некоторого $C_4 > 0$ и любого $K > K^*$. Так как $h_K = 2^{-K}$, то $\log_2 \Delta_{t^*}^K(n) \leqslant \leqslant \log_2 C_4 - K$. Рисунок 4 демонстрирует зависимость $\log_2 \Delta_{t^*}^K$ от K ($K = \overline{7,20}$) для всех ком понент вектора. Графики зависимостей являются вогнутыми, что соответствует утверждению теоремы 1, и демонстрируют консервативный характер оценки (8): порядок аппроксимации, полученный в рассмотренном примере, выше чем $1/2$.

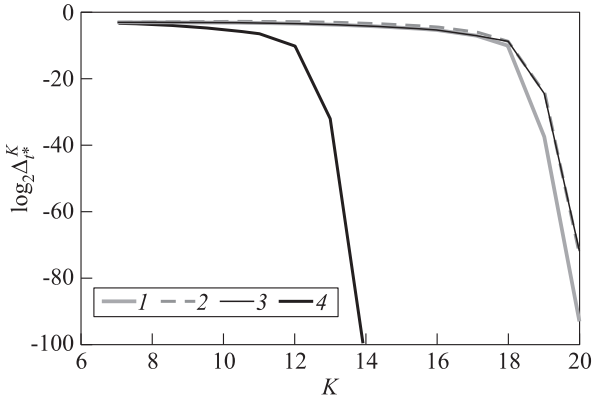

Рис. 4 Зависимость СКОА $\Delta_{t^*}^K$ от K: *1* — $n = 1$; *2* — 2; *3* — 3; *4* — $n = 4$

5 Заключение

Вторая часть статьи посвящена разработке численного алгоритма, реализующего решение задачи байесовской классификации.

Предлагается отказаться от непосредственной численной аппроксимации решения стохастической дифференциальной системы, определяющей искомую оценку классификации, в пользу построения байесовской оценки по дискретизованным наблюдениям. Помимо формул, определяющих численную схему аппроксимации искомой оценки, в работе также исследована точность этой аппроксимации. Предложенный численный пример иллюстрирует свойства оценки и качество ее аппроксимаций.

Полученные в обеих частях работы теоретические результаты и схемы численной аппроксимации могут рассматриваться как часть алгоритмического обеспечения систем информатики при решении задач вероятностно-статистического экспресс-моделирования и оценивания по высокочастотным данным.

Приложение

Доказательство теоремы 1. Доказательство теоремы опирается на формулы (2), (3), (5) и (6). Заметим, что на промежутке времени $(0, \xi(n))$ производная $d\langle Y, Y\rangle_t/dt$ существует почти всюду по мере Лебега и также почти всюду верно равенство

$$\frac{d\langle Y, Y\rangle_t}{dt} \equiv G_t(n)\,.$$

Введем следующие обозначения:

$$\psi(n) \triangleq \int_0^{\mathrm{T}} \left[\ln|G_s(n)| + \mathrm{tr}\left(G_s(\omega)G_s^{-1}(n)\right)\right]\, ds\,;$$

$$\eta(n) \triangleq -\int_0^{\mathrm{T}} \left[\frac{1}{2}\|f_s(n)\|^2_{G_s^{-1}(n)} ds - f_s^{\top}(n)G_s^{-1}(n)\, dY_s\right]\,.$$

Зафиксируем $n \in \{1, \ldots, N\}$ и построим множества

$$S_T \triangleq \{\ell:\ 1 \leqslant \ell \leqslant N:\ \mathbf{G}_t(\ell) \equiv \langle Y, Y\rangle_t,\ t \in [0, T]\};$$

$$\overline{S}_T \triangleq \{1, \ldots, N\}\backslash S_T\,.$$

Очевидно, что $S_T \neq \varnothing$, так как гарантированно $n \in S_T$. Из определения случайного момента $\xi(n)$ следует, что $\{\omega:\ \xi(n) \geqslant T\} = \{\omega:\ X(\omega) = e_q,\ q \in S_T\}$. При этом из доказательства леммы 1 [1] следует, что на множестве $\{\omega:\ X(\omega) = e_q,\ q \in S_T\}$ $\psi(\ell) - \psi(q) \equiv 0\ \forall\ \ell \in S_T$ и $\psi(j) - \psi(q) > 0\ \forall\ j \in \overline{S}_T$.

Далее без ограничения общности для упрощения выкладок будем считать, что

$$\mathbf{P}\{\omega:\ \xi(n) < T\} > 0\,.$$

Оценим сверху средний квадрат ошибки аппроксимации $I^K(n) \triangleq \mathbf{E}\left\{(\widehat{X}^K(n) - \widehat{X}(n))^2\right\}$:

$$I^K(n) = \sum_{q \in S_T} \mathbf{E}\left\{(\widehat{X}^K(n) - \widehat{X}(n))^2 \mathbf{I}_{\{X = e_q\}}(\omega)\right\} + \\ + \sum_{j \in \overline{S}_T} \mathbf{E}\left\{(\widehat{X}^K(n) - \widehat{X}(n))^2 \mathbf{I}_{\{X = e_j\}}(\omega)\right\}.$$

Сначала для фиксированного $q \in S_T$ рассмотрим на множестве $\{\omega:\ X(\omega) = e_q\}$ случайную величину $|\widehat{X}^K(n) - \widehat{X}(n)|\mathbf{I}_{\{X=e_q\}}(\omega)$, учитывая, что на этом множестве

$$\widehat{X}(n) = \frac{\widetilde{X}(n)}{\sum_{\ell\in S_T}\widetilde{X}(\ell)} = \frac{p_n e^{\eta(n)}}{\sum_{\ell\in S_T} p_\ell e^{\eta(\ell)}}\,.$$

Тогда верна следующая цепочка неравенств:

$$\begin{aligned}
&\left|\widehat{X}^K(n) - \widehat{X}(n)\right|\mathbf{I}_{\{X=e_q\}}(\omega) = \\
&\quad = \widehat{X}(n)\mathbf{I}_{\{X=e_q\}}(\omega)\left|\frac{\widehat{X}^K(n)}{\widehat{X}(n)} - 1\right| = \\
&\quad = \widehat{X}(n)\mathbf{I}_{\{X=e_q\}}(\omega)\left|\frac{\widetilde{X}^K(n)\sum_{\ell\in S_T}\widetilde{X}(\ell)}{\sum_{m=1}^N \widetilde{X}^K(m)\widetilde{X}(n)} - 1\right| = \\
&\quad = \widehat{X}(n)\mathbf{I}_{\{X=e_q\}}(\omega)\left(\left|\sum_{\ell\in S_T}\left(\frac{\widetilde{X}^K(n)}{\widetilde{X}(n)}\,\widetilde{X}(\ell) - \right.\right.\right. \\
&\quad \left.\left.\left. - \widetilde{X}^K(\ell)\right) - \sum_{j\in\overline{S}_T}\widetilde{X}^K(j)\right|\right) \Bigg/ \sum_{m=1}^N \widetilde{X}^K(m) \leqslant \\
&\quad \leqslant \mathbf{I}_{\{X=e_q\}}(\omega)\left(\left|\sum_{\ell\in S_T}\left(\frac{\widetilde{X}^K(n)}{\widetilde{X}(n)}\,\widetilde{X}(\ell) - \right.\right.\right. \\
&\quad \left.\left.\left. - \widetilde{X}^K(\ell)\right) - \sum_{j\in\overline{S}_T}\widetilde{X}^K(j)\right|\right) \Big/ \widetilde{X}^K(q) \leqslant \\
&\quad \leqslant \mathbf{I}_{\{X=e_q\}}(\omega)\left[\sum_{\ell\in S_T}\left|\frac{\widetilde{X}^K(n)}{\widetilde{X}^K(q)}\,\frac{\widetilde{X}(\ell)}{\widetilde{X}(n)} - \frac{\widetilde{X}^K(\ell)}{\widetilde{X}^K(q)}\right| + \right. \\
&\quad \left. + \sum_{j\in\overline{S}_T}\frac{\widetilde{X}^K(j)}{\widetilde{X}^K(q)}\right] = \\
&\quad = \mathbf{I}_{\{X=e_q\}}(\omega)\left(\sum_{\ell\in S_T} e^{\eta(\ell)-\eta(q)-\phi^K(q)}\left|e^{\phi^K(n)} - e^{\phi^K(q)}\right| + \right. \\
&\quad \left. + \sum_{j\in\overline{S}_T} e^{-((\psi(j)-\psi(q))/(2h_K))+\eta(j)-\eta(q)+\phi^K(j)-\phi^K(q)}\right) \leqslant \\
&\quad \leqslant \mathbf{I}_{\{X=e_q\}}(\omega)\left(\sum_{\ell\in S_T} e^{\eta(\ell)-\eta(q)-\phi^K(q)}\left|e^{\phi^K(n)} - e^{\phi^K(q)}\right| + \right. \\
&\quad \left. + e^{-(\mu(q)/(2h_K))}\sum_{j\in\overline{S}_T} e^{\eta(j)-\eta(q)+\phi^K(j)-\phi^K(q)}\right),
\end{aligned}$$

где

$$\mu(q) \triangleq \min_{j\in\overline{S}_T}\int_0^T\left[\ln\frac{|G_s(j)|}{|G_s(q)|} + \operatorname{tr}\left(G_s(q)G_s^{-1}(j)\right)\right]ds > 0\,.$$

Тогда $\forall\ q \in S_T$

$$\begin{aligned}
&\mathbf{E}\left\{(\widehat{X}^K(n) - \widehat{X}(n))^2\mathbf{I}_{\{X=e_q\}}(\omega)\right\} \leqslant \\
&\quad \leqslant N\sum_{\ell\in S_T}\mathbf{E}\left\{\mathbf{I}_{\{X=e_q\}}(\omega)e^{2(\eta(\ell)-\eta(q)-\phi^K(q))}\times\right. \\
&\quad \left.\times\left(e^{\phi^K(n)} - e^{\phi^K(q)}\right)^2\right\} + Ne^{-\mu(q)/h_K}\times \\
&\quad \times\sum_{j\in\overline{S}_T}\mathbf{E}\left\{\mathbf{I}_{\{X=e_q\}}(\omega)e^{2(\eta(j)-\eta(q)+\phi^K(j)-\phi^K(q))}\right\}.
\end{aligned}$$

При условии $X(\omega) = e_q$ величины $\{\phi^K(\ell)\}_{\ell=\overline{1,N}}$ имеют гауссовское распределение. Можно показать, что в случае, когда $\mathbf{E}\left\{(\phi^K(\ell))^2\right\} = O(h_K)$, оценка $\mathbf{E}\left\{(\phi^K(\ell) - \phi^K(m))^4\right\} = O(h_K^2)$ верна для любых $\ell, m \in \{1,\ldots,N\}$ и любой зависимости $\phi^K(\ell)$ и $\phi^K(m)$. При этом случайные величины $\{e^{4(\eta(\ell)-\eta(q)-\phi^K(q))}\}_{\ell\in S_T}$ и $\{e^{2(\eta(j)-\eta(q)+\phi^K(j)-\phi^K(q))}\}_{j\in\overline{S}_T}$ имеют логнормальное распределение с ограниченными по ℓ и j математическими ожиданиями. Одновременно с этим функция $e^{-(\mu(q)/h_K)}$ стремится к 0 при $h_K \to 0$ быстрее любой степени h_K. Используя все вышесказанное, а также неравенство Коши–Буняковского, можно построить следующую цепочку неравенств:

$$\begin{aligned}
&\mathbf{E}\left\{(\widehat{X}^K(n) - \widehat{X}(n))^2\mathbf{I}_{\{X=e_q\}}(\omega)\right\} \leqslant \\
&\leqslant N\sum_{\ell\in S_T}\sqrt{\mathbf{E}\left\{\mathbf{I}_{\{X=e_q\}}(\omega)e^{4(\eta(\ell)-\eta(q)-\phi^K(q))}\right\}}C_{nq}h_K + \\
&\quad + Ne^{-(\mu(q)/h_K)}\times \\
&\quad \times\sum_{j\in\overline{S}_T}\mathbf{E}\left\{\mathbf{I}_{\{X=e_q\}}(\omega)e^{2(\eta(j)-\eta(q)+\phi^K(j)-\phi^K(q))}\right\} \leqslant \\
&\quad \leqslant Q_{nq}h_K \qquad (9)
\end{aligned}$$

для некоторых положительных констант C_{nq} и Q_{nq}, что и доказывает истинность оценки (8).

Далее рассмотрим величину $|\widehat{X}^K(n) - \widehat{X}(n)|\mathbf{I}_{\{X=e_q\}}(\omega)$ на множестве $\{\omega:\ X(\omega) = e_q\}$ для некоторого фиксированного $q \in \overline{S}_T$. Из (2) и (3) следует, что $\widehat{X}(n)\mathbf{I}_{\{X=e_q\}}(\omega) = 0$ $\mathcal{P}$-п. н., поэтому

$$\begin{aligned}
&\left|\widehat{X}^K(n) - \widehat{X}(n)\right|\mathbf{I}_{\{X=e_q\}}(\omega) = \\
&\quad = \mathbf{I}_{\{X=e_q\}}(\omega)\frac{p_n\widetilde{X}^K(n)}{\sum_{\ell=1}^N p_n\widetilde{X}^K(\ell)} \leqslant \mathbf{I}_{\{X=e_q\}}(\omega)\frac{p_n}{p_q}\times \\
&\quad \times e^{-((\psi(n)-\psi(q))/(2h_k))+\eta(n)-\eta(q)+\phi^K(n)-\phi^K(q)}\,, \qquad (10)
\end{aligned}$$

причем из леммы 1 [1] следует, что $\psi(n) - \psi(q) > 0$ на множестве $\{\omega:\ X(\omega) = e_q\}$. Вычисляя математические ожидания квадратов левой и правой частей (10) и используя ту же аргументацию, можно видеть, что оценка (9) справедлива также и для всех $q \in \overline{S}_T$. Таким образом,

$$I^K(n) \leqslant \sum_{q=1}^N Q_{nq}h_K\,,$$

что доказывает истинность оценки (8). Теорема 1 доказана.

Доказательство теоремы 2. Воспользуемся приемами доказательства теоремы 1. Прежде всего, $\widehat{X}_{\xi(n)+H_K}(n) = 0$ $\mathcal{P}$-п. н., а также

$$\widehat{\varkappa}^K_{\xi(n)+H_K}(n)\mathbf{I}_{\{\xi(n)<T\}}(\omega) = \sum_{q:q\neq n} \widehat{\varkappa}^K_{\xi(n)+H_K}(n)\mathbf{I}_{\{X=e_q\}}(\omega) =$$

$$= \sum_{q:q\neq n} \mathbf{I}_{\{X=e_q\}}(\omega) \left(\widehat{\varkappa}^K_{u(q,n)}(n) \exp\left\{ -\frac{\psi_{H_K}(n)}{2h_K} + \right.\right.$$

$$\left.\left. + \eta_{H_K}(n) + \phi^K_{H_K}(n) \right\}\right) \Bigg/ \sum_{\ell=1}^{N} \widehat{\varkappa}^K_{u(q,n)}(\ell) \exp\left\{ -\frac{\psi_{H_K}(\ell)}{2h_K} + \right.$$

$$\left. + \eta_{H_K}(\ell) + \phi^K_{H_K}(\ell) \right\},$$

где

$$\psi_{H_K}(\ell) \triangleq \int_{u(q,n)}^{u(q,n)+H_K} \left[|G_s(n)| + \operatorname{tr}\left(G_s(q) G_s^{-1}(n) \right) \right] ds;$$

$$\eta_{H_K}(\ell) \triangleq \int_{u(q,n)}^{u(q,n)+H_K} \left(\frac{1}{2} \|f_s(n)\|^2_{G_s^{-1}(n)} - \right.$$

$$\left. - f_s^{\top}(n) G_s^{-1}(n)\, dY_s \right),$$

а $\{\phi^K_{H_K}(\ell)\}_{\ell=\overline{1,N}}$ — набор таких случайных последовательностей, что $\mathbf{E}\left\{ (\phi^K_{H_K}(\ell))^2 \right\} = O(h_K)$.

Тогда в силу условия (в) теоремы верно следующее неравенство:

$$\widehat{\varkappa}^K_{\xi(n)+H_K}(n)\mathbf{I}_{\{\xi(n)<T\}}(\omega) \leqslant$$

$$\leqslant \mathbf{I}_{\{X=e_q\}}(\omega) e^{-C_1/(2h_K)} \sum_{q:q\neq n} \frac{\widehat{\varkappa}^K_{u(q,n)}(n)}{\widehat{\varkappa}^K_{u(q,n)}(q)} \exp\left\{ \eta_{H_K}(n) + \right.$$

$$\left. + \phi^K_{H_K}(n) - \eta_{H_K}(q) - \phi^K_{H_K}(q) \right\}.$$

Далее

$$\mathbf{E}\left\{ (\widehat{\varkappa}^K_{\xi(n)+H_K}(n) - \widehat{X}_{\xi(n)+H_K}(n))^2 \mathbf{I}_{\{\xi(n)<T\}}(\omega) \right\} \leqslant$$

$$\leqslant e^{-C_1/h_K} (N-1) \times$$

$$\times \sum_{q:q\neq n} \mathbf{E}\left\{ \mathbf{I}_{\{X=e_q\}}(\omega) \left(\frac{\widehat{\varkappa}^K_{u(q,n)}(n)}{\widehat{\varkappa}^K_{u(q,n)}(q)} \right)^2 \exp\left\{ 2\left(\eta_{H_K}(n) + \right.\right.\right.$$

$$\left.\left.\left. + \phi^K_{H_K}(n) - \eta_{H_K}(q) - \phi^K_{H_K}(q) \right) \right\} \right\}. \quad (11)$$

Математические ожидания в правой части (11) ограничены по $K \in \mathbb{N}$ некоторой положительной константой Q, поэтому первое утверждение теоремы непосредственно следует из (11), если обозначить $C_2 = \sqrt{(N-1)Q}$. Второе утверждение теоремы доказывается абсолютно аналогично путем замены константы C_1 на $C_3 H_K$.

Теорема 2 доказана.

Литература

1. *Борисов А. В.* Классификация по непрерывным наблюдениям с мультипликативными шумами I: формулы байесовской оценки // Информатика и её применения, 2017. Т. 11. Вып. 1. С. 11–19.
2. *Стоянов Й.* Контрпримеры в теории вероятностей / Пер. с англ. — М.: МЦНМО, 2014. 296 с. (*Stoyanov J.* Counterexamples in probability. — New York, NY, USA: John Wiley, 1987. 313 p.)
3. *Липцер Р. Ш., Ширяев А. Н.* Теория мартингалов. — М.: Наука, 1986. 512 с.
4. *Platen E., Bruti-Liberati N.* Numerical solution of stochastic differential equations with jumps in finance. — New York, NY, USA: Springer, 2010. 868 p.
5. *Platen E., Rendek R.* Quasi-exact approximation of hidden Markov chain filters // Commun. Stoch. Anal., 2010. Vol. 4. No. 1. P. 129–142.
6. *Липцер Р. Ш., Ширяев А. Н.* Статистика случайных процессов. — М.: Наука, 1974. 512 с.

Поступила в редакцию 19.12.16

CLASSIFICATION BY CONTINUOUS-TIME OBSERVATIONS IN MULTIPLICATIVE NOISE II: NUMERICAL ALGORITHM

A. V. Borisov

Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: This is the second part of the paper “Classification by continuous-time observations in multiplicative noise I: Formulae for Bayesian estimate” published in “Informatics and Applications,” 2017, 11(1). Investigations are aimed at estimation of a finite-state random vector given continuous-time noised observations. The key feature

is that the observation noise intensity is a function of the estimated vector, which makes useless the known results in the optimal filtering. In the first part of the paper, the required estimate is obtained both in the explicit integral form and as a solution to a stochastic differential system with some jump processes in the right-hand side. The second part contains a numerical algorithm of the estimate approximate calculation together with its accuracy analysis. An example illustrating the performance of the proposed estimate is also presented.

Keywords: optimal filtering; identifiability; recursive scheme; approximation order; time discretization

DOI: 10.14357/19922264170204

Acknowledgments

The work was supported in part by the Russian Foundation for Basic Research (projects Nos. 15-37-20611 and 16-07-00677).

References

1. Borisov, A. V. 2017. Klassifikatsiya po nepreryvnym nablyudeniyam s mul'tiplikativnymi shumami I: Formuly Bayesovskoy otsenki [Classification by continuous-time observations in multiplicative noise I: Formulae for Bayesian estimate] // *Informatika i ee Primeneniya — Inform. Appl.* 11(1):11–19.
2. Stoyanov, J. 1987. *Counterexamples in probability.* New York, NY: John Wiley. 313 p.
3. Liptser, R. Sh., and A. N. Shiryayev. 1989. *Theory of martingales.* New York, NY: Springer. 812 p.
4. Platen, E., and N. Bruti-Liberati. 2010. *Numerical solution of stochastic differential equations with jumps in finance.* New York, NY: Springer. 868 p.
5. Platen, E., and R. Rendek. 2010. Quasi-exact approximation of hidden Markov chain filters. *Commun. Stoch. Anal.* 4(1):129–142.
6. Liptser, R. Sh., and A. N. Shiryayev. 2001. *Statistics of random processes: I. General theory.* Berlin: Springer. 427 p.

Received December 19, 2016

Contributor

Borisov Andrey V. (b. 1965) — Doctor of Science in physics and mathematics, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; aborisov@frccsc.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 42–49

ИНФОРМИРОВАННОСТЬ УЧАСТНИКОВ И СУЩЕСТВОВАНИЕ РАВНОВЕСИЯ В ПОЗИЦИОННЫХ МНОГОШАГОВЫХ ИГРАХ МНОГИХ ЛИЦ

Н. С. Васильев[1]

Аннотация: В позиционных играх изучаются динамические модели принятия решений в условиях конфликта интересов участников и при информированности о текущей позиции игры. Каждый игрок может оказывать некоторое допустимое управляющее воздействие на общую для всех динамическую систему. Выбираемая игроком стратегия управления — это функция, определенная на фазовом пространстве системы. Отслеживая траекторию движения, все стороны конфликта имеют неявное представление о стратегиях партнеров. Принцип рационального поведения всех игроков состоит в стремлении достичь ситуации равновесия Нэша. Доказано, что к нему можно прийти в результате коллективных усилий по выбору совместного программного управления системой. Устойчивость решения обеспечивается угрозой наказания, применяемого к игроку, не выполняющего эту программу. Контроль за соблюдением согласованной траектории движения и некоторая дополнительная информация позволяют игрокам идентифицировать нарушителя. Его наказание реализуется с запаздыванием, связанным с необходимостью обнаружения «виновника». Теорема существования равновесия применена к исследованию экономико-математической модели.

Ключевые слова: динамическая система; дифференциальная игра; позиционная многошаговая игра; программное управление; позиционная стратегия; контрстратегия; стратегия наказания; гарантирующая стратегия; ситуация равновесия Нэша; эффективность по Парето

DOI: 10.14357/19922264170205

1 Введение

Значение информированности игроков для принятия рациональных решений было выявлено в исследованиях Ю. Б. Гермейера, Н. Н. Моисеева и их учеников (В. А. Горелика, А. Ф. Кононенко, Н. С. Кукушкина, В. В. Фёдорова и др.) [1–3]. Аксиоматизация понятия стратегии формализует динамику принятия решений в статических играх, моделирующую рефлексию игроков при обмене информацией [4]. В дифференциальных и многошаговых играх добавляется другой тип динамики. Игроки принимают решения, управляя и наблюдая за траекторией движения системы. В каждой текущей позиции разыгрывается «локальная» статическая игра, участники которой получают неявное знание о стратегиях партнеров в форме изменения позиций игры. Стратегии игроков являются воздействиями на управляемую систему, распределенными по ее фазовому пространству.

Начало изучению дифференциальных игр положили задачи преследования–убегания [5]. В трудах Л. С. Понтрягина и его учеников эта теория развивалась в рамках программного управления [6, 7]. Позиционные стратегии (управления в форме синтеза) потребовали обобщить понятие движения динамической системы, которое, вообще говоря, составляет ансамбль траекторий [8]. В классе произвольных позиционных стратегий было установлено существование седловой точки [8]. Позиционные дифференциальные игры развивались в трудах Н. Н. Красовского, А. И. Субботина [8], В. А. Горелика, А. Ф. Кононенко, Ю. Е. Чистякова [9–13], В. И. Жуковского [14], Л. А. Петросяна [15] и других авторов. При этом рассматривались игры как антагонистические [5–8], так и с непротивоположными интересами игроков [2, 9–15]. Равновесие Нэша является одним из принципов рационального поведения игроков [1]. Существование ситуации равновесия в дифференциальных играх установлено в работах [10, 11].

Развитие сетевых технологий и компьютерной техники стимулирует исследование моделей коллективного поведения, в которых все участники игры заинтересованы в рациональном разрешении конфликта интересов. Принятие решений существенно усложняется в позиционной игре с n, $n \geq 2$, участниками, которые управляют движением динамической системы, описываемой дифференциальными или разностными уравнениями. Игро-

[1]Московский государственный технический университет им. Н. Э. Баумана, nik8519@yandex.ru

ки преследуют, вообще говоря, несовпадающие цели, стремясь по возможности оптимизировать свои функции выигрыша. Выбираемые игроками стратегии определяют траекторию системы, а вместе с нею и результат игры. Всякая текущая позиция игры (t, x_t) предполагается известной всем игрокам. Поэтому стратегиями являются отображения $\tilde{u} : \{(t, x)\} \to U$ фазового пространства системы (совокупности позиций игры (t, x)) в некоторое множество U, которое ограничивает управляющие возможности игроков.

Подстановка в дифференциальное уравнение произвольного управления в форме синтеза $\tilde{u} = u(t, x)$ создает проблему, связанную с существованием и единственностью решения задачи Коши. Неслучайно, что в ранних исследованиях по дифференциальным играм классы допустимых стратегий игроков ограничивались, например, гладкими отображениями $\tilde{u}_t : x \to u(t, x)$ или даже программными управлениями $\tilde{u} : t \to u(t)$. Для приложений предпочтительней сохранить единственность траектории движения системы и не ограничивать игроков в выборе своих стратегий. Поэтому далее рассматриваются игровые многошаговые позиционные задачи, в которых, в отличие от работ [9, 15], не фиксируется порядок ходов игроков. Этот класс задач является дискретным аналогом дифференциальных игр, нуждающимся в самостоятельном исследовании. В этой работе использован подход Ю. Б. Гермейера и А. Ф. Кононенко, применяемый для поиска равновесий Нэша в дифференциальных играх двух лиц [11, 13]. Равновесие обеспечивается угрозой наказания [2, 9–13]. В играх многих лиц реализация этого подхода опирается на использование разрешающего правила, которое служит обнаружению партнера по игре, нарушающего «принимаемое» всеми соглашение о совместном программном управлении системой. Предложенное в статье правило опирается на сбор «минимальной» дополнительной информации о ходе игры.

2 Многошаговая позиционная игра

2.1 Постановка игровой задачи

Пусть в игровой операции принимают участие игроки $i = 1, 2, \ldots, n$. Они выбирают стратегии $\tilde{u}^i = u^i_t(x) = u(t, x)$, которые на промежутке времени $T = \{t_0, t_0 + 1, \ldots, T - 1\}$ определяют движение динамической системы,

$$x_{t+1} = g\left(t, x_t, \tilde{u}\right),\ \tilde{u} = \left(\tilde{u}^1, \tilde{u}^2, \ldots, \tilde{u}^n\right),\ t \in T, \quad (1)$$

стартующей из начальной позиции $(t_0, x_{t_0}) \in \in N \times R^m$. Позиционные стратегии игроков — это отображения $\tilde{u}^i : T \times R^m \to U^i$, принимающие значения в заданных множествах $U^i = U^i(t, x)$. Такие управляющие воздействия игроков называются допустимыми. Интерес каждого участника игры состоит в максимизации по возможности своего критерия эффективности

$$w^i = f^i(x_T),\ \ i = 1, 2, \ldots, n. \quad (2)$$

Функции выигрыша (2), зависящие от терминального состояния системы (1) x_T, отвечающего моменту времени T, не ограничивают общности задачи. Для платежей игроков (2) введем дополнительное обозначение $w^i = F^i(\tilde{u})$, в явном виде отражающее зависимость результатов игры от выбираемых стратегий. Если анализируется семейство игр (1), (2), в котором произвольно варьируется начальная позиция (t, x), то для критерия (2), $f = (f^1, \ldots, f^n)$, применяется более подробная запись $w^i = F^i_{t,x}(\tilde{u})$, $i = 1, 2, \ldots, n$. При этом вдоль траектории системы (1) $F^i_{t,x_t}(\tilde{u}) = F^i_{t+1,x_{t+1}}(\tilde{u})$.

Согласно (1), (2) выигрыш любого игрока определяется не только его действиями $\tilde{u}^i$. Выбор параметров $\tilde{u}^i$, $j \neq i$, контролируется остальными участниками игры. Поэтому ни один из игроков не имеет возможности точно прогнозировать величину получаемого выигрыша и вынужден выбирать свою стратегию в условиях «хаоса» будущих результатов игры. Применяемые позиционные стратегии $\tilde{u}^i$ могут зависеть от некоторой дополнительной информации I, которой располагают игроки, и тогда стратегии $\tilde{u}^i = u^i_t(x; I)$ [2, 4]. Коалиционное поведение игроков не рассматривается.

Всякая ситуация игры $\tilde{u}$ определяет некоторый результат (2). Наибольший интерес представляет собой нахождение ситуаций, отвечающих принципам оптимального поведения [2]. Исследуем вопрос нахождения ситуация равновесия Нэша, реализующего принцип устойчивости. Напомним, что равновесная ситуация $\tilde{u}^*$ обладает следующим свойством:

$$(\forall\, i)\left(\forall\, \tilde{u}^i\right) F^i\left(\tilde{u}^* \left| \tilde{u}^i \right.\right) \leq F^i\left(\tilde{u}^*\right). \quad (3)$$

Здесь $\tilde{u}^* \left| \tilde{u}^i \right. = \left(\tilde{u}^{*1}, \ldots, \tilde{u}^{*i-1}, \tilde{u}^i, \tilde{u}^{*i+1}, \ldots, \tilde{u}^{*n}\right)$. Далее показано, что участники позиционных игр вполне могут рассчитывать на выбор эффективных по Парето равновесий, отвечающих и принципу выгодности [2, 4].

2.2 Основные определения и допущения

Никто из игроков «не согласится» получить выигрыш, меньший того, который он может сам себе

гарантировать. Игру (1), (2) естественно рассматривать в рамках семейства задач управления системой (1), отличающихся лишь начальной позицией (t,x). Если существует ситуация равновесия, то в ней платежи игроков не могут быть меньше наилучших гарантированных результатов [2, 4], равных

$$L^i(t,x) = \max_{\tilde{u}^i \in \tilde{U}^i} \min_{\tilde{u}^j \in \tilde{U}^j, j \neq i} F^i(\tilde{u})\,. \qquad (4)$$

Величина $L = (L^i, i = 1,2,\ldots,n)$ зависит от классов допустимых стратегий игроков $\tilde{U}^i$, которые зафиксированы в постановке задачи и поэтому опущены в обозначениях. Будем считать, что все ограничивающие множества U^i, $i = 1,2,\ldots,n$, конечны. Тогда в определении (4) и во всех последующих экстремальных задачах максимумы и минимумы достигаются. Поэтому, например, игроки имеют *гарантирующие* стратегии $\tilde{u}^{\Gamma i}$, доставляющие максимумы в задачах (4), $i = 1,2,\ldots,n$. Всех игроков объединяет общее стремление — достичь множества $X_T(L) = \{x_T : \ f(x_T) \geq L\} \neq \emptyset$ [4]. Более того, целесообразно попасть в ту его часть P_T, элементы которой $x_T^* \in P_T$ являются эффективными решениями статической игры $\Gamma_T = (f, X_T)$. Напомним [2], что векторы $x_T^* \in P_T$, отвечающие принципу эффективности Парето, обладают свойством

$$\neg\,(\exists\, x_T\,(x_T \in X_T) \wedge (f(x_T) \geq f(x_T^*)) \wedge \\ \wedge\,(f(x_T) \neq f(x_T^*)))\,.$$

В позиционной игре (1)–(3) всякая ситуация $\tilde{u}^*$ эффективна, если управление $\tilde{u}^*$ задает траекторию x_t^*, для которой $x_T^* \in P_T$. Целесообразно дальнейшее сужение целевого терминального множества P_T вплоть до подмножества $X_T(M) \cap P_T$, где вектор $M = M(t,x)$ имеет координаты $M(T,x) = L(T,x) \stackrel{\triangle}{=} f(x)$ и

$$M^i(t,x) = \min_{\tilde{u}^j \in \tilde{U}^j, j \neq i} \max_{\tilde{u}^i \in \tilde{U}^i} F^i(\tilde{u})\,, \quad i = 1,2,\ldots,n. \qquad (5)$$

Напомним, что $M \geq L$ [4]. Решение задачи (5) $\tilde{u}^{\mathrm{н}j}(i) \equiv u^{\mathrm{н}j}(t,x;i)$, $j \neq i$, называется *стратегией наказания* i-го игрока [2]. При исполнении наказания его оптимальная стратегия $\tilde{u}^{\mathrm{a}i}$ доставляет максимум в (5). Для сравнения: гарантирующая стратегия игрока обеспечивает ему результат $w^{\Gamma i}$, который при движении системы (1) располагается в диапазоне значений (см. леммы 1 и 2)

$$L^i(t-1,x_{t-1}) \leq L^i(t,x_t) \leq w^{\Gamma i} \equiv \\ \equiv F^i_{t,x_t}\left(\tilde{u}^{\Gamma i}, \tilde{u}^{\mathrm{н}}(i)\right) \leq M^i(t-1,x_{t-1}) \leq M^i(t,x_t)\,.$$

Зафиксируем произвольное программное управление $\overline{u}_t = (\overline{u}_t^1, \overline{u}_t^2, \ldots, \overline{u}_t^n)$ и рассмотрим множество позиций $D^i_\tau(\overline{u})$, $t_0 + 1 \leq \tau \leq T-1$, $i = 1,2,\ldots,n$, достижимых системой (1) в момент τ с помощью управлений вида $u = (\overline{u}|u^i_\tau)$, $u^i_\tau \in U^i$, т. е.

$$u^i(t,x) = \begin{cases} \overline{u}_t^j\,, & t \leq \tau - 1\,,\ j \neq i \vee t \leq \tau - 2\,; \\ u^i\,, & t = \tau - 1\,. \end{cases}$$

Управление системой (1) происходит на промежутке $t_0 \leq t \leq \tau - 1$. Введем величины $(i = 1,2,\ldots,n)$:

$$\left.\begin{aligned} \overline{M}^i(\overline{u}) &= \max_{(t,x) \in D^i(\overline{u})} M^i(t,x)\,; \\ D^i(\overline{u}) &= \bigcup_{\tau=t_0+1}^{T-1} D^i_\tau(\overline{u})\,. \end{aligned}\right\} \qquad (6)$$

Из них составим вектор

$$\overline{M}(\overline{u}) = \left(\overline{M}^1(\overline{u}), \overline{M}^2(\overline{u}), \ldots, \overline{M}^n(\overline{u})\right)\,.$$

Пусть множество $A \subset T \times R^m$. Рассмотрим движение $t \to \overline{x}_t$ системы (1), $u = \overline{u}$, для которого $(t,\overline{x}_t) \in A$, $t = t_0,\ldots,T-2$. Допустим, что любой i-й игрок вместо $\overline{u}^i$ в момент времени t начинает использовать произвольную допустимую позиционную стратегию $\tilde{u}^i$. Если у остальных игроков имеется такая контрстратегия $\widehat{u}(i) = \left(\widehat{u}^j(i), j \neq i\right)$, применяемая с момента $t+1$, что траектория $\tilde{x}$ системы (1), $\tilde{u} = \left(\widehat{u}(i), \tilde{u}^i\right)$, не покидает множества A, т. е. $(\tau,\tilde{x}_\tau) \in A$, $\tau = t+1,\ldots,T-1$, то будем говорить о *стабильности* движения $t \to \overline{x}_t$ системы (1) относительно множества $A \subset T \times R^m$. Введенное определение является дискретным аналогом понятия стабильного моста, используемого в дифференциальных играх [8].

Предположим, что игроки приняли соглашение применять программное управление u^*, которому отвечает движение системы X^*. Первое отклонение $\| x_t - x_t^* \| \neq 0$ системы (1) от траектории X^* становится известным всем игрокам и служит сигналом нарушения в момент времени $t-1$ достигнутой договоренности. *Разрешающим правилом* назовем такую «процедуру», которая позволяет любому участнику конфликта идентифицировать игрока-нарушителя. Это правило должно опираться на использование каких-либо дополнительных сведений I о ходе игры: информации о текущей позиции игры (t,x_t), вообще говоря, недостаточно. Потребуем выполнения следующих допущений.

Д1. Пусть существует известное всем игрокам такое программное управление $u(t) = u_t^*$ системой (1), что $x_T^* \in X_T(\overline{M}(u^*))$.

Д2. У игроков имеется разрешающее правило, применяемое к управлению u^*.

Д3. Ситуация u^*_{T-1} — равновесие Нэша в игре $\Gamma_{T-1} = (f \circ g, U), U = U^1 \times U^2 \times \cdots \times U^n$.

Рассмотрим произвольное программное управление $\overline{u}_t$ и связанное с ним подмножество фазового пространства системы (1) вида

$$S(\overline{u}) = \bigcup_i \left\{ (t,x) :\ M^i(t,x) \le \overline{M}^i(\overline{u}) \right\} .$$

Лемма 1. *В предположении* Д2, *применяемом к* $\overline{u}$, *всякое движение* $t \to \overline{x}_t$ *динамической системы* (1) *стабильно относительно множества позиций* $S(\overline{u})$.

Доказательство. Во-первых, в соответствии с (6) «трубка» $\bigcup_i D^i(\overline{u}) \subset S(\overline{u})$. Во-вторых, согласно предположению Д2, в качестве контрстратегии $\widehat{u}(i)$ на стратегию i-го игрока $\tilde{u}^i \in \tilde{U}^i$ можно взять стратегию наказания $\tilde{u}^{\text{н}}(i) = (u^{\text{н}j}(t,x;i),\ j \ne i) \in \prod\limits_{j \ne i} \tilde{U}^j$ (см. (5)). Покажем, что она обеспечивает стабильность движения $t \to \overline{x}_t$ относительно S.

Применение наказания начинается в той позиции игры $(t, \tilde{x}_t) \in D^i_t(\overline{u})$, для которой впервые $\tilde{x}_t \ne \overline{x}_t$. Пусть $\tilde{x}_{t+1}$ — следующее состояние системы (1), отвечающее управлению $\tilde{u}^i, \tilde{u}^{\text{н}}(i)$. Тогда в силу (2), (5) и (6)

$$M^i(t+1, \tilde{x}_{t+1}) \le \max_{\substack{\tilde{u}^i_\tau \in \tilde{U}^i, \\ \tau \ge t+1}} F^i_{t,\tilde{x}} \left(\tilde{u}^i, \tilde{u}^{\text{н}}(i) \right) \le$$

$$\le \max_{\substack{\tilde{u}^i_\tau \in \tilde{U}^i, \\ \tau \ge t}} F^i_{t,\overline{x}}(\tilde{u}^i, \tilde{u}^{\text{н}}(i)) \stackrel{\triangle}{=} M^i(t, \tilde{x}_t) \le \overline{M}^i(\overline{u}) .$$

Неравенство можно продолжить далее для всех моментов времени $t+2, \ldots, T-1$ и получить цепь соотношений:

$$M^i(T-1, \tilde{x}_{T-1}) \le \cdots \le M^i(t+1, \tilde{x}_{t+1}) \le$$

$$\le M^i(t, \tilde{x}_t) \le \overline{M}^i(\overline{u}) .$$

Это доказывает, что движение системы $t \to \tilde{x}_t$ проходит по множеству $S(\overline{u})$.

Лемма 2. $M^i(t, \overline{x}_t) = M^i(t+1, \overline{x}_{t+1})$ *вдоль траектории системы* (1), $\tilde{u} = (\tilde{u}^{\text{а}i}, \tilde{u}^{\text{н}}(i))$.

Доказываемое равенство является следствием определения (5) стратегии $\tilde{u}$. Если равновесие существует, то результат игры $w^* = f(x^*_T) \ge M(t_0, x_0)$. Предположение Д3 необходимо для равновесия и обусловлено «краевым» эффектом, связанным с конечностью процесса управления системой (1), а именно: предпоследней позиции $(T-1, x^*_{T-1})$ динамической игровой задачи (1), (2) отвечает статическая игра Γ_{T-1}, в которой никому из игроков не удается повлиять на выбор партнеров. Угроза наказания не действует, так как о нарушении соглашения u^* становится известно в момент времени $t = T$, когда игра уже закончилась и уже ничего нельзя изменить.

Предположение Д3 становится излишним, если следующим образом изменить классы стратегий игроков, а именно: в игре Γ_{T-1} все партнеры могут применять смешанные стратегии [2]. От краевого эффекта можно избавиться, изменив правила игры, считая, что факт $u^i_{T-1} \ne u^{*i}_{T-1}$ становится мгновенно известным всем игрокам.

2.3 Существование равновесия Нэша

В соответствии с предположением Д1 рассмотрим траекторию движения $X^* = \{(t, x^*_t), t = t_0, \ldots, T\}$ системы (1) при $u(t,x) = u^*_t$. Через $j(I)$ обозначим игрока-нарушителя программы u^*_t, идентифицируемого всеми игроками $i \ne j$ согласно Д3.

Теорема 1. *Пусть выполняются допущения* Д1–Д3. *Тогда в позиционной игре многих лиц существует равновесие Нэша, равное*

$$\tilde{u}^{*i} = u^{*i}_t(x, I) = \begin{cases} u^*_t , & x = x^*_t ; \\ \tilde{u}^{\text{н}i}_{t,x}(j(I)) , & x \ne x^*_t , \end{cases} \qquad (7)$$

$t \in \{t_0, \ldots, T-2\}$, *для всех* $i = 1, 2, \ldots, n$.

Доказательство. По условию теоремы соотношения (7) определяют допустимую ситуацию игры. Согласно лемме 1 траектория X^* стабильна относительно множества $S(u^*)$. Придерживаясь программного управления u^*_t, игроки получают выигрыши, равные $w^{*i} = f^i(x^*_T)$, $i = 1, 2, \ldots, n$. Пусть теперь кто-то из них изменил свою стратегию и вместо $\tilde{u}^{*j}$ с момента времени t, $t \in \{t_0, \ldots, T-2\}$, начал применять другую допустимую стратегию $\tilde{u}^j \ne \tilde{u}^{*j}$. Зная позицию $(t+1, \tilde{x}_{t+1}) \in S$, все партнеры $i \ne j$ это обнаруживают и согласно предположению Д3 идентифицируют нарушителя $j = j(I)$. В соответствии с определением стратегий $\tilde{u}^{*i}$ (см. (7)) с момента $t+1$ игроки $i \ne j$ применяют контрстратегию $\tilde{u}^{\text{н}i}(j)$. Оценим выигрыш j-го игрока в этой ситуации. Согласно предположению Д1 и последнему неравенству из доказательства леммы 1, примененному к $\overline{u} = u^*$, получаем:

$$\tilde{w}^j = f^j(\tilde{x}_T) \le M^i(T-1, \tilde{x}_{T-1}) \le$$

$$\le M^i(t, \tilde{x}_t) \le \overline{M}^i(u^*) \le f^j(x^*_T) = w^{*j} .$$

Осталось проанализировать случай, когда отход от программного управления u^* состоится в предпоследний момент времени $T-1$. Тогда разыгрывается игра Γ_{T-1}, которая по предположению Д3

имеет равновесие Нэша, совпадающее с выбором управления $\tilde{u}^*_{T-1}$. И здесь $\tilde{w}^j \le \tilde{w}^{*j}$ по определению равновесия (3). Итак, доказано выполнение неравенств (3) при всех $j = 1, 2, \ldots, n$ в ситуации (7). Следовательно, она равновесна по Нэшу.

Применим теорему 1 к следующей экономической модели, параметры которой, считаем, удовлетворяют условию:

$$\exp(C-1) + \exp(n(C-U)) \le \frac{1}{2} + \exp\frac{n}{2}\,, \qquad U^i = U\,. \quad (8)$$

Пример 1. Модель свободного сырьевого рынка. Имеются добывающие компании $i = 1, 2, \ldots, n$, которые в моменты времени $t = 0, 1, \ldots, T-1$ могут выходить на рынок с предложением сырья. Объемы продаж u^i ограничены количеством добываемых ресурсов U^i, $U^i > 1$. Реализация сырья происходит по цене $c(u)$, зависящей от предложения товара $u = \sum_i u^i$ в момент торгов. Каждая компания заинтересована в максимизации своей суммарной прибыли.

Пусть x^i_t — прибыль компании $i = 1, 2, \ldots, n$, извлеченная к моменту времени $t = 0, 1, \ldots, T-1$. Тогда эти величины изменяются согласно разностным уравнениям:

$$x^i_{t+1} = x^i_t + c(u_t)u^i_t\,,\ \ x^i_0 = 0\,,\ \ t = 0, 1, \ldots, T-1\,.$$

Каждая компания стремится максимизировать общую прибыль $f^i(\tilde{u}) = x^i_T$. Это игрок, выбирающий свои допустимые позиционные стратегии $\tilde{u}^i = u^i_t(x, I)$, $0 \le u^i_t \le U^i$, где $x = (x^1, \ldots, x^n)$. Пусть зависимость цены товара от предложения равна

$$c(u) = \begin{cases} 1\,, & 0 \le u \le v_0\,; \\ \exp\left(-n + v_0\right)\,, & u > v_0\,. \end{cases}$$

Объем предложения $u \le v_0$ характеризует сбалансированность спроса и предложения на рынке, обеспечивающую стабильную цену сырья, равную единице. Будем считать, что константа $C = v_0/n$ удовлетворяет неравенству $1/2 \le C < 1$.

Равновесие Нэша построим, исходя из результатов игроков, равных

$$x^{*i}_T \equiv \frac{v_0}{n}\,(T-1) + \exp\left(-n + v_0\right)\,,$$

достижимых на траектории системы, управляемой программой вида ($i = 1, 2, \ldots, n$):

$$u^{*i}_t = \begin{cases} \dfrac{v_0}{n}\,, & t = 0, 1, \ldots, T-2\,; \\ 1\,, & t = T-1\,. \end{cases}$$

В любой момент времени $t = 1, 2, \ldots, T-1$ всякий игрок может самостоятельно обнаружить факт нарушения общего плана действий. Если $x_t \ne x^*_t$, то указанное событие случилось при $\tau = t-1$. За это следует наказать нарушителя j. Начиная с момента времени $\tau = t$ участники игры $i \ne j$ будут использовать свои позиционные стратегии наказания, имеющие вид $u^{\text{н}i}_\tau(x; j) \equiv U^i$, $x \ne x^*$. Последние одинаковы для всех $j \ne i$, поэтому даже не нужно идентифицировать нарушителя! В статической игре Γ_{T-1} имеется равновесие Нэша, равное $u^*_{T-1} = 1$.

Выполнены все условия теоремы 2.1, так что равновесие $\tilde{u}^*$ в позиционных стратегиях имеет вид (7). Справедливость сказанного обусловлена действенностью наказания, уменьшающего «оговоренный» выигрыш игрока x^{*j}_T. Это происходит при соотношении (8) между параметрами модели. Сравним найденное позиционное решение игровой задачи с программным равновесием $(\forall\, i) u^{\text{p}i}_t \equiv \equiv 1$, дающим результат игры $x^{\text{p}i}_T = T\exp(-n + v_0)$. Ввиду справедливости неравенств $(\forall\, i) x^{\text{p}i}_T < x^{*i}_T$, ситуация u^{p} не эффективна по Парето.

3 Разрешающее правило

Предположение Д2 излишне ($I = \emptyset$) для игры двух лиц и для игр специального вида, в которых система (1) распадается на подсистемы:

$$x^i_{t+1} = \overline{g}^i(t, x_t, \tilde{u}^i)\,, \qquad i = 1, 2, \ldots, n\,,\ \ t = t_0, t_0+1, \ldots, T-1\,.$$

Каждому игроку достаточно обнаружить выполнение неравенства $x^j_t \ne x^{*j}_t$ и приступить к наказанию j-го игрока.

В общем случае придется увеличить информированность игроков. Целесообразно расширить фазовое пространство и вместо исходного многошагового уравнения (1) перейти к управляемой системе, построенной с помощью программы u^*_t:

$$z^i_{t+1} = g\left(t, z^i_t, u^*_t | \tilde{u}^i\right)\,, \qquad i = 1, 2, \ldots, n\,,\ \ t = t_0, t_0+1, \ldots, T-1\,. \quad (9)$$

В полученной модели все действия игроков полностью контролируются. Они выбирают свои допустимые стратегии в классе отображений $\tilde{u}^i : (t, z) \to u^i(t, z)$, так что $I = \{(t, z)\} \ne \emptyset$. Нарушитель $j = j(I)$ соглашения u^*_t идентифицируется с помощью условия $x^{*j}_t \ne z^j_t$. Если все игроки придерживаются выбора u^*_t, то траектории движения

всех систем (9), $i = 1, 2, \ldots, n$, и (1) совпадают. При нарушении программы u_t^{*j} движение только одной из подсистем (9) совпадает с движением исходной системы.

Пример 2. Позиционные игры с влиятельными игроками. Пусть в игре имеется n_1 игроков, каждый из которых $k = 1, 2, \ldots, n_1$ оказывает влияние на группу игроков $L(k)$, непосредственно воздействуя на управляемые системы вида:

$$x_{t+1}^l = \overline{g}^l \left(t, x_t^l, \tilde{u}^l, \tilde{u}^k\right) , \quad l \in L(k) , \ \ t = t_0, t_0 + 1, \ldots, T - 1 . \quad (10)$$

У каждого игрока $k = 1, 2, \ldots, n_1$ имеется свой объект управления

$$x_{t+1}^k = \overline{\overline{g}}^k \left(t, x_t, \tilde{u}, \tilde{v}\right) , \quad (11)$$

где

$$\tilde{u} = \left(u^k, k = 1, 2, \ldots, n_1\right) ; \ \ \tilde{v} = \left(\tilde{u}^l, l \in L(k)\right) ,$$

с состояниями $x = ((x^k, k = 1, 2, \ldots, n_1), (x^l, l \in \in L(k)))$, зависящими от управляющих воздействий «подчиненных» игроков $l \in L(k)$.

В игровой задаче (10), (11) с платежными функциями игроков (2) и с принципом оптимальности (3) выделено n_1 иерархических структур [1]. Для реализации ситуации равновесия из теоремы 1 можно воспользоваться следующим разрешающим правилом, предполагающим наличие у игроков лишь *частичного* знания позиций игры. Так, каждый влиятельный игрок $k = 1, 2, \ldots, n_1$ должен иметь информацию о текущих состояниях систем (11) всех влиятельных партнеров и систем (10) всех «подчиненных» ему игроков $l \in L(k)$. Все игроки $l \in L(k)$ должны знать текущие состояния $(x^k, (x^l, l \in L(k)))$ своей иерархической подсистемы. Равновесие (7) поддерживается идентификацией нарушителя $j(x_t)$ программы u_t^* согласно правилу:

$$j(x_t) = \begin{cases} l, & (\exists_1 k)\,(\exists_1 l \in L(k))\, x_t^l \neq x_t^{*l},\ x_t^k \neq x_t^{*k}; \\ k, & (\exists_1 k)\,(\exists_{\geq 1} l \in L(k))\, x_t^l \neq x_t^{*l},\ x_t^k \neq x_t^{*k}. \end{cases}$$

При необходимости каждый влиятельный игрок $k = 1, 2, \ldots, n_1$ наказывает одного из подчиненных $l \in L(k)$ либо участвует в наказании одного из влиятельных участников конфликта $k' = j(x_t)$, $k' \neq k$. Каждый подчиненный игрок $l \in L(k)$ угрожает наказанием лишь игрокам из k-й иерархической подсистемы.

Пример 3. Позиционные игры с различимыми воздействиями игроков на управляемую систему. В условиях теоремы 1 заменим допущение Д2 следующим предположением. Для любых значений $u^k \in U^k(t, x_t^*) \backslash \{u_t^{*k}\}$, $k = i, j, t \in T$,

$$g\left(t, x_t^*, u_t^* | \, u^i\right) \neq g\left(t, x_t^*, u_t^* | \, u^j\right) , \ \ i \neq j . \quad (12)$$

Пусть один из игроков в момент времени t не выполняет программу u_t^*. Тогда система (1) может быть переведена лишь в одну из позиций $(t, x) \in \in \bigcup_i D_t^i(u^*)$. Согласно (12),

$$D_t^i\,(u^*) \cap D_t^j\,(u^*) = \emptyset, \ \ i \neq j .$$

Это позволяет ввести отображение $I = I_{t,x} : \{0, 1, \ldots, n\} \to \{0, 1, \ldots, n\}$, задающее информационное множество игрока в зависимости от текущей позиции, в которой пребывает система (1), и от значения k этой функции, полученного в предыдущей позиции:

$$I_{t_0,x_0}(k) \equiv 0 ;$$

$$I_{t,x}(k) = \begin{cases} 0, & x = x_t^* ; \\ k, & x \neq x_t^*,\ k \neq 0 ; \\ j, & x \neq x_t^*,\ k = 0,\ (t, x) \in D_t^j(u^*), \\ & \qquad\qquad t > t_0 . \end{cases}$$

Нарушитель соглашения u_t^* выделяется ненулевым номером $I_{t,x}(k)$. Нулевое же значение этой функции отвечает случаю, когда все соблюдают договоренность, применяя программное управление u_t^*. Итак, в формуле равновесия (7) $j(I) = I_{t,x}$.

4 Заключение

По сравнению с дифференциальными играми многошаговые динамические задачи допускают более простое исследование и численное решение. В теореме 1 обосновано существование ситуации равновесия Нэша в позиционной игре, в которой угроза наказания реализуется при необходимости с некоторым запаздыванием. Приведены примеры разрешающих правил идентификации нарушителя совместного программного поведения. Выделены специальные классы игр, в которых это правило не предполагает сбора дополнительной информации о ходе игры.

Литература

1. *Моисеев Н. Н.* Элементы теории оптимальных систем. — М.: Наука, 1975. 527 с.
2. *Гермейер Ю. Б.* Игры с непротивоположными интересами. — М.: Наука, 1976. 326 с.
3. *Кононенко А. Ф.* Структура оптимальной стратегии в управляемых динамических системах // Ж. вычисл. мат. мат. физ., 1980. Т. 20. № 5. С. 1105–1116.

4. *Васильев Н. С.* Коалиционно устойчивые эффективные равновесия в моделях коллективного поведения с обменом информацией // Информатика и её применения, 2015. Т. 9. Вып. 2. С. 2–13.
5. *Айзекс Р.* Дифференциальные игры / Пер. с англ. — М.: Мир, 1967. 480 с. (*Isaacs R.* Differntial games. — New York, NY, USA: John Wiley and Sons, 1965. 416 p.)
6. *Понтрягин Л. С.* Линейные дифференциальные игры преследования // Мат. сб., 1980. Т. 112(154). Вып. 3(7). С. 307–330.
7. *Никольский М. С.* Первый метод Л. С. Понтрягина в дифференциальных играх. — М.: МГУ, 1984. 64 с.
8. *Красовский Н. Н., Субботин А. И.* Позиционные дифференциальные игры. — М.: Наука, 1974. 458 с.
9. *Кононенко А. Ф.* О многошаговых конфликтах с обменом информацией // Ж. вычисл. мат. мат. физ., 1977. Т. 17. № 4. С. 922–931.
10. *Чистяков Ю. Е.* Задача о ситуациях равновесия по Нэшу в игре многих лиц с памятью // Прикладная математика и механика, 1987. № 2. С. 201–214.
11. *Кононенко А. Ф., Чистяков Ю. Е.* О равновесных позиционных стратегиях в дифференциальных играх многих лиц // ДАН СССР, 1988. Т. 299. № 5. С. 1053–1056.
12. *Горелик В. А., Горелов М. А., Кононенко А. Ф.* Анализ конфликтных ситуаций в системах управления. — М.: Радио и связь, 1991. 286 с.
13. *Горелов М. А., Кононенко А. Ф.* Динамические модели конфликтов II. Равновесия // Автоматика и телемеханика, 2014. Т. 75. № 12. С. 56–77.
14. *Жуковский В. И.* Введение в дифференциальные игры при неопределенности. Равновесие по Нэшу. — М.: URSS, 2010. 168 с.
15. *Петросян Л. А.* Теория игр. — СПб: БХВ-Петербург, 2012. 432 с.

Поступила в редакцию 22.02.17

PARTICIPANTS' INFORMATION AWARENESS AND EXISTENCE OF EQUILIBRIUM IN POSITIONAL ITERATION GAMES OF MANY PLAYERS

N. S. Vasilyev

N. E. Bauman Moscow State Technical University, 5 Baumanskaya 2nd Str.,Moscow 105005, Russian Federation

Abstract: In positional games, dynamical decision-making models are studied for the situation when there is a conflict of interests and participants know the current position of the game. Each player is able to control the dynamical system partially. The control strategy chosen by a player is a function defined on the system's phase space. Players check the system's movement and obtain an implicit idea about strategies applied by their partners. The principle of players' rational behavior consists in trying to achieve the situation of Nash equilibrium. It is proved that an equilibrium can be reached as a result of collective efforts to choose the system's general program control. Stability of the solution is reached by using the threat of punishment to those who refuse to fulfill the program. Positions control and some additional information give players the possibility to identify the guilty player. Then, after a delay, he/she is punished by all other players. The theorem of existence of an equilibrium is applied to economic and mathematical model.

Keywords: system; differential game; positional iteration game; program control; positional strategy; counter strategy; punishment strategy; guaranty strategy; Nash equilibrium situation; Pareto effectiveness

DOI: 10.14357/19922264170205

References

1. Moiseev, N. N. 1975. *Elementy teorii optimal'nykh sistem* [Elements of the optimal systems theory]. Moscow: Nauka. 527 p.
2. Germeyer, Yu. B. 1976. *Igry s neprotivopolozhnymi interesami* [Games with nonantogonistic interests]. Moscow: Nauka. 326 p.
3. Kononenko, A. F. 1980. Struktura optimal'noy strategii v upravlyaemykh dinamicheskikh sistemakh [Optimal strategy structure in control dynamic systems]. *Zh. Vyichislitel'noy matematiki i matematicheskoy fiziki* [Comput. Math. Math. Phys.] 20(5):1105–1116.
4. Vasilyev, N. S. 2014. Koalitsionno ustoychivye effektivnye ravnovesiya v modelyakh kollektivnogo povedeniya s obmenom informatsiey [Coalitionally stable effective equilibria in collective behavior models with information exchange]. *Informatika i ee Primeneniya — Inform. Appl.* 8(2):2–13.
5. Isaacs, R. 1965. *Differential games*. New York, NY: John Wiley and Sons. 416 p.

6. Pontryagin, L. S. 1980. Lineynye differentsial'nye igry presledovaniya [Linear differential games of pursuit]. *Mat. sb.* [Mathematical Collection] 112(3):307–330.
7. Nikol'skiy, M. S. 1984. Pervyy metod L. S. Pontryagina v differentsial'nykh igrakh [L. S. Pontryagin's first method in differential games]. Moscow: MGU. 64 p.
8. Krasovskiy, N. N., and A. I. Subbotin. 1974. *Pozitsionnye differentsial'nye igry* [Positional differential games]. Moscow: Nauka. 458 p.
9. Kononenko, A. F. 1977. O mnogoshagovykh konfliktakh s obmenom informatsiey [On many steps conflicts with information exchange]. *Zh. Vychislitel'noy matematiki i matematicheskoy fiziki* [Comput. Math. Math. Phys.] 17(4):922–931.
10. Chistyakov, Yu. E. 1987. Zadacha o situatsiyakh ravnovesiya po Neshu v igre mnogikh lits s pamyat'yu [Nash's equilibrium situation problem in the game of many players with memory]. *Prikladnaya matematika i mekhanika* [Appl. Math. Mech.] 2:201–214.
11. Kononenko, A. F., and Yu. E. Chistyakov. 1988. O ravnovesnykh pozitsionnykh stpategiyakh v diffepentsial'nykh igpakh mnogikh lits [On equilibrium positional strategies in many players' differential games]. *Dokl. USSR Akad. Sci.* 299(5):1053–1056.
12. Gorelik, V. A., Gorelov M. A., and A. F. Kononenko. 1991. *Analiz konfliktnykh situatsiy v sistemakh upravleniya* [Analysis of conflict situations in control systems]. Moscow: Radio i svyaz. 286 p.
13. Gorelov, M. A., and A. F. Kononenko. 2014. Dynamic models of conflicts. II. Equilibria. *Automat. Rem. Contr.* 75(12):2135–2151.
14. Zhukovskiy, V. I. 2010. *Vvedenie v differentsial'nye igry pri neopredelennosti. Ravnovesiya po Neshu* [Introduction in differential games under uncertainty. Nash's equilibria]. Moscow: URSS. 168 p.
15. Petrosyan, L. A. 2012. *Teoriya igr* [Theory of games]. St. Peterburg: BHV-Peterburg. 432 p.

Received February 22, 2017

Contributor

Vasilyev Nikolai S. (b. 1952) — Doctor of Science in physics and mathematics, professor, N. E. Bauman Moscow State Technical University, 5 Baumanskaya 2nd Str.,Moscow 105005, Russian Federation; nik8519@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 50–58

МОДЕЛИРОВАНИЕ ОТНОШЕНИЯ СИГНАЛ/ИНТЕРФЕРЕНЦИЯ В МОБИЛЬНОЙ СЕТИ СО СЛУЧАЙНЫМ БЛУЖДАНИЕМ ВЗАИМОДЕЙСТВУЮЩИХ УСТРОЙСТВ*

Ю. В. Гайдамака[1], Ю. Н. Орлов[2], Д. А. Молчанов[3], А. К. Самуйлов[4]

Аннотация: Целью исследования является анализ отношения сигнал/интерференция (SIR, signal-to-interference ratio) при прямом взаимодействии устройств в мобильных сетях 5-го поколения (5G) с учетом перемещения приемопередающих устройств в зоне обслуживания. Величина SIR на приемнике ассоциированной пары исследуется как изменяющийся во времени случайный процесс, а математическая модель движения задана кинетическим уравнением с учетом скорости перемещения устройств, их пространственной плотности и максимально допустимого радиуса взаимодействия. Задача численного анализа решается методом имитационного моделирования. В качестве функционала величины SIR исследуется плотность распределения случайной величины (СВ) длительности периода наличия (отсутствия) связи. Показано, что вероятность обрыва связи логарифмически растет с увеличением как скорости перемещения взаимодействующих устройств, так и их числа в заданной зоне обслуживания.

Ключевые слова: беспроводная сеть; отношение сигнал/интерференция; взаимодействие устройств; стохастическая геометрия; модель движения; кинетическое уравнение; показатель эффективности; вероятность обрыва связи; длительность периода наличия связи

DOI: 10.14357/19922264170206

1 Введение

Методы анализа SIR как ключевого показателя, характеризующего качество предоставления услуг в беспроводной сети [1, 2], разрабатывались начиная примерно с 2005 г. Хотя большинство исследователей используют для анализа этой характеристики методы имитационного моделирования, известны также точные [3–6] и приближенные [7–9] аналитические методы расчета SIR, использующие аппарат стохастической геометрии и предназначенные для анализа характеристик сети в случае неподвижных абонентов [10, 11]. Авторам не известны публикации, где проводится анализ SIR, меняющегося во времени, а именно в этих предположениях следует исследовать интерференцию при анализе показателей эффективности беспроводных сетей с перемещающимися абонентами. Первые шаги в этом направлении предприняты в [12].

В данной статье отношение SIR на приемнике ассоциированной пары исследуется как случайный процесс, состояние которого зависит от скорости движения приемопередающих устройств, плотности размещения передатчиков в зоне обслуживания и расстояния в ассоциированной паре приемник–передатчик. Для случая однородных абонентов, перемещающихся на плоскости, получены распределения длительности периода наличия и периода отсутствия связи между устройствами пары. В исследуемой модели задано кинетическое уравнение относительно плотности распределения случайных приращений координат для описания движения абонента, плотность абонентов, а также максимальный допустимый радиус взаимодействия для прямых соединений. На основе построенной в разд. 2 модели прямых соединений в разд. 3 разработан метод расчета отношения сигнал/интерференция, а в разд. 4 определены показатели эффективности модели. С использованием траекторий SIR во времени в разд. 5 численно исследованы плотности распределения длительностей периода наличия и периода отсутствия связи.

2 Системная модель

Рассматривается сценарий, где в ограниченной зоне обслуживания V, представляющей собою область в M-мерном пространстве, перемещаются носители идентичных приемопередающих

*Публикация подготовлена при финансовой поддержке Минобрнауки России (№ 2.3397.2017).

[1]Российский университет дружбы народов; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, gaydamaka_yuv@rudn.university

[2]Институт прикладной математики им. М. В. Келдыша Российской академии наук, yuno@kiam.ru

[3]Российский университет дружбы народов, molchanov_da@rudn.university

[4]Российский университет дружбы народов, samuylov_ak@rudn.university

устройств. Такой сценарий пригоден, например, для описания хаотичного перемещения зрителей на городской площади, на стадионе во время массовых мероприятий ($M = 2$) или для описания движения покупателей в многоуровневом торговом центре ($M = 3$). Для исследования интерференции на приемнике произвольной ассоциированной пары взаимодействующих устройств, создаваемой передатчиками других пар, рассмотрено N пар устройств, взаимодействующих напрямую (D2D, device-to-device) на основе одной из чувствительных к интерференции (interference-limited) технологий, например LTE, Release 13 или Wi-Fi Direct Connect [13, 14]. Перемещение передатчиков сочетает целеполагающее поступательное движение и хаотическое блуждание и определяется кинетическим уравнением типа Фоккера–Планка с заданной скоростью сноса u и коэффициентом диффузии α. Ассоциированный приемник движется вместе с передатчиком внутри окружности максимально допустимого радиуса взаимодействия d с центром в точке расположения передатчика. Мощность сигнала на приемнике определяется из стандартной модели распространения сигнала $\phi(r) = Ar^{-\gamma}$, где r — расстояние в паре приемник–передатчик; A — константа, учитывающая излучаемую мощность и коэффициенты усиления приемной и передающей антенны; γ — коэффициент распространения сигнала.

Для оценки интерференции на приемнике используется SIR, вычисляемое по формуле:

$$\mathrm{SIR} = \frac{\phi(r_0)}{\sum_{n=1}^{N-1} \phi(d_n)}, \qquad (1)$$

где r_0 — расстояние между приемником и передатчиком в исследуемой паре; d_n — расстояние между приемником из исследуемой пары и передатчиком из n-й интерферирующей пары.

Также задан порог SIR*, определяющий минимальное значение SIR, необходимое для поддержания соединения. При падении SIR на приемнике ниже порогового значения SIR* связь в паре прерывается до момента, когда SIR вновь превысит данный порог. Таким образом, задачей исследования является нахождение плотностей распределения длительностей периодов наличия и отсутствия связи.

3 Метод расчета отношения сигнал/интерференция

Моделирование траекторий проводится для случая нестационарного блуждания на задаваемом временно́м горизонте $t \in [1, T]$ в дискретном времени с единичным шагом по времени, следуя результатам [15–17]. На каждом шаге $k = 1, 2, \ldots, T$ для каждого устройства генерируется приращение координат в соответствии с решением уравнения Фоккера–Планка, которое для случая моделирования одномерного блуждания ($M = 1$) имеет вид:

$$\frac{\partial f(x,t)}{\partial t} + \frac{\partial}{\partial x}\left(u(x,t) f(x,t)\right) - \frac{\alpha(t)}{2} \frac{\partial^2 f(x,t)}{\partial x^2} = 0\,. \qquad (2)$$

Здесь $f(x,t)$ — плотность распределения приращений координат x положения устройств в момент времени t, а параметры уравнения — скорость сноса $u(x,t)$ и нестационарный в общем случае неотрицательный коэффициент диффузии $\alpha(t)$ — в данной работе были построены по моделям типичных нестационарных процессов, описывающих изменения положений случайно блуждающих объектов, обсуждаемых в [17]. Для полученного ансамбля траекторий расстояние $r_{ij}(k)$ между точками i с координатами $R^i(k)$ и j с координатами $R^j(k)$, которые находятся на разных траекториях в некоторой области V в M-мерном пространстве в один и тот же момент времени $t = k$, вычисляются по формуле:

$$r_{ij}^2(k) = \sum_{m=1}^{M} \left(R_m^i(k) - R_m^j(k)\right)^2 .$$

Введем функцию от расстояния между двумя точками $\phi_{ij} = \phi(r_{ij})$, соответствующую определенной выше стандартной модели распространения сигнала $\phi(r) = Ar^{-\gamma}$. Для произвольной пары точек, например $i = 1$ и $j = 2$, функционал (1) SIR от расстояния $r_{12}(k)$ во введенных обозначениях в момент времени $t = k$ определяется формулой:

$$S\left(r_{12}(k), N\right) = \frac{\phi_{12}(k)}{\sum_{j=3}^{N} \phi_{1j}(k)}\,. \qquad (3)$$

С точки зрения анализа устойчивости соединения основной задачей является вычисление суммы мощностей сигналов в знаменателе формулы (3), поскольку мощность сигнала в числителе ограничивается максимально допустимым радиусом взаимодействия d.

Изложим далее общий сценарий численного моделирования величины SIR на одном из N взаимодействующих в области V устройств, например на приемнике, расположенном в точке $i = 1$.

Сумма в знаменателе формулы (3) за вычетом величины ϕ_{12} представляет собой умноженное на

$N-1$ среднее по ансамблю значение функции связи с первой точкой, определяемое как

$$U\left(r_{12}(k),k\right)=\int\limits_V \phi\left(|r_{12}(k)-r'|\right) f\left(r',k\right)\,dr'\,,$$

где $f(r',k)$ есть плотность распределения N точек в области V в момент времени $t=k$ в соответствии с решением кинетического уравнения (2).

Положим $r(k)=r_{12}(k)$ и рассмотрим все остальные точки в системе отсчета, связанной со второй точкой. Тогда функционал (3) примет вид:

$$S(k)\equiv S(k;u,\alpha)=\frac{\phi(r(k))}{(N-1)U(r(k),k)-\phi(r(k))}=\\ =\frac{\phi(r(k))}{NU(r(k),k)}+o\left(\frac{1}{N}\right)\,.$$

Отметим, что зависимость функции $U(r(t),t)$ от времени двоякая: во-первых, эта зависимость неявно определяется тем, что положение точек получено генерацией их из распределения, построенного как решение кинетического уравнения (2) в момент времени $t=k$; во-вторых, эта зависимость определяется и самой плотностью функции распределения $f(r,t)$. Учитывая это, будем для краткости писать далее

$$U\equiv U(r,t)=\int\limits_V \phi\left(|r-r'|\right) f\left(r',t\right)\,dr'\,. \qquad (4)$$

Заметим теперь, что выбранные выше точки 1 и 2, связь между которыми изучается с помощью функционала (4), движутся по тому же статистическому закону, что и прочие точки системы, т. е. в каждый момент времени они случайно перемещаются в пространстве в соответствии с тем, какая конкретная выборочная траектория реализована для моделирования движения каждой из них. Поскольку выбранная пара точек произвольна, то качество связи между двумя точками определяется средним по ансамблю значением $q(t)$ функционала $S(r(t),t)$:

$$q(t)=\frac{1}{N}\int\limits_V \frac{\phi(r)}{U(r,t)}\,f(r,t)\,dr\,. \qquad (5)$$

Выведем уравнение эволюции среднего значения функционала SIR, т. е. величины (5). Из (4) имеем:

$$N\frac{dq}{dt}=\int\limits_V \frac{\phi(r)}{U(r,t)}\,\frac{\partial f(r,t)}{\partial t}\,dr-\\ -\int\limits_V \frac{\phi(r)}{U^2(r,t)}\,\frac{\partial U(r,t)}{\partial t}\,f(r,t)\,dr\,. \qquad (6)$$

Считая, что на границе области функция распределения обращается в ноль, для производной $\partial U/\partial t$ получаем после подстановки производной $\partial f/\partial t$ из уравнения (2), интегрирования по частям и замены возникающей после этого действия производной функции $\phi(|r-r'|)$ по r' на производную по r следующее выражение:

$$\frac{\partial U}{\partial t}=\frac{\alpha}{2}\,\Delta U-\operatorname{div}\mathbf{J}\,,$$

где

$$\mathbf{J}=\int\limits_V \phi\left(|r-r'|\right)u\left(r',t\right)f\left(r',t\right)\,dr'\,.$$

Далее первое слагаемое в (6) с использованием (2) преобразуется к виду:

$$\int\limits_V \frac{\phi(r)}{U(r,t)}\,\frac{\partial f(r,t)}{\partial t}\,dr=\\ =-\int\limits_V \frac{\phi(r)}{U(r,t)}\operatorname{div}\left(uf\right)dr+\frac{\alpha}{2}\int\limits_V \frac{\phi(r)}{U(r,t)}\,\Delta f\,dr\,,$$

где Δ обозначен оператор Лапласа. После интегрирования по частям получаем:

$$\int\limits_V \frac{\phi}{U}\,\frac{\partial f}{\partial t}\,dr=\int\limits_V f\left(u\nabla+\frac{\alpha}{2}\,\Delta\right)\left(\frac{\phi}{U}\right)\,dr\,.$$

В результате уравнение (6) принимает вид:

$$N\frac{dq}{dt}=\int\limits_V\left(\left(u\nabla+\frac{\alpha}{2}\,\Delta\right)\left(\frac{\phi}{U}\right)-\right.\\ \left.-\frac{\phi}{U^2}\left(\frac{\alpha}{2}\,\Delta U+\operatorname{div}\mathbf{J}\right)\right)f(r,t)\,dr\,.$$

Это уравнение сложным нелинейным образом зависит от функции распределения точек, т. е. от плотности ансамбля их выборочных траекторий.

Следовательно, актуальной задачей дальнейших исследований является численное моделирование статистик, связанных с распределением величины $S(r(t),t)$ зависимости SIR от расстояния до точки 2, и величины $q(t)$, которая представляет собой среднее значение функционала $S(r(t),t)$ по ансамблю траекторий.

4 Показатели эффективности

Для определения показателей эффективности передачи данных в канале между взаимодействующими устройствами ассоциированной пары в зоне обслуживания на рис. 1 приведены примеры численного моделирования SIR для некоторой пары устройств по описанной методике.

Для оценки вероятности обрыва связи между взаимодействующими устройствами следует учитывать количество выбросов SIR ниже порога SIR* на длительном интервале наблюдения. Кроме вероятности обрыва связи интересными для анализа исследуемой системы являются характеристики СВ τ_l^- — длительности периода отсутствия и СВ τ_l^+ — длительности периода наличия связи между устройствами, которые показаны на рис. 1 вместе с необходимыми для их определения моментами t_l^- и t_l^+ пересечения графиком величины SIR порогового значения отношения SIR*. В первую очередь интерес представляет плотность распределения СВ τ_l^- и τ_l^+ в предположении об их независимости в совокупности и одинаковом распределении.

В качестве примеров могут быть рассмотрены два типа приложений, использующих прямые соединения при предоставлении услуг — приложения в реальном времени и приложения с кэшированием данных. К приложениям в реальном времени относятся, например, набирающие популярность игровые приложения и голосовые приложения. Для таких приложений критична возможность непрерывного поддержания соединения в ассоциированной паре, поэтому ключевым показателем качества предоставления услуги является вероятность того, что SIR упадет ниже заданного порога SIR* на определенном отрезке времени (вероятность выброса величины SIR вниз). В этом случае соединение в такой ассоциированной паре прерывается и планировщиком базовой станции сети LTE или контролирующим узлом D2D соединения для передачи данных в паре приемник–передатчик должна быть выбрана новая радиочастота. Таким образом, для приложений в реальном времени длительность интервала до прерывания соединения совпадает с длительностью периода наличия связи τ_1^+ до первого обрыва связи в момент t_1^-.

В отличие от приложений в реальном времени, для кэшируемых приложений, к которым в первую очередь относится передача видео, обрыв связи не всегда приводит к прерыванию соединения. Если качество соединения в ассоциированной паре позволяет передавать данные на скорости, превышающей требуемую скорость воспроизведения, то при кратковременном отсутствии связи буферизация обеспечит непрерывное воспроизведение видео. Таким образом, для кэшируемых приложений падение SIR ниже заданного порога, приводящее к обрыву связи, не является критичным для предоставления услуги. Для таких приложений важны длительности периодов наличия и отсутствия

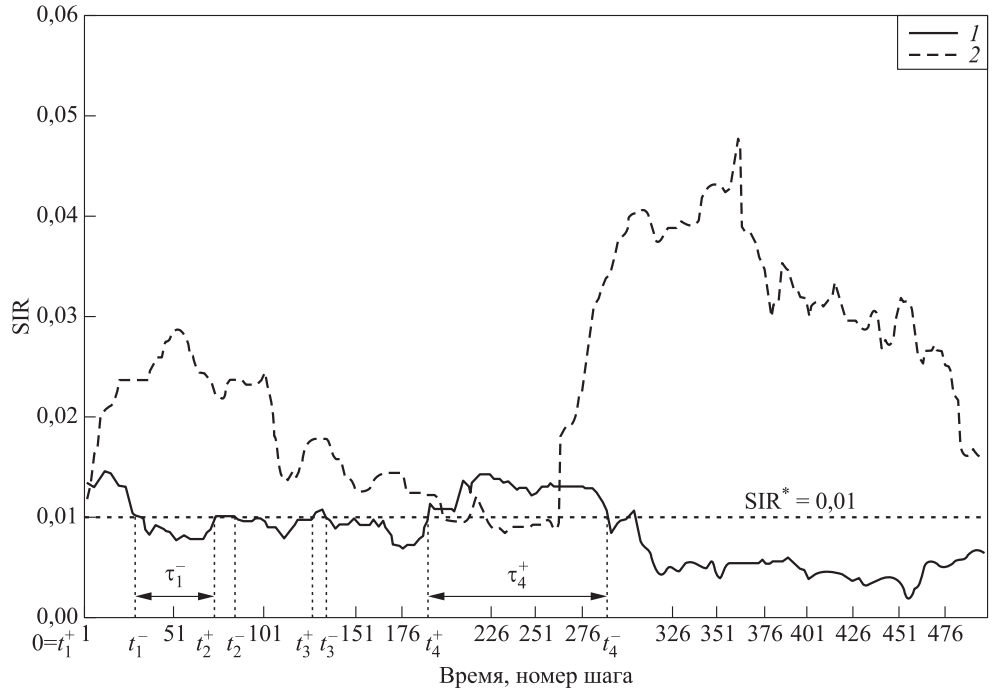

Рис. 1 Фрагмент временны́х рядов SIR для различной плотности устройств: *1* — $N = 100$; *2* — $N = 10$

связи, поскольку новая радиочастота для поддержания соединения в паре приемник–передатчик должна быть назначена, только если временное отсутствие связи не может быть компенсировано механизмом буферизации. Для кэшируемых приложений длительность интервала времени до прерывания соединения определяется моментом обрыва связи $t^-_{L(\tau^*)}$, $L(\tau^*) \geq 1$, соответствующий интервал $\tau^-_{L(\tau^*)}$ отсутствия связи после которого превысит порог τ^* — задержку, которая может быть сглажена с помощью механизма буферизации. В этом случае длительность интервала времени до прерывания соединения равна $t^-_{L(\tau^*)} + \tau^*$ и определяется формулой $\sum_{l=1}^{L(\tau^*)}(\tau_l^+ + \min(\tau_l^-, \tau^*))$, где $L(\tau^*) = \inf\left\{l:\ \tau^-_{L(\tau^*)} \geq \tau^*\right\}$.

Таким образом, основными метриками, определяющими качество функционирования кэшируемых приложений, являются характеристики СВ τ_l^+ (τ_l^-) длительности периода наличия (отсутствия) связи между взаимодействующими устройствами. Обозначим $F_{\tau^+}(x) = P\left\{\tau_l^+ < x\right\}$ ($F_{\tau^-}(x) = P\{\tau_l^- < x\}$) и $f_{\tau^+}(x) = F'_{\tau^+}(x)$ ($f_{\tau^-}(x) = F'_{\tau^-}(x)$) функции и плотности распределения этих СВ соответственно.

Задачу численного анализа в следующем разделе статьи будем решать изложенным выше методом с использованием имитационного моделирования. Анализ вероятностных характеристик будем проводить в зависимости от средней скорости движения устройств $v = \int_V u(x,t)f(x,t)\,dx$, предполагая для простоты эту скорость постоянной, а также в зависимости от числа N пар приемопередающих устройств.

5 Численный анализ

В качестве примера численного анализа рассматривается сценарий перемещения носителей устройств внутри торгового центра (см. таблицу).

Графики плотностей распределения длительности периодов наличия $f_{\tau^+}(x)$ и отсутствия $f_{\tau^-}(x)$ связи между устройствами в зависимости от средней скорости v передвижения устройств и числа N пар взаимодействующих устройств приведены на рис. 2 и 3.

Рисунки 2, *а* и 3, *а* иллюстрируют плотность $f_{\tau^+}(x)$ распределения длительности периода наличия связи при уровне порога $\mathrm{SIR}^* = 0{,}01$ для различных значений средней скорости и числа пар устройств. Следует отметить, что качественное поведение графика плотности остается неизменным для различных значений скоростей и количества пар. Кроме того, статистика свидетельствует о характерных для показательного распределения длительностях периодов наличия связи.

Аналогично периодам наличия связи, плотность $f_{\tau^-}(x)$ распределения периода отсутствия связи (см. рис. 2, *б* и 3, *б*) имеет ярко выраженный показательный характер. Эти наблюдения хорошо согласуются с теоретическими результатами, представленными в [18], где показано, что выбросы траекторий стационарного нормального случайного процесса имеют показательный характер.

Представленные результаты численного анализа плотностей распределения длительности периодов наличия и отсутствия связи позволяют сделать важные практические выводы относительно качества обслуживания прямых соединений в сетях беспроводной связи. Известно, что показательное распределение характеризуется отсутствием памяти. Таким образом, если можно утверждать о достаточно большом значении SIR в момент установления прямого соединения, то сделать обоснованный количественный вывод о продолжительности периода наличия устойчивой связи не представляется возможным вследствие характера плотности распределения периода наличия устойчивой связи. Аналогичные выводы можно сделать и о периоде отсутствия связи. Наконец, необходимо отметить, что указанное поведение характерно для выбран-

Параметры системной модели

Параметр	Значение
Область обслуживания в ($M = 2$)-мерном пространстве, V	50×50 кв. м
Число пар «приемник–передатчик», N	10, 30, 50, 100
Средняя скорость сноса, v	1, 3, 5, 10, 40 м/с
Параметр диффузии, α	2
Константа распространения, A	1
Коэффициент распространения, γ	3
Максимальное расстояние от приемника до передатчика в ассоциированной паре, d	5 м
Расстояние от приемника до передатчика в ассоциированной паре, r_0	$0 < r_0 \leq d$
Диффузия приемника в ассоциированной паре	1 м/с
Уровень отсутствия связи, SIR^*	0,01

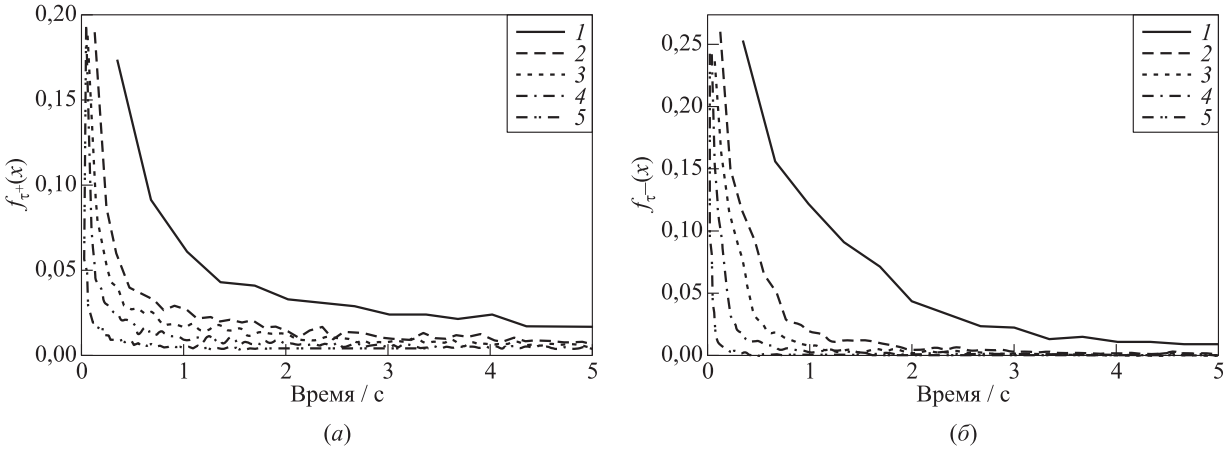

Рис. 2 Плотность распределения длительности периода наличия (*а*) и отсутствия (*б*) связи при различных скоростях v ($N = 10$): *1* — $v = 1$; *2* — 3; *3* — 5; *4* — 10; *5* — $v = 40$

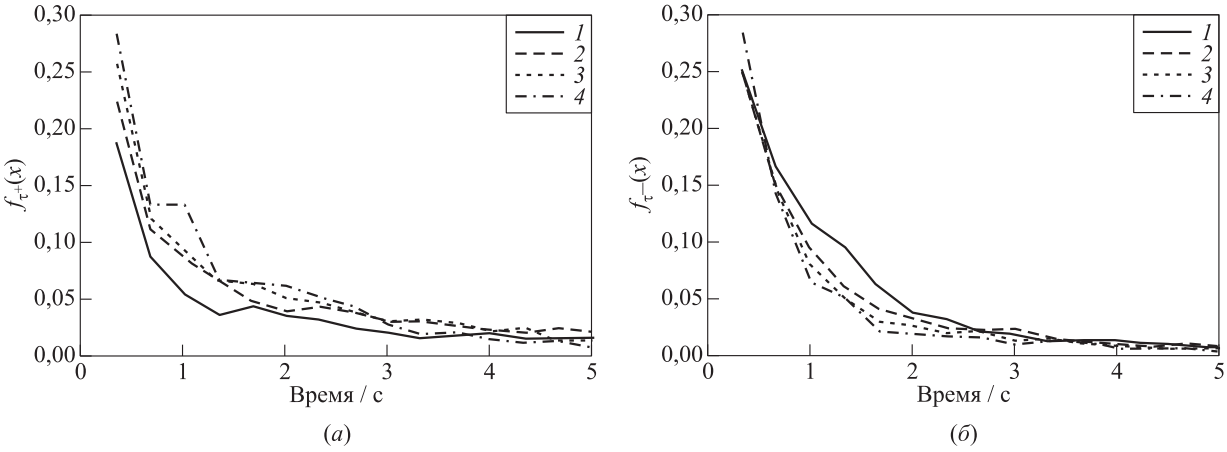

Рис. 3 Плотность распределения длительности периода наличия (*а*) и отсутствия (*б*) связи для различного числа пар N ($v = 5$ м/с): *1* — $N = 10$; *2* — 30; *3* — 50; *4* — $N = 100$

ной модели движения и может отличаться для других моделей, что является предметом дальнейших исследований.

6 Заключение

В представленной работе построена модель для анализа характеристик беспроводных соединений с технологией прямого взаимодействия устройств, которая является неотъемлемой частью сетей связи 5G.

С использованием аппарата кинетических уравнений движения получены характеристики, определяющие качество обслуживания для приложений в реальном времени и кэшируемых приложений. В частности, исследованы плотности распределения длительности периодов наличия и отсутствия связи как функции от скорости перемещения абонентов и числа пар взаимодействующих устройств.

Результаты численного анализа, представленные в работе, показывают, что вероятность обрыва связи экспоненциально уменьшается при увеличении как числа пар взаимодействующих устройств, так и скорости перемещения абонентов. Важным результатом для длительности периодов наличия и отсутствия связи является показательный характер их распределений.

Проведенный численный анализ позволяет сделать вывод о том, что поддержка приложений в реальном времени в сетях D2D возможна только в тех областях, где средняя скорость устройств не превышает нескольких метров в секунду, например внутри зданий, в пешеходных зонах на открытой местности. Кроме того, при принятии решения об установлении нового прямого соединения должно учитываться количество уже установленных соединений. Выбор оптимального значения N для некоторого порога SIR^* может быть сделан на основе

предложенной в статье методологии. Для кэшируемых приложений при дальнейших исследованиях интервал времени до прерывания соединения может быть представлен как сумма случайных величин, где число слагаемых имеет геометрическое распределение.

В дальнейшем интересно провести качественный и количественный анализ вероятностно-временны́х характеристик периодов наличия и отсутствия связи при различных типах случайного блуждания абонентов. Отдельным направлением может также стать расширение сценария на случай наличия базовых станций. Кроме того, комментарий по дальнейшим исследованиям содержится в заключении разд. 3 статьи.

В заключение авторы благодарят проф. К. Е. Самуйлова за обсуждение постановки задачи исследований.

Литература

1. *Rong Z., Rappaport T. S.* Wireless communications: Principles and practice. —1st ed. — Upper Saddle River, NJ, USA: Prentice Hall, 1996. 641 p.
2. *Andrews J. G., Claussen H., Dohler M., Rangan S., Reed M. C.* Femtocells: Past, present, and future // IEEE J. Sel. Area. Comm., 2012. Vol. 30. Iss. 3. P. 497–508. doi: 10.1109/JSAC.2012.120401.
3. *Samuylov A., Gaidamaka Yu., Moltchanov D., Andreev S., Koucheryavy Y.* Random triangle: A baseline model for interference analysis in heterogeneous networks // IEEE Trans. Veh. Technol., 2015. Vol. 65. Iss. 8. P. 6778–6782. doi: 10.1109/TVT.2016.2596324.
4. *Гайдамака Ю. В., Самуйлов А. К.* Метод расчета характеристик интерференции двух взаимодействующих устройств в беспроводной гетерогенной сети // Информатика и её применения, 2015. Т. 9. Вып. 1. С. 9–14. doi:10.14357/19922264150102.
5. *Гайдамака Ю. В., Андреев С. Д., Сопин Э. С., Самуйлов К. Е., Шоргин С. Я.* Анализ характеристик интерференции в модели взаимодействия устройств с учетом среды распространения сигнала // Информатика и её применения, 2016. Т. 10. Вып. 4. С. 2–10. doi: 10.14357/19922264160401.
6. *Samuylov A., Ometov A., Begishev V., Kovalchukov R., Moltchanov D., Gaidamaka Yu., Samouylov K., Andreev S., Koucheryavy Y.* Analytical performance estimation of network-assisted D2D communications in urban scenarios with rectangular cells // Trans. Emerg. Telecommun. Technol., 2017. Vol. 28. No. 2. P. 2999-1–2999-15. doi: 10.1002/ett.2999. (Version of record online: November 12, 2015.)
7. *Gong Z., Haenggi M.* Interference and outage in mobile random networks: Expectation, distribution, and correlation // IEEE Trans. Mobile Comput., 2014. Vol. 13. P. 337–349. doi: 10.1109/TMC.2012.253.
8. *Etezov Sh. A., Gaidamaka Yu. V., Samouylov K. E., Moltchanov D. A., Samuylov A. K., Andreev S. D., Koucheryavy E. A.* On distribution of SIR in dense D2D deployments // 22nd European Wireless Conference Proceedings. — VDE, 2016. P. 333–337.
9. *Petrov V., Komarov M., Moltchanov D., Jornet J. M., Koucheryavy Y.* Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas // IEEE Trans. Wireless Commun., 2017. Vol. 16. P. 1791–1808.
10. *Baccelli F., Blaszczyszyn B.* Stochastic geometry and wireless networks // Found. Trends Netw., 2010. Vol. 3. No. 3-4. P. 249–449. doi:10.1561/1300000006; Vol. 4. No. 1-2. P. 1–312. doi:10.1561/1300000026.
11. *Haenggi M.* Stochastic geometry for wireless networks. — Cambridge: Cambridge University Press, 2012. 298 p.
12. *Orlov Yu. N., Fedorov S. L., Samuylov A. K., Gaidamaka Yu. V., Molchanov D. A.* Simulation of devices mobility to estimate wireless channel quality metrics in 5G networks // AIP Conference Proceedings: 12th Conference (International) of Numerical Analysis and Applied Mathematics. — New York, NY, USA: AIP Publishing, 2017 (in press).
13. 3GPP LTE Release 10 & beyond (LTE-Advanced). — December 2009. ftp://www.3gpp.org/workshop/2009-12-17_ITU-R_IMT-Adv_eval/docs/pdf/REV-090006.pdf.
14. Wi-Fi Peer-to-Peer (P2P) Technical Specification v1.7. — Wi-Fi Alliance, 2010. https://www.wi-fi.org/downloads-registered-guest/Wi-Fi_P2P_Technical_Specification_v1.7.pdf/29559.
15. *Босов А. Д., Кальметьев Р. Ш., Орлов Ю. Н.* Моделирование нестационарного временного ряда с заданными свойствами выборочного распределения // Мат. моделирование, 2014. Т. 26. № 3. С. 97–107.
16. *Орлов Ю. Н., Федоров С. Л.* Генерация нестационарных траекторий временного ряда на основе уравнения Фоккера–Планка // Труды МФТИ, 2016. Т. 8. № 2. С. 126–133.
17. *Орлов Ю. Н., Федоров С. Л.* Методы численного моделирования процессов нестационарного случайного блуждания. — М.: МФТИ, 2016. 112 с.
18. *Тихонов В. И., Хименко В. И.* Выбросы траекторий случайных процессов. — М.: Наука, 1987. 304 с.

Поступила в редакцию 15.04.17

MODELING THE SIGNAL-TO-INTERFERENCE RATIO IN A MOBILE NETWORK WITH MOVING DEVICES

Yu. V. Gaidamaka[1,2], Yu. N. Orlov[3], D. A. Molchanov[3], and A. K. Samuylov[3]

[1] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation
[2] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[3] Keldysh Institute of Applied Mathematics of the Russian Academy of Sciences, 4 Miusskaya Sq., Moscow 125047, Russian Federation

Abstract: The goal of the study is to analyze the signal-to-interference ratio (SIR) for device-to-device interaction of devices communication in the 5th generation mobile networks, taking into account the movement of the receiving and transmitting devices in the service area. The SIR value at the receiver of the associated pair is studied as a time-varying random process, and the mathematical model of motion is given by a kinetic equation taking into account the given average speed of the devices, their spatial density, and the maximum allowable communication radius. The measures of performance quality were studied by numerical analysis using SIR simulation of a key channel. The measures are the following: the signal interruption probability for the receiver–transmitter pair, the probability density function of the random variables for the duration of the availability period, and the period of absence of communication. It is shown that the signal interruption probability grows logarithmically as either the average device speed or the number of devices in the service area increases.

Keywords: wireless network; signal-to-interference ratio; device-to-device; stochastic geometry; motion model; kinetic equation; performance measure; signal interruption probability

DOI: 10.14357/19922264170206

Acknowledgments

The publication was supported by the Ministry of Education and Science of the Russian Federation (project No. 2.3397.2017).

References

1. Rong, Z., and T. S. Rappaport. 1996. *Wireless communications: Principles and practice*. 1st ed. Upper Saddle River, NJ: Prentice Hall. 641 p.
2. Andrews, J. G., H. Claussen, M. Dohler, S. Rangan, and M. C. Reed. 2012. Femtocells: Past, present, and future. *IEEE J. Sel. Area. Comm.* 30(3):497–508. doi: 10.1109/JSAC.2012.120401.
3. Samuylov, A., Yu. Gaidamaka, D. Moltchanov, S. Andreev, and Y. Koucheryavy. 2015. Random triangle: A baseline model for interference analysis in heterogeneous networks. *IEEE Trans. Veh. Technol.* 65(8):6778–6782. doi: 10.1109/TVT.2016.2596324.
4. Gaydamaka, Yu. V., and A. K. Samuylov. 2009. Metod rascheta kharakteristik interferentsii dvukh vzaimodeystvuyushchikh ustroystv v besprovodnoy geterogennoy seti [The method of calculation of the characteristics of the interference of two interacting devices in a wireless heterogeneous network]. *Informatika i ee Primeneniya — Inform. Appl.* 9(1):9–14. doi:10.14357/19922264150102.
5. Gaydamaka, Yu. V., S. D. Andreev, E. S. Sopin, K. E. Samouylov, and S. Ya. Shorgin. 2016. Analiz kharakteristik interferentsii v modeli vzaimodeystviya ustroystv s uchetom sredy rasprostraneniya signala [Analysis of the characteristics of the interference in the model of interaction between devices taking into account the signal propagation environment]. *Informatika i ee Primeneniya — Inform. Appl.* 10(4):2–10. doi: 10.14357/19922264160401.
6. Samuylov, A., A. Ometov, V. Begishev, R. Kovalchukov, D. Moltchanov, Yu. Gaidamaka, K. Samouylov, S. Andreev, and Y. Koucheryavy. 2015. Analytical performance estimation of network-assisted D2D communications in urban scenarios with rectangular cells. *Trans. Emerg. Telecommun. Technol.* 28(2):2999-1–2999-15. doi: 10.1002/ett.2999. (Version of record online: November 12, 2015.)
7. Gong, Z., and M. Haenggi. 2014. Interference and outage in mobile random networks: Expectation, distribution, and correlation. *IEEE Trans. Mobile Comput.* 13:337–349. doi: 10.1109/TMC.2012.253.
8. Etezov, Sh., Yu. Gaidamaka, K. Samouylov, D. Moltchanov, A. Samuylov, S. Andreev, and E. Koucheryavy. 2016. On distribution of SIR in dense D2D deployments. *22nd European Wireless Conference Proceedings*. VDE. 333–337.
9. Petrov, V., M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy. 2017. Interference and SINR in mil-

limeter wave and terahertz communication systems with blocking and directional antennas. *IEEE Trans. Wireless Commun.* 16:1791–1808.

10. Baccelli, F., and B. Blaszczyszyn. 2010. Stochastic geometry and wireless networks. *Found. Trends Netw.* 3(3-4):249–449. doi:10.1561/1300000006; 4(1-2):1–312. doi:10.1561/1300000026.
11. Haenggi, M. 2012. *Stochastic geometry for wireless networks*. Cambridge: Cambridge University Press. 298 p.
12. Orlov, Yu. N., S. L. Fedorov, A. K. Samuylov, Yu. V. Gaidamaka, and D. A. Molchanov. 2017 (in press). Simulation of devices mobility to estimate wireless channel quality metrics in 5G networks. *AIP Conference Proceedings: 12th Conference (International) of Numerical Analysis and Applied Mathematics*. New York, NY: AIP Publishing.
13. 3GPP LTE Release 10 & beyond (LTE-Advanced). December 2009. Available at: ftp://www.3gpp.org/workshop/2009-12-17_ITU-R_IMT-Adv_eval/docs/pdf/REV-090006.pdf (accessed April 20, 2017).
14. Wi-Fi Peer-to-Peer (P2P) Technical Specification v1.7. Wi-Fi Alliance, 2010. Available at: https://www.wi-fi.org/downloads-registered-guest/Wi-Fi_P2P_Technical_Specification_v1.7.pdf/29559 (accessed April 20, 2017).
15. Bosov, A. D., R. S. Kalmetiev, and Yu. N. Orlov. 2014. Modelirovanie nestatsionarnogo vremennogo ryada s zadannymi svoystvami vyborochnogo raspredeleniya [Sample distribution function construction for nonstationary time-series forecasting]. *Matematicheskoe modelirovanie* [Mathematical Simulation] 26(3):97–107.
16. Orlov, Yu. N., and S. L. Fedorov. 2016. Generatsiya nestatsionarnykh traektoriy vremennogo ryada na osnove uravneniya Fokkera–Planka [Generation of nonstationary trajectories of the time series based on the Fokker–Planck equation]. *Trudy MFTI* [Proceedings of MIPT] 8(2):126–133.
17. Orlov, Yu. N., and S. L. Fedorov. 2016. Metody chislennogo modelirovaniya protsessov nestatsionarnogo sluchaynogo bluzhdaniya [Methods of a numerical simulation of nonstationary random walk]. Moscow: MFTI. 112 p.
18. Tikhonov, V. I., and V. I. Khimenko. 1987. *Vybrosy traektoriy sluchaynykh protsessov* [Emissions of trajectories of random processes]. Moscow: Nauka. 304 p.

Received April 15, 2017

Contributors

Gaidamaka Yuliya V. (b. 1971) — Candidate of Science (PhD) in physics and mathematics, associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; gaydamaka_yuv@rudn.university

Orlov Yurii N. (b. 1964) — Doctor of Science in physics and mathematics, professor, head of sector, Keldysh Institute of Applied Mathematics of the Russian Academy of Sciences, 4 Miusskaya Sq., Moscow 125047, Russian Federation; yuno@kiam.ru

Molchanov Dmitri A. (b. 1978) — Candidate of Science (PhD) in technology, associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; molchanov_da@rudn.university

Samuylov Andrey K. (b. 1988) — Candidate of Science (PhD) in physics and mathematics, associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; samuylov_ak@rudn.university

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 59–64

МНОГОМЕРНЫЙ РЕФЕРЕНСНЫЙ РЕГИОН ВЫСОКОЙ ПЛОТНОСТИ

М. П. Кривенко[1]

Аннотация: Рассматриваются принципы построения многомерных референсных регионов (MRR — multivariate reference region). Предложен оригинальный метод построения региона на основе областей с высокой плотностью точек и аппроксимации распределения данных с помощью смеси нормальных распределений. Для оценки порога для плотности распределения используется бутстреп-метод. В качестве эксперимента рассмотрена задача построения и использования эталонной области для прогнозирования типа мочевого камня. Обработка реальных данных продемонстрировала преимущества предлагаемых решений.

Ключевые слова: многомерный референсный регион; область высокой плотности; бутстреп-метод; смесь многомерных нормальных распределений

DOI: 10.14357/19922264170207

1 Введение

Многомерный референсный регион был предложен в литературе по клинической химии в начале 1970-х гг. как альтернатива одномерным референсным интервалам [1]. Там излагались преимущества предлагаемых множественных тестов, хоть и имеющих упрощенный вид, но снижающих (по отношению к одномерным вариантам) число ложных положительных результатов. Появление MRR оказалось особенно привлекательным для интерпретации результатов наборов медицинских тестов. Тем не менее возникали трудности в построении и использовании процедур многомерного анализа (см., например, [2]), связанные, в частности, с быстрым увеличением числа параметров, которые должны быть оценены. Немногие лаборатории использовали MRR в своей практике, причем в экспериментальном режиме, и, как следствие, на сегодняшний день имеется относительно малое количество соответствующих публикаций.

2 Многомерный референсный регион на основе расстояния Махалонобиса

Одномерный референсный интервал, полученный статистическим путем, использует центральную часть значений анализируемого показателя, обычно соответствующую 95% некоторой популяции — совокупности особей определенного вида (например, здоровой части населения определенного пола из некоторого диапазона возрастов). Одномерные референсные интервалы применялись в течение многих лет в качестве стандартного приема интерпретации лабораторных данных. Они легко формируются, хранятся, извлекаются и передаются в лабораторных информационных системах, просты в понимании, хорошо воспринимаются медицинским сообществом в ходе длительного использования. Тем не менее одномерные референсные интервалы при классификации данных могут дать большое число ложно аномальных результатов. Этот далеко не единственный недостаток однофакторного референсного интервала может быть полностью или частично устранен с помощью MRR.

Простейшим и весьма распространенным способом построения MRR является использование прямого произведения отдельных референсных интервалов в предположении, что они статистически независимы. Пусть $(1-\alpha)$ — вероятность попадания в MRR, а p_0 — вероятность попадания в референсный интервал для любого из d признаков, тогда $p_0 = \sqrt[d]{1-\alpha}$. С ростом размерности d значения p_0 быстро приближаются к 1, что фактически лишает смысла применение MRR.

Как и в одномерном случае, отправной точкой для построения MRR может стать нормальное распределение. Идеи центрального расположения референсного региона и заданной вероятности попадания в него приводят для d-мерного нормального

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, mkrivenko@ipiran.ru

распределения, имеющего плотность распределения

$$\varphi(y,\mu,\Sigma) = \\ = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left(-\frac{(y-\mu)^{\mathrm{T}}\,\Sigma^{-1}(y-\mu)}{2}\right),$$

где величина $(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)$ есть квадрат так называемого расстояния Махаланобиса между y и μ, к использованию многомерного эллипсоида

$$(2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left(-\frac{(y-\mu)^{\mathrm{T}}\,\Sigma^{-1}(y-\mu)}{2}\right) = \\ = const$$

или, что то же самое,

$$(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu) = const\,.$$

Его называют эллипсоидом равной плотности распределения (или просто эллипсоидом равной вероятности).

Если задаться вероятностью $(1-\alpha)$ попадания в эллипсоид равной вероятности вида $(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu) = \rho$, то параметр ρ удовлетворяет уравнению $\Pr\left\{\chi^2_d \le \rho\right\} = 1-\alpha$.

Использование эллипсоида в качестве MRR будет оправдано только тогда, когда исходное распределение данных есть многомерное нормальное. Поэтому становятся актуальными критерии подгонки, а также использование процедур нормализации распределения данных в многомерном случае. Если с помощью тестов выявляется, что распределение не является нормальным, то Международная федерация клинической химии и лабораторной медицины рекомендует, согласно [3], использовать двухступенчатую процедуру нормализации. Следует обратить внимание, что многошаговость здесь относится не к многомерности, а касается лишь покоординатного преобразования распределения данных к нормальному.

Первые же попытки применения MRR на основе расстояния Махалонобиса (фактически это означает принятие модели нормального распределения референсных значений) выявили ряд недостатков (более подробно смотри в [4, разд. 6.2]):

- проявление «проклятий» размерности при механическом увеличении d, в особенности если игнорируется этап анализа состава признаков [1, 5, 6];
- из-за небольших объемов обучающей выборки невысокая устойчивость при применении, в частности чувствительность к увеличению неточностей измерений после того, как регион был установлен [5, 7].
- предположение о нормальном распределении и попытки «подправить» действительность с помощью преобразований реальных данных для их нормализации при увеличении размерности данных становятся все более шаткими [5];
- представление и интерпретация выводов на основе MRR трудно понимаемы не только для специалистов в предметной области [8].

3 Многомерный референсный регион высокой плотности

Заметим, что в случае нормального распределения референсных значений для точек внутри построенного эллипсоида значения плотности распределения больше, чем на границе, а вне — меньше. Это замечание позволяет предложить другой подход к построению MRR.

Определение. Если плотность распределения референсных значений есть $f(y)$, то MRR есть область $A_t = \left\{y \in \mathcal{R}^d | f(y) \ge t\right\}$ для некоторого порогового значения t.

Для нормального распределения это уже упомянутый эллипсоид равной вероятности. Если задается вероятность $(1-\alpha)$ попадания в A_t, то пороговое значение t есть решение уравнения $\int_{A_t} f(u)\,du = 1-\alpha$, получить которое аналитически в случае произвольной плотности распределения вряд ли возможно. Здесь присутствуют две проблемы: вычисление многомерного интеграла и зависимость области интегрирования от неизвестного значения. Для решения их предлагается привлечь метод моделирования.

Сгенерируем выборку из $f(y)$, которую обозначим как $Y^f = \left\{y_1^f, \ldots, y_m^f\right\}$. Для оценки $\int_{A_t} f(u)\,du$ используем отношение:

$$\frac{\left|\left\{y_i^f | y_i^f \in A_t\right\}\right|}{m} = \frac{\left|\left\{y_i^f | f\left(y_i^f\right) \ge t\right\}\right|}{m} = \\ = 1 - \frac{\left|\left\{y_i^f | f(y_i^f) < t\right\}\right|}{m} = 1 - F_m(t)\,,$$

где $F_m(t)$ — эмпирическая функция распределения случайной величины $f(y)$, т. е. случайной величины, являющейся результатом преобразования с помощью функции $f(\cdot)$ случайной величины, имеющей плотность распределения $f(u)$.

Таким образом, искомая оценка t^* должна удовлетворять уравнению $F_m(t^*) = \alpha$ и может быть получена как непараметрическая оценка квантиля

порядка α из распределения $F_m(\cdot)$. Если обозначить $f_i = f(y_i^f)$, то t^* есть $f_{(r)}$, где

$$r = \begin{cases} m\alpha, & m\alpha \text{ — целое}; \\ \lfloor m\alpha + 1 \rfloor, & m\alpha \text{ — не целое}. \end{cases}$$

Заметим, что для такой оценки можно указать доверительный интервал.

Для построения MRR необходимо знать распределение данных. При реализации принципа точек высокой плотности в первую очередь следует обратиться к параметрическим моделям, в частности к смеси нормальных распределений, имеющей плотность распределения

$$f(u) = \sum_{j=1}^{k} p_j \varphi\left(u, \mu_j, \Sigma_j\right) .$$

Если $\hat{f}(u)$ — оценка смеси, то t^* строится следующим образом:

- генерируется выборка $\left\{y_1^f, \ldots, y_m^f\right\}$ из $\hat{f}(u)$ и для каждого ее i-го элемента подсчитывается значение $\hat{f}\left(y_i^f\right)$;
- в качестве t^* берется непараметрическая оценка квантиля порядка α (в случае необходимости дополнительно находится непараметрическая оценка доверительного интервала для t^*, что может характеризовать правильность выбранного объема для генерируемой выборки).

Пусть для $f(u)$ имеется A_t, а также получена $\hat{f}(u)$ и соответствующий MRR вида $\hat{A}_t$. Качество аппроксимации A_t с помощью $\hat{A}_t$ можно оценить с помощью вероятности совпадения этих областей, т. е.

$$P_c = \int\limits_{\{u \in A_t\} \cup \{u \in \hat{A}_t\}} f(u)\, du + \int\limits_{\{u \notin A_t\} \cup \{u \notin \hat{A}_t\}} f(u)\, du .$$

Для оценки P_c можно использовать величину

$$\hat{P}_c = \frac{\left|\left\{y_i^f | y_i^f \in \left\{\left\{y_i^f \in A_t\right\} \cup \left\{y_i^f \in \hat{A}_t\right\}\right\}\right\}\right|}{m} + \\ + \frac{\left|\left\{y_i^f | y_i^f \in \left\{\left\{y_i^f \notin A_t\right\} \cup \left\{y_i^f \notin \hat{A}_t\right\}\right\}\right\}\right|}{m} .$$

Использование MRR высокой плотности для диагностирования сводится к реализации так называемого слабого критерия значимости для наблюденного значения x: нулевая гипотеза заключается в том, что $x \in A_t$, статистика критерия есть $\hat{f}(x)$ и решение о принадлежности критической области A_t принимается при больших значениях $\hat{f}(x)$.

Для медицинской практики важна возможность использования референсного региона при интерпретации результатов обследования некоторого пациента с вектором признаков x. В подобных случаях сложившейся практикой для слабых критериев значимости является использование критического уровня α_{cr} (более распространенным в медицине является употребление термина p-значение) $\alpha_{\text{cr}} = \\ = \Pr\left\{\hat{f}(y) \le \hat{f}(x)\right\}$, где y — случайная величина, имеющая плотность распределения $\hat{f}(u)$, а $\hat{f}(x)$ — значение плотности распределения $\hat{f}(u)$ в точке x. Эта характеристика дает представление о том, насколько сильно данное наблюденное значение x противоречит гипотезе (или подкрепляет ее) о принадлежности данных MRR. При выбранном же заранее уровне значимости с помощью α_{cr} сразу же можно принять конкретное решение.

4 Эксперименты

Для демонстрации возможностей MRR использовались данные по прогнозу химического состава мочевых камней по метаболическим показателям мочи и сыворотки крови, а также антропологическим характеристикам пациентов [9]. В качестве исходной классификации камней рассматривалась следующая: чисто оксалатные (далее обозначены как O), чисто уратные (U), чисто фосфатные (P), смесь только оксалатных и уратных (OU), смесь только оксалатных и фосфатных (OP), смесь только уратных и фосфатных (UP), все остальные. Данная классификация была построена в [10] на основе доминирующих частот встречаемости основных компонентов. В качестве референсных значений рассматривались наборы метаболических и антропологических показателей (их всего было 14), соответствующих определенному классу камней.

Для каждого из основных классов O, U, P, OU, OP и UP перед построением MRR проводилась селекция признаков и принималось то значение размерности признакового пространства d и соответствующий набор показателей, которые позволяли прогнозировать состав камней без потери качества (методика описана в [9] и привела к значению $d = \\ = 9$). В качестве модели данных в первую очередь рассматривалась смесь многомерных нормальных распределений из пяти элементов (подбор числа элементов смеси проводился с помощью AIC — Akaike information criterion), для соответствующего региона было принято обозначение MRR(5). Для сравнения также использовалась модель нормального распределения, которой соответствовал MRR(1). Полученные результаты приводятся частично в таблице, где N — объем классифицируемых

Качество классификации с помощью MRR

Тип камня	N	$(1-\alpha)$, %	MRR(5)		MRR(1)	
			$(1-\hat{\alpha})$, %	$\hat{\beta}$, %	$(1-\hat{\alpha})$, %	$\hat{\beta}$, %
O	82	95	100	71	90	24
		85	96	78	89	36
		75	91	85	77	44
		65	76	88	74	50
U	76	95	100	75	91	24
		85	99	85	80	35
		75	82	89	74	48
		65	71	91	68	56
P	83	95	100	66	87	25
		85	94	78	86	33
		75	86	82	82	41
		65	77	87	75	47

данных; $\hat{\alpha}$ — оценка для α; $\hat{\beta}$ — оценка мощности критерия при определении типа камня на основании MRR.

Одной из базовых характеристик является вероятность попадания в MRR $(1-\alpha)$ и ее оценка $(1-\hat{\alpha})$. Сравнение соответствующих столбцов с учетом значений N и ориентировочных значений разброса (стандартные отклонения на основе биномиального распределения) не позволило выявить явных отклонений. Необходимо, правда, отметить, что во всех проанализированных случаях для MRR(5) оказалось, что $1-\hat{\alpha} \geq 1-\alpha$.

Назначение MRR, заключающееся в сжатом представлении референсных значений, в многомерном случае практически не проявляется. Для задания MRR(5) необходимо указать следующие величины: $1-\alpha$, t, $p_1, \ldots, p_{k-1}$, $\mu_1, \Sigma_1, \ldots, \mu_k, \Sigma_k$, общее количество которых равно $[2+(k-1)+k(d+$ $+\, d(d+1)/2)]$ и, в частности, в рассматриваемых экспериментах — 276. Для MRR(1) это значение меньше и равно 56. При этом для обрабатываемой обучающей выборки в зависимости от класса камней речь идет о порядка 10^2 векторах данных (см. столбец со значениями N), что приблизительно дает 10^3 скалярных величин.

Другое назначение MRR состоит в его использовании для диагностирования (классификации). В этой связи в первую очередь проводился сравнительный анализ MRR(1) (фактически это означает, что построение региона осуществляется на основе расстояния Махаланобиса) и MRR(5) (модель смеси нормальных распределений и предложенный в данной работе метод оценивания параметров региона). Показателем информативности метода построения многомерного региона выступала мощность соответствующего слабого критерия значимости, а именно: вероятность не попасть в MRR при условии, что данные берутся из дополнения к классу, для которого построена MRR. Сравнение соответствующих столбцов говорит о явном преимуществе двух предложенных моментов: усложнение модели данных путем перехода от нормального распределения к смеси нормальных распределений и построение региона высокой плотности.

Использование критического уровня можно продемонстрировать с помощью зависимости результатов сравнения двух классов от того, какой класс взять за основу. Введем для возможных значений p-величины три интервала: $(-\infty, 1\%)$, $[1\%, 5\%)$, $[5\%, 100\%)$ с соответствующей интерпретацией положения наблюденного набора показателей для пациента относительно построенного MRR: уверенное непопадание, неуверенное попадание, уверенное попадание. Если MRR построить для оксалатных камней, то результаты для анализа пациентов с фосфатными камнями дадут следующий вектор относительных частот попадания p-величин в указанные интервалы: $(60\%, 18\%, 22\%)$. Если же MRR строить для фосфатных камней, то получим $(71\%, 5\%, 24\%)$. Таким образом, для классификации указанных камней при приблизительно одинаковых частотах попадания в MRR (22% или 24%) уверенный отказ от референсного региона происходит чаще, если принять за базовый MRR регион для фосфатных камней. Построение шкалы, подобной рассмотренной, является прерогативой специалистов в предметной области, в данной работе она использовалась только для иллюстрации.

5 Заключение

На настоящий момент имеется относительно мало примеров применения MRR в клинической практике. Тому есть несколько причин. Математическое обеспечение, необходимое для получения и применения MRR, не отвечает возможностям

большинства клинических лабораторий. Лаборатории слабо оснащены программными средствами для реализации достаточно сложного математического аппарата многомерного анализа, а еще важнее, что отсутствуют методики, инструкции по использованию соответствующих средств. Лишь немногие клинические применения демонстрируют преимущества MRR, хотя свидетельств неудачных попыток больше.

Несмотря на сложности внедрения многомерного анализа референсных значений, можно сформулировать некоторые рекомендации по исследованию и разработке MRR. Во-первых, эффективная размерность в MRR должна быть как можно меньше, чтобы избежать затенения диагностически полезной информации тестами, создающими шум. Низкая размерность также должна уменьшить неблагоприятные последствия увеличения неточности результатов в связи с ростом числа анализируемых показателей. Во-вторых, показатели (тесты), включенные в MRR, должны быть физиологически релевантными исследуемому кругу расстройств, чтобы максимизировать информацию, полученную от MRR. В-третьих, чтобы учесть эффекты долгосрочной лабораторной изменчивости, данные, используемые для получения MRR, должны быть собраны и проанализированы в течение достаточно большого периода времени (от нескольких недель до нескольких месяцев). В-четвертых, представление результатов лабораторных исследований следует осуществлять в графическом виде, чтобы помочь врачам лучше понять MRR. Различные подходы к уменьшению размерности помогут выполнить это требование.

Необходима дальнейшая разработка пояснительных инструментов, способных воспринять результаты анализа MRR. При этом дополнительно необходима информация о том, какие именно тесты являются важнейшими факторами нарушения нормы. Надо признать, что соответствующий математический аппарат еще предстоит разработать. Решение перечисленных вопросов играет важную роль для обеспечения постоянного клинического применения MRR.

Литература

1. *Boyd J. C.* Reference regions of two or more dimensions // Clin. Chem. Lab. Med., 2004. Vol. 42. No. 7. P. 739–746.
2. *Winkel P.* Patterns and clusters — multivariate approach for interpreting clinical chemistry results // Clin. Chem., 1973. Vol. 19. No. 12. P. 1329–1333.
3. IFCC. Expert panel on theory of reference values. Approved recommendation on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits // J. Clin. Chem. Clin. Biochem., 1987. Vol. 25. No. 9. P. 645–656.
4. *Кривенко М. П.* Статистические методы представления и предварительной обработки референсных значений. — М.: ФИЦ ИУ РАН, 2016. 160 с.
5. *Boyd J. C., Lacher D. A.* The multivariate reference range: An alternative interpretation of multi-test profiles // Clin. Chem., 1982. Vol. 28. No. 2. P. 259–265.
6. *Albert A., Harris E. K.* Multivariate interpretation of clinical laboratory data. — New York, NY, USA: CRC Press, 1987. 328 p.
7. *Linnet K.* Influence of sampling variation and analytical errors on the performance of the multivariate reference region // Meth. Inf. Med., 1988. Vol. 27. No. 1. P. 37–42.
8. *Durbridge T. C.* Clinical acceptance of a multi-test reference region for biochemical-panel results // Clin. Chem., 1983. Vol. 29. No. 10. P. 1724–1726.
9. *Кривенко М. П.* Критерии значимости отбора признаков классификации // Информатика и её применения, 2016. Т. 10. Вып. 3. С. 32–40.
10. *Кривенко М. П., Голованов С. А., Сивков А. В.* Анализ однородности данных о химическом составе камней при уролитиазе // Информатика и её применения, 2013. Т. 7. Вып. 4. С. 94–104.

Поступила в редакцию 5.12.16

HIGH-DENSITY MULTIVARIATE REFERENCE REGION

M. P. Krivenko

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The paper considers the principles of construction of multivariate reference regions. An original method of construction of a region on the basis of areas of high density of points and approximation of data distribution with a mixture of normal distributions is suggested. To estimate the threshold for the probability density, the bootstrap method is used. As an experiment, the paper considers the problem of description and use of the reference region for predicting the type of urinary stones. Real data treatment demonstrated the benefits of the proposed solutions.

Keywords: multivariate reference region; high-density region; bootstrap method; multivariate normal mixture

DOI: 10.14357/19922264170207

References

1. Boyd, J. C. 2004. Reference regions of two or more dimensions. *Clin. Chem. Lab. Med.* 42(7):739–746.
2. Winkel, P. 1973. Patterns and clusters — multivariate approach for interpreting clinical chemistry results. *Clin. Chem.* 19(12):1329–1333.
3. IFCC. 1987. Expert panel on theory of reference values. Approved recommendation on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. *J. Clin. Chem. Clin. Biochem.* 25(9):645–656.
4. Krivenko, M. P. 2016. *Statisticheskie metody predstavleniya i predvaritel'noy obrabotki referensnykh znacheniy* [Statistical methods for representation and preliminary processing of reference values]. Moscow: FRC CSC RAS. 160 p.
5. Boyd, J. C., and D. A. Lacher. 1982. The multivariate reference range: An alternative interpretation of multi-test profiles. *Clin. Chem.* 28(2):259–265.
6. Albert, A., and E. K. Harris. 1987. *Multivariate interpretation of clinical laboratory data*. New York, NY: CRC Press. 328 p.
7. Linnet, K. 1988. Influence of sampling variation and analytical errors on the performance of the multivariate reference region. *Meth. Inf. Med.* 27(1):37–42.
8. Durbridge, T. C. 1983. Clinical acceptance of a multitest reference region for biochemical-panel results. *Clin. Chem.* 29(10):1724–1726.
9. Krivenko, M. P. 2016. Kriterii znachimosti otbora priznakov klassifikatsii [Significance tests of feature selection for classification]. *Informatika i ee Primeneniya — Inform. Appl.* 10(3):32–40.
10. Krivenko, M. P., S. A. Golovanov, and A. V. Sivkov. 2013. Analiz odnorodnosti dannykh o khimicheskom sostave kamney pri urolitiaze [Analysis of data homogeneity of the chemical compositions of stones in case of urolithiasis]. *Informatika i ee Primeneniya — Inform Appl.* 7(4):94–104.

Received December 5, 2016

Contributor

Krivenko Michail P. (b. 1946) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; mkrivenko@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 65–73

ПРИМЕНЕНИЕ КВАЗИСЛУЧАЙНОГО ПОДХОДА И АНСАМБЛЕВЫХ ВЫЧИСЛЕНИЙ ДЛЯ ОПРЕДЕЛЕНИЯ ОПТИМАЛЬНЫХ НАБОРОВ ЗНАЧЕНИЙ ПАРАМЕТРОВ КЛИМАТИЧЕСКОЙ МОДЕЛИ*

В. П. Пархоменко[1]

Аннотация: В условиях неопределенности значений большого числа параметров гидродинамической трехмерной глобальной климатической модели реализована процедура их одновременной оценки для близости результатов моделирования к данным наблюдений. Модель включает блоки атмосферы, термохалинной крупномасштабной циркуляции океана и морского льда. В квазислучайном подходе по методу латинского гиперкуба генерируется ансамбль из 200 расчетов путем равномерного полного покрытия диапазона изменения каждого из 12 параметров модели. Параметры определяют перемешивание и перенос в атмосфере, океане и морском льду, но их комбинации выбираются случайным образом. Исследование количественной меры ошибки модели позволило решить обратную задачу оценки параметров модели и прямую задачу прогнозных расчетов по модели.

Ключевые слова: глобальная климатическая модель; оценка параметров; метод латинского гиперкуба

DOI: 10.14357/19922264170208

1 Введение

Климатические модели имеют ряд настраиваемых параметров, значения которых не всегда определяются из теории или данных наблюдений при исследовании соответствующих процессов [1]. Даже характер физических процессов может быть неясен и зависеть от пространственного разрешения модели, а параметризации подсеточных процессов представляют собой самые различные физические явления (вихри и мелкомасштабные движения, инерционные гравитационные волны, приливы и т. п.). В таких случаях значения параметров могут быть определены путем выбора оптимального ансамбля модельных результатов для соответствия данным наблюдений. Это, естественно, влечет за собой поиск оптимальных квазистационарных решений в многомерном пространстве всех параметров модели. Использование стандартного метода Монте Карло потребует десятков или сотен тысяч интегрирований модели до достижения квазистационарных состояний.

Для моделей с высоким или умеренным разрешением вычислительные затраты даже одного такого расчета могут оказаться непомерно высокими [2, 3]. Вместо этого большие модели, как правило, настроены на последовательность расчетов для подробного исследования влияния одного параметра. Однако взаимозависимость параметров почти наверняка означает, что даже порядок, в котором такие исследования проводятся, повлияет на конечный результат и, следовательно, на модельные прогнозы.

Вычислительно эффективные модели имеют значительный потенциал для выполнения большого числа расчетов за разумное время и позволяют исследовать большие диапазоны в пространстве их параметров. Если параметры имеют явную физическую интерпретацию или близкие аналоги в модели с более высоким пространственным разрешением, то результаты могут иметь и более общее значение. Вычислительно эффективные модели также полезны для понимания долгосрочной естественной изменчивости климата, в этом случае оптимальный баланс сложности блоков модели может зависеть от временны́х масштабов, интересующих исследователя.

В статье рассматривается модель океана с произвольным рельефом дна в глобальной постановке в геострофическом приближении с фрикционным членом и с расширением за счет добавления энерго- и влагобалансовой модели атмосферы и динамической и термодинамической модели морского льда [4]. В данной реализации увеличено гори-

*Работа выполнена при поддержке РФФИ (проекты 16-01-0466, 17-01-00693, 17-07-00035).

[1]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук; Московский государственный технический университет им. Н. Э. Баумана, parhom@ccas.ru

зонтальное разрешение модели до 72×72 расчетных ячеек [5]; тем не менее, учитывая достаточно простое представление процессов в атмосфере, в результате совместная модель имеет высокую вычислительную эффективность.

2 Описание модели

Представлена глобальная модель климата, которая включает полностью трехмерную, с трением геострофическую модель океана, обладающая высокой эффективностью интегрирования по сравнению со значительно более ресурсоемкими климатическими моделями с трехмерными примитивными уравнениями океана. Модель включает также динамическую и термодинамическую модель морского льда и энерго- и влагобалансовую модель атмосферы.

Система уравнений модели океана рассматривается в геострофическом приближении с фрикционным членом в уравнениях импульса по горизонтали [4, 5]. Значения температуры и солености удовлетворяют адвекционно-диффузионным уравнениям, что позволяет описать термохалинную циркуляцию океана. Приближенным образом учитываются также конвективные процессы. Таким образом, система основных уравнений, записанных для наглядности в локальных декартовых координатах (x, y, z), где x, y — горизонтальные координаты и z — высота, направленная вверх, имеет следующий вид:

– уравнения импульса по горизонтали

$$-lv + \lambda u = -\frac{1}{\rho}\frac{\partial p}{\partial x} + \frac{1}{\rho}\frac{\partial (k_w \tau_x)}{\partial z}\,;$$

$$lu + \lambda v = -\frac{1}{\rho}\frac{\partial p}{\partial y} + \frac{1}{\rho}\frac{\partial (k_w \tau_y)}{\partial z}\,;$$

– уравнение неразрывности

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial y} = 0\,;$$

– уравнение гидростатики

$$\frac{\partial p}{\partial z} = -\rho g\,;$$

– уравнение состояния морской воды

$$\rho = \rho(S, T)\,;$$

– уравнение переноса и диффузии трассеров X (температуры и солености)

$$\frac{d}{dt}X = k_h \nabla^2 X + \frac{\partial}{\partial z}\left(k_v \frac{\partial X}{\partial z}\right) + C\,,$$

в которых u, v и w — горизонтальные и вертикальная компоненты вектора скорости соответственно; λ — переменный в пространстве параметр, увеличивающийся к береговым границам и экватору и определяющий влияние фрикционного члена; T, S и p — температура, соленость и давление соответственно; τ_x и τ_y — компоненты напряжения трения ветра; ρ — плотность воды; l — параметр Кориолиса; g — ускорение свободного падения; k_h и k_v — коэффициенты турбулентной диффузии трассеров по горизонтали и вертикали соответственно; C — источники.

Указанная система уравнений решается в сферической системе координат для всего Мирового океана с реальной аппроксимированной глубиной. На границах материков принимаются равными нулю нормальные составляющие потоков тепла и солей. Океан подвергается воздействию напряжения трения ветра на поверхности. Потоки T и S у дна полагаются равными нулю, а на поверхности определяются взаимодействием с атмосферой.

В термодинамической модели морского льда динамические уравнения решаются для сплоченности льда и для средней толщины льда. Рост и таяние льда в модели зависят только от разности между потоком тепла из атмосферы в морской лед и потока тепла изо льда в океан. Для температуры поверхности льда решается диагностическое уравнение.

Для описания процессов, протекающих в атмосфере, используется энерго- и влагобалансовая модель. В модели решается вертикально проинтегрированное уравнение для температуры, определяющее баланс приходящего и уходящего радиационных потоков, явных (турбулентных) обменов потоками тепла с подстилающей поверхностью, высвобождения скрытого тепла из-за осадков и простой однослойной параметризации горизонтальных процессов переноса. Источники в уравнении переноса для удельной влажности определяются осадками, испарением и сублимацией с подстилающей поверхности.

Все блоки модели связаны между собой обменом импульсом, теплом и влагой. Используется реальная конфигурация материков и распределение глубин Мирового океана [5]. Уравнения в сферической системе координат решаются численным конечно-разностным методом. По горизонтали применяется равномерная по долготе и синусу широты расчетная сетка размерностью 72×72. Глубина океана представляется в виде восьмиуровневой логарифмической шкалы до максимального значения 5000 м. Начальное состояние системы характеризуется постоянными температурами океана, атмосферы и нулевыми скоростями течений океана. Числен-

ные эксперименты показывают, что модель выходит на равновесие за период около 2000 лет [5].

3 Постановка задачи оценки параметров и результаты

В предлагаемом квазислучайном подходе генерируется ансамбль расчетов путем равномерного полного покрытия диапазона изменения каждого индивидуального параметра модели, которые перечислены далее, но комбинации параметров выбираются случайным образом. Это соответствует равномерному разбиению вероятностного пространства значений параметров при равномерном распределении плотности вероятности. Таким образом, при M расчетах и N параметрах каждый параметр принимает M значений, равномерно (или по логарифмическому закону) покрывающих весь диапазон его изменения, но порядок, в котором выбираются эти значения, определяется случайным образом. Это соответствует понятию так называемого «латинского гиперкуба» в статистике и планировании эксперимента [6]. Выборки из латинских гиперкубов начали активно применяться после удачных решений в области планирования эксперимента, где их использование позволяет уменьшить взаимную зависимость факторов без увеличения числа экспериментов [6]. Каждый расчет представляет собой отдельное интегрирование модели на 2000 лет от однородного состояния климатической системы с нулевыми скоростями течений до установившегося состояния при стандартных условиях, соответствующих современному климату [5].

Как показывают расчеты, окончательное квазистационарное состояние может быть не единственным для данного набора параметров. Другие квазистационарные состояния могут быть получены с использованием различных начальных условий, в частности различных начальных температур океана. Однако в настоящей работе прежде всего изучается влияние изменения параметров модели и поэтому фиксируется начальная температура океана на 20 °C. Такая постановка приводит к быстрому конвективному механизму начала процессов установления в океане.

Для обработки результатов такого большого количества численных экспериментов необходимо определить объективную меру ошибки модели. Для этого используется взвешенная среднеквадратическая ошибка на множестве всех динамических переменных в океане и атмосфере по сравнению с интерполированными данными наблюдений, а именно: температуры и влажности воздуха на поверхности (1000 мб), в среднем за период с 1948 до 2002 гг., и температуры и солености океана [7].

В таблице перечислены 12 основных параметров модели (первый столбец) и принимаемые диапазоны их возможного изменения (второй и третий столбцы) [8]. Если изменять каждый из этих параметров в отдельности, то будет изучена только очень ограниченная область пространства параметров. Поэтому допускаем, чтобы все 12 параметров изменялись сразу в указанных диапазонах, которые приведены в таблице. Предельные значения выбираются таким образом, чтобы покрывать или превышать диапазон разумного выбора соответствующих значений для такой модели.

Диапазон изменения параметров модели климата для ансамблевых экспериментов

	Параметр модели	Минимум	Максимум	Приемлемый диапазон
	Океан			
1.	Горизонтальная диффузия, м^2/с	300	10^4	4200–8500
2.	Вертикальная диффузия, м^2/с	$2 \cdot 10^{-6}$	$2 \cdot 10^{-4}$	$3 \cdot 10^{-5}$–$1{,}9 \cdot 10^{-4}$
3.	Коэффициент трения, сут$^{-1}$	1/5	2	0,6–1,90
4.	Ветровое воздействие	1	3	1,14–2,58
	Атмосфера			
5.	Диффузия тепла, м^2/с	10^6	10^7	$4{,}35 \cdot 10^6$–$9 \cdot 10^6$
6.	Угловой коэффициент, рад	0,5	2	0,7–1,45
7.	Коэффициент наклона	0	0,25	0,023–0,230
8.	Диффузия влажности, м^2/с	$5 \cdot 10^6$	$5 \cdot 10^6$	$1 \cdot 10^5$–$3 \cdot 10^5$
9.	Коэффициент адвекции тепла	0	1	0,050–0,815
10.	Коэффициент адвекции влажности	0	1	0,255–0,850
11.	Поток между океанами, S_v	0	0,64	0–0,75
	Морской лед			
12.	Диффузия морского льда, м^2/с	300	10^4	300–9320

Всего по модели было проведено $M = 200$ расчетов. Для определения ошибки модельных результатов используется взвешенная среднеквадратичная ошибка, вычисляемая по набору всех динамических переменных для атмосферы и океана при сравнении с данными наблюдений:

$$\varepsilon^2 = \sum_{i=1}^{n} w_i \left(X_i - D_i\right)^2 ,$$

где X_i и D_i — соответственно модельные результаты и данные наблюдений для этих переменных (температура и влажность атмосферы, температура и соленость океана). Суммирование ведется по всем точкам трехмерной сетки и по всем указанным переменным ($n = 30\,008$). Величины $w_i = 1/(n\sigma_X^2)$ — весовые множители, зависящие от соответствующей переменной X_i, но не зависящие от точки сетки; σ_X — среднеквадратичная ошибка данных наблюдений. Вычисляется также альтернативная ошибка ε_A — по той же формуле, но только для расчетных точек и переменных атмосферы.

На рис. 1 и 2 приведены 12 графиков со значениями вычисленных ошибок в зависимости от исследуемых параметров. На этих рисунках символами *1* отмечены значения параметров с ошибками $\varepsilon > 0{,}6$; *2–4* соответствуют меньшим значениям ошибки. Среди последних символами *2* отмечены значения параметров с ошибкой $\varepsilon_A > 0{,}1$; *3* (всего 4 штуки) отмечены значения $\varepsilon < 0{,}6$ и $\varepsilon_A < 0{,}1$ одновременно, при этом исследование климатических распределений показывает, что достигнуто состояние климатической системы, не соответствующее современному. Эти результаты исключаются из рассмотрения. Наконец, символами *4* (всего 7 штук) отмечены приемлемые результаты расчетов ($\varepsilon < 0{,}6$ и $\varepsilon_A = 0{,}1$) с минимальными ошибками, описывающие современный климат.

Таким образом, результаты показывают, что сформулированным критериям удовлетворяют 7 наборов значений 12 параметров. Граничные значения ошибок $\varepsilon = 0{,}6$ и $\varepsilon_A = 0{,}1$ соответствуют ошибкам данных наблюдений. По этой причине нет оснований в расчетах предпочесть только один набор значений параметров. Предлагается вести ансамблевые расчеты по модели сразу с 7 оптимальными наборами параметров и в качестве результатов рассматривать средние по ансамблю и отклонения от них. В соответствии с таблицей и рис. 1 и 2 в наборы параметров входят значения параметров, меняющиеся в широком диапазоне (последний столбец в таблице). Это может означать, что предположение о постоянных значениях параметров достаточно грубое. В зависимости от расчетных ха-

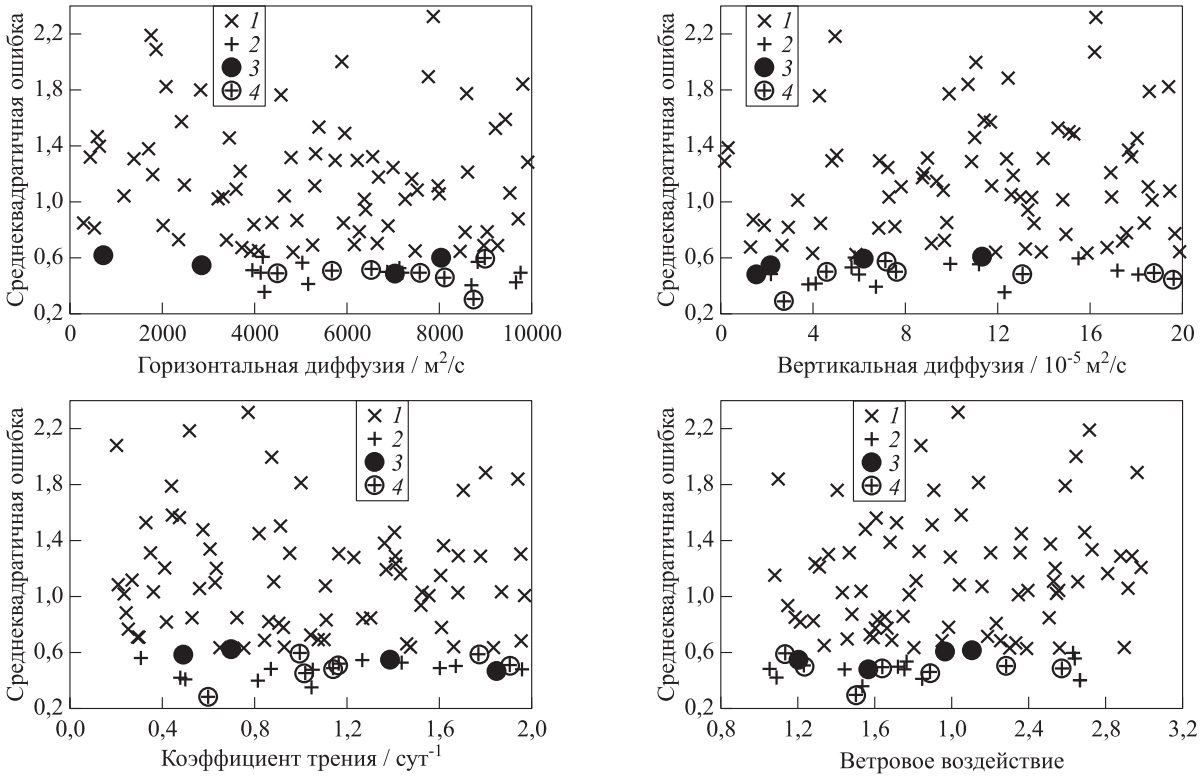

Рис. 1 Среднеквадратичные ошибки в зависимости от величины исследуемых параметров под номерами 1–4 из таблицы

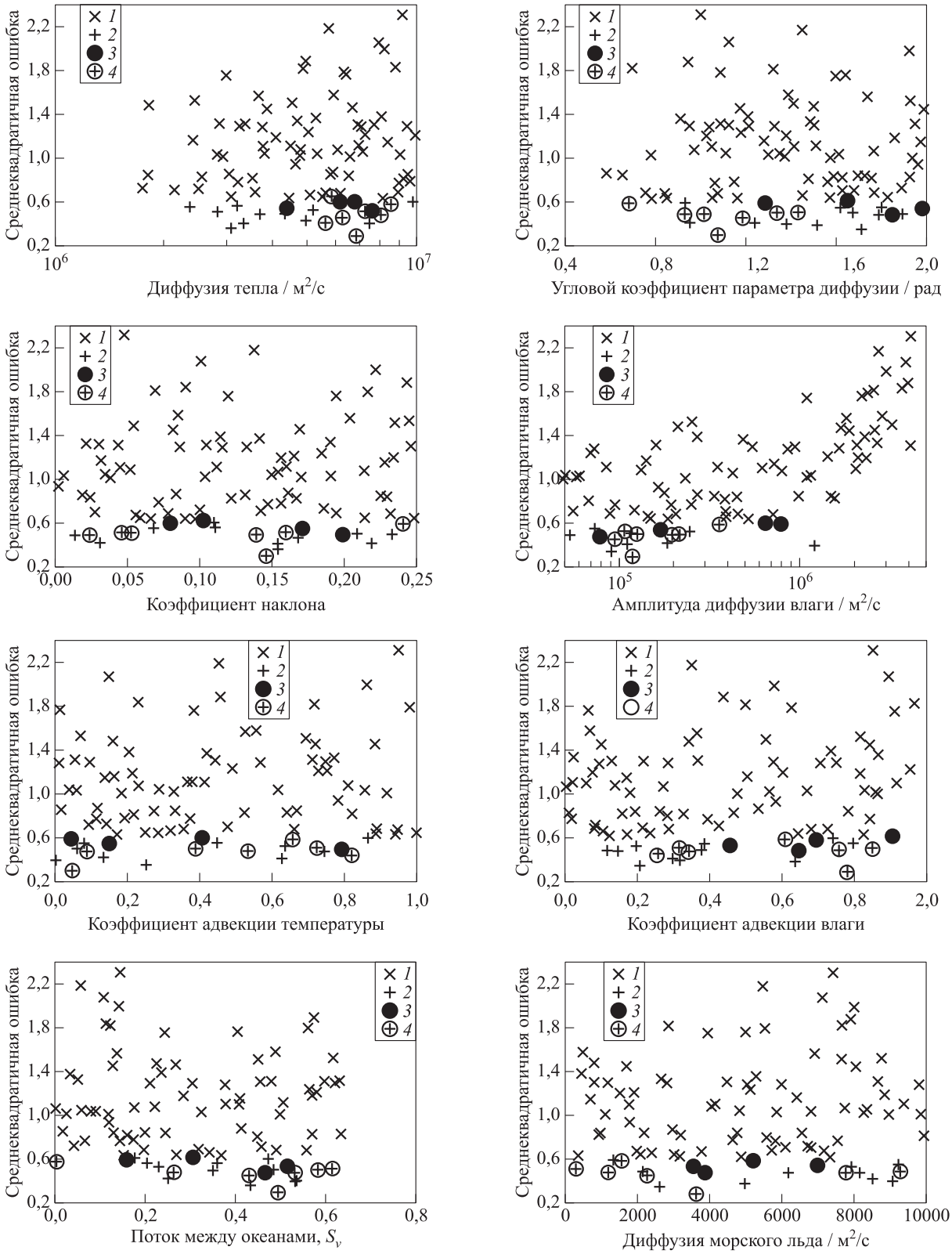

Рис. 2 Среднеквадратичные ошибки в зависимости от величины исследуемых параметров под номерами 5–12 из таблицы

рактеристик климата, градиентов, географических координат и некоторых других причин значения параметров могут меняться во времени и пространстве.

В силу заложенных в постановку ограничений модели и сложности описываемых процессов эти зависимости неизвестны. Однако предлагаемая процедура проведения ансамблевых расчетов в некоторой степени учитывает эти зависимости и позволяет уточнить результаты, поскольку дает диапазон изменения климатических характеристик в рамках ансамбля.

4 Ансамблевые расчеты с оптимальными наборами параметров модели

Далее приведены результаты расчетов по модели с использованием 7 приемлемых наборов пара-

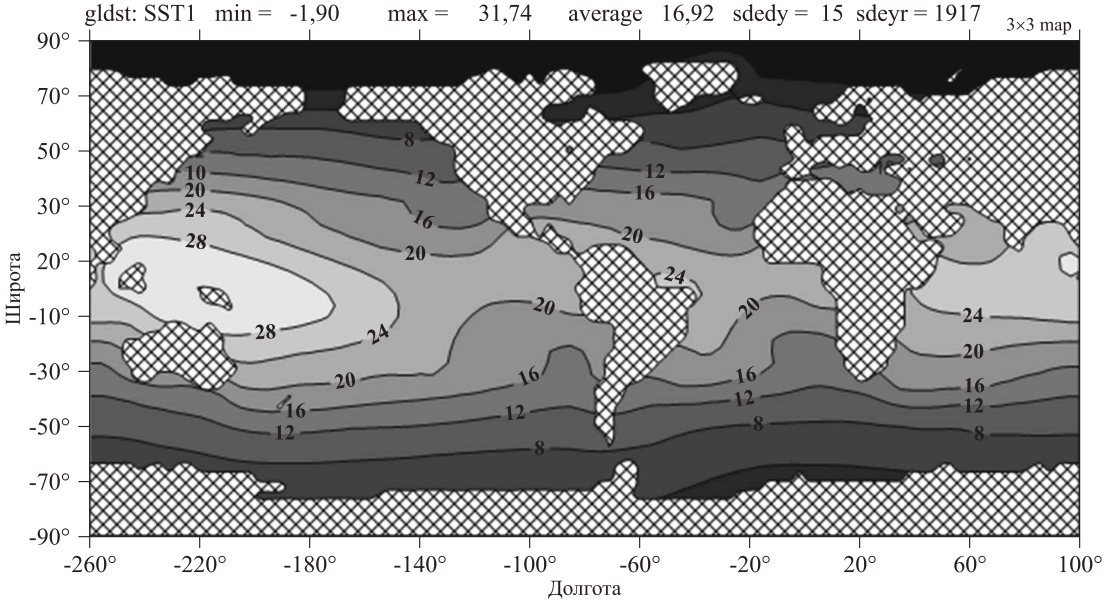

Рис. 3 Температура поверхности океана, осредненная по результатам 7 расчетов с минимальной ошибкой

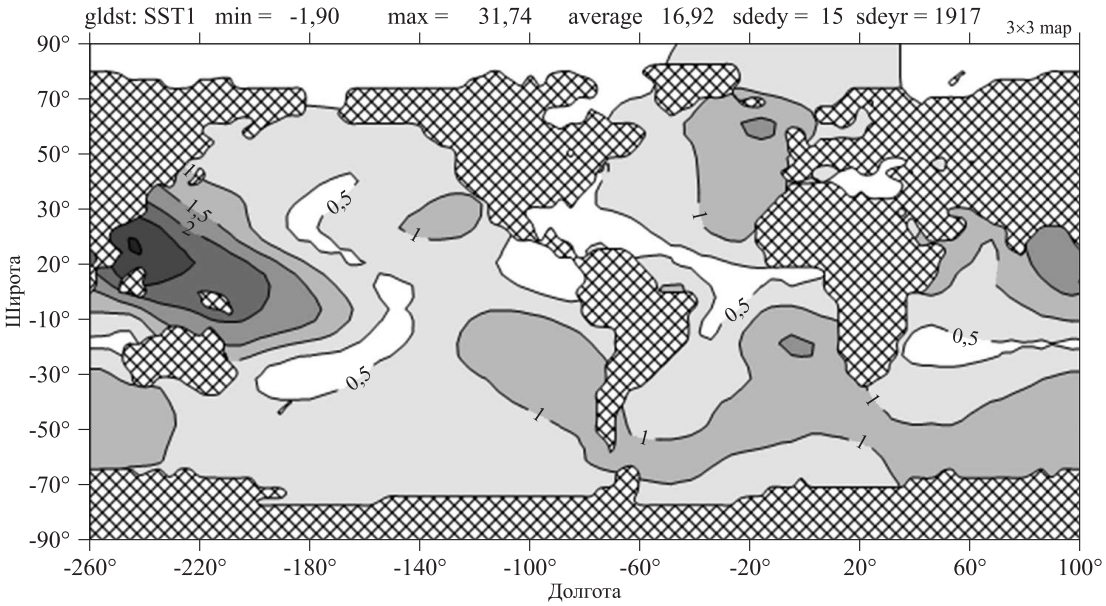

Рис. 4 Среднеквадратичное отклонение температуры поверхности океана в январе, вычисленное по набору 7 расчетов с минимальной ошибкой

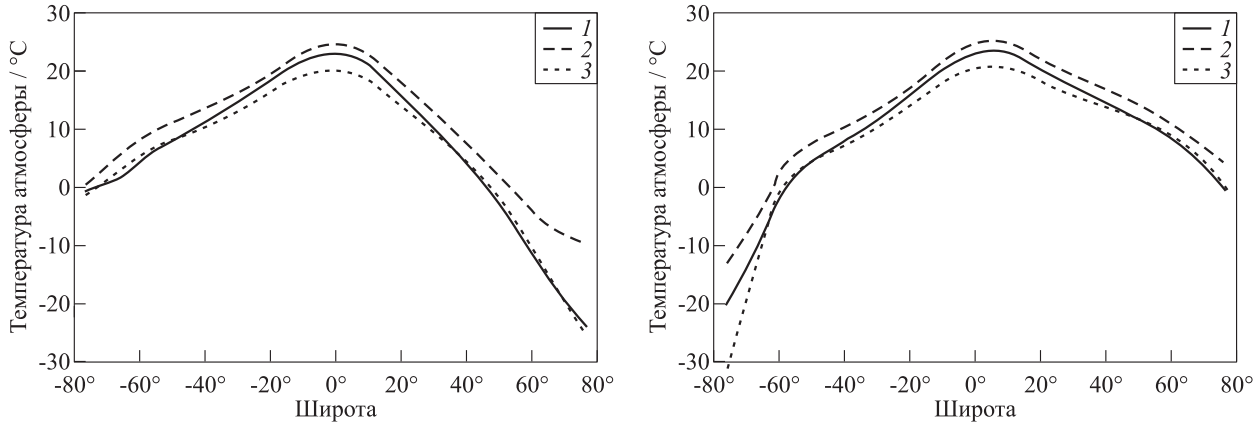

Рис. 5 Распределение зонально осредненной температуры атмосферы для января (*а*) и июля (*б*): *1* — данные наблюдений; *2* — максимум в ансамбле расчетов; *3* — минимум в ансамбле расчетов

метров, обеспечивающих минимальную ошибку по сравнению с данными наблюдений. Расчеты ведутся в постановке, описанной выше, до установившегося состояния, соответствующего современному климату (рис. 3). Сравнение с данными наблюдений показывает хорошее совпадение (рис. 4 и 5). Среднеквадратичное отклонение температуры поверхности океана, вычисленное по набору 7 расчетов с минимальной ошибкой, практически во всей области не превышает 0,5–1,0 °C (см. рис. 4).

Расчетный разброс климатических откликов на глобальное потепление при 100-летнем прогнозе (для приземной температуры воздуха разброс около 0,3 °C, см. рис. 6) является существенным, учитывая, что он представляет собой диапазон предсказаний, возникающий только с изменением параметров перемешивания и транспорта в модели.

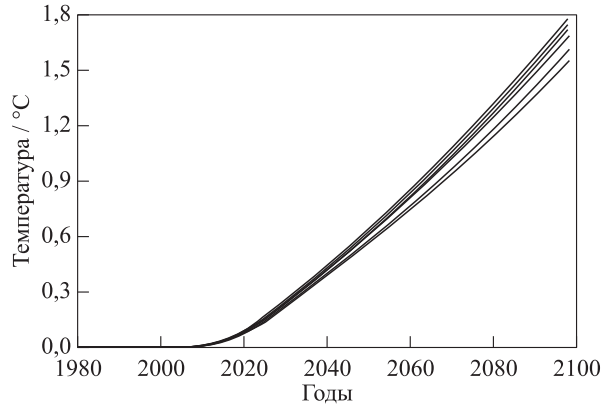

Рис. 6 Изменение средней глобальной температуры атмосферы для 7 расчетов с минимальной ошибкой при прогнозируемом увеличении концентрации CO_2 с 2010 до 2100 г.

5 Заключение

Посредством анализа случайным образом сгенерированных расчетов на 2000 лет рассмотрены неопределенности, связанные с 12 параметрами модели, определяющими перемешивание и перенос в атмосфере, океане и морском льду. Исследование количественной меры ошибки модели позволило решить обратную задачу оценки параметров модели и прямую задачу прогнозных расчетов по модели. Результаты представляют собой попытку настройки трехмерной климатической модели жестко определенной процедурой, но в которой, тем не менее, рассматривается соответствующее пространство квазислучайного изменения параметров модели. Этот подход обеспечивает соответствие результатов моделирования данным наблюдений, хотя модельные входные параметры исходно точно не известны и могут меняться в широких пределах. Неопределенность предсказаний модели преодолевается двумя различными способами: во-первых, рассмотрением множества прогнозов по подмножеству примерно одинаково правдоподобных моделей и, во-вторых, достаточно статистически обоснованной процедурой взвешивания всех результатов в соответствии со средней ошибкой. Меньшее значение ошибки, вероятно, означает лучшее качество моделирования, и поэтому, если модель в динамике надежна, приемлемые прогнозы находятся в пределах неопределенности порядка ошибки.

Литература

1. *Edwards N. R., Marsh R.* Uncertainties due to transport-parameter sensitivity in an efficient 3-D ocean-climate model // Clim. Dynam., 2005. Vol. 24. No. 4. P. 415–433.

2. *Randall D. A.* General circulation model development. — Gardners Books, 2010. 416 p.

3. *Satoh M.* Atmospheric circulation dynamics and general circulation models. — Berlin: Springer-Verlag, 2014. 730 p.

4. *Marsh R., Edwards N. R., Shepherd J. G.* Development of a fast climate model (C-GOLDSTEIN) for Earth System Science // SOC, 2002. No. 83. 54 p.

5. *Пархоменко В. П.* Глобальная модель климата с описанием термохалинной циркуляции Мирового океана // Математическое моделирование и численные методы, 2015. № 1. С. 94–108.

6. *Montgomery D. C.* Design and analysis of experiments. — 5th ed. — New York, NY, USA: John Wiley & Sons, 2001. 684 p.

7. *Levitus S., Boyer T. P., Conkright M. E., O'Brien T., Antonov J., Stephens C., Stathoplos L., Johnson D., Gelfeld R.* Noaa Atlas Nesdis 18, World Ocean database. — Washington, D.C., USA: U.S. Government Printing, 1998. Vol. 1. 346 p.

8. *Parkhomenko V.* Ensemble calculations application for estimation and optimization of climate model parameters // 3rd Conference (International) on Optimization Methods and Applications Proceedings. — Moscow: Computing Center of RAS, 2012. P. 203–207.

Поступила в редакцию 26.01.17

APPLICATION OF QUASI-RANDOM ENSEMBLE CALCULATIONS FOR DETERMINATION OF CLIMATE MODEL OPTIMAL PARAMETERS

V. P. Parkhomenko[1,2]

[1]A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation
[2]N. E. Bauman Moscow State Technical University, 5 Baumanskaya 2nd Str., Moscow 105005, Russian Federation

Abstract: By analyzing a randomly generated set of runs, each 2000 years in length, the author has considered the uncertainty in 12 mixing and transport parameters. Constructing a quantitative measure for the model error made it possible to address both the inverse problem of estimation of model parameters and the direct problem of model predictions. The results represent an attempt at tuning a three-dimensional climate model by a strictly defined procedure which, nevertheless, considers the whole of the appropriate parameter space. The modeling approach is thus to match model outputs to observations while model inputs (parameters) are initially only weakly constrained.

Keywords: global climate model; model parameters estimation; latin hypercube sampling

DOI: 10.14357/19922264170208

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (projects 16-01-0466, 17-01-00693, and 17-07-00035).

References

1. Edwards, N. R., and R. Marsh. 2005. Uncertainties due to transport-parameter sensitivity in an efficient 3-D ocean-climate model. *Clim. Dynam.* 24(4):415–433.

2. Randall, D. A. 2010. *General circulation model development*. Gardners Books. 416 p.

3. Satoh, M. 2014. *Atmospheric circulation dynamics and general circulation models*. Berlin: Springer-Verlag. 730 p.

4. Marsh, R., N. R. Edwards, and J. G. Shepherd. 2002. Development of a fast climate model (C-GOLDSTEIN) for Earth System Science. *SOC* 83. 54 p.

5. Parkhomenko, V. P. 2015. Global'naya model' klimata s opisaniem termokhalinnoy tsirkulyatsii Mirovogo okeana [Global climate model including description of thermohaline circulation of the World Ocean]. *Matematicheskoe modelirovanie i chislennye metody* [Mathematical Modeling and Numerical Methods] 1:94–108.

6. Montgomery, D. C. 2001. *Design and analysis of experiments*. 5th ed. New York, NY: John Wiley & Sons, Inc. 684 p.

7. Levitus, S., T. P. Boyer, M. E. Conkright, T. O'Brien, J. Antonov, C. Stephens, L. Stathoplos, D. Johnson, and R. Gelfeld. 1998. *Noaa Atlas Nesdis 18, World Ocean database 1998*. Washington, D.C.: U.S. Government Printing. Vol. 1. 346 p.

8. Parkhomenko, V. 2012. Ensemble calculations application for estimation and optimization of climate model parameters. *3rd Conference (International) on Optimization Methods and Applications Proceedings*. Moscow: Computing Center of RAS P. 203–207.

Received January 26, 2017

Contributor

Parkhomenko Valery P. (b. 1951) — Candidate of Science (PhD) in physics and mathematics, head of laboratory, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; associate professor, N. E. Bauman Moscow State Technical University, 5 Baumanskaya 2nd Str., Moscow 105005, Russian Federation; parhom@ccas.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 74–84

МОДИФИКАЦИЯ ФУНКЦИОНАЛА КАЧЕСТВА В ЗАДАЧАХ НЕЛИНЕЙНОЙ РЕГРЕССИИ ДЛЯ УЧЕТА ГЕТЕРОСКЕДАСТИЧНЫХ ПОГРЕШНОСТЕЙ ИЗМЕРЯЕМЫХ ДАННЫХ

Г. И. Рудой[1]

Аннотация: Рассматривается задача восстановления нелинейной регрессионной зависимости по данным, имеющим погрешности определения как зависимых, так и независимых переменных, при этом распределения погрешностей разных измерений могут иметь разную дисперсию. Предлагается модифицированный функционал качества, учитывающий погрешности определения независимых переменных и разные распределения погрешностей в разных точках. Приводятся результаты численного моделирования на данных, полученных в ходе эксперимента по измерению зависимости мощности лазера от прозрачности резонатора. Рассматривается сходимость вектора параметров, минимизирующего предлагаемый функционал качества, к оптимальному для классического функционала среднеквадратичной ошибки. Сравнивается сходимость параметров, оптимальных для предлагаемого и классического функционалов, к некоторым «истинным» параметрам модели на данных, сгенерированных согласно этим «истинным» параметрам и зашумленным согласно предположениям о погрешностях измерений, в зависимости от параметров этих погрешностей.

Ключевые слова: гетероскедастичные ошибки; ошибки измерения независимых переменных; символьная регрессия; нелинейная регрессия

DOI: 10.14357/19922264170209

1 Введение

В ряде приложений (см., например, [1,2]) возникает задача нахождения оптимальных коэффициентов $\boldsymbol{\omega}$ некоторой регрессионной модели f, заданной в виде аналитической формулы, по набору экспериментальных данных. Для этого в предположении о нормальном распределении регрессионных остатков строится функционал $\sum_i (y_i - f(x_i, \boldsymbol{\omega}))^2$, представляющий сумму квадратов отклонений экспериментальных точек y_i от значения регрессионной кривой $f(x, \boldsymbol{\omega})$ в точке x_i, и находится вектор параметров $\boldsymbol{\omega}$, его минимизирующий.

Однако данный функционал корректен только для точно измеренных независимых переменных и гомоскедастичных ошибок измерения зависимой переменной: в частности, для линейных моделей соответствующая оценка параметров является несмещенной, состоятельной и наиболее эффективной только при выполнении этих условий. В случае нелинейных моделей вывод функционала среднеквадратичной ошибки согласно методу наибольшего правдоподобия также опирается на эти предположения (с обобщением в виде взвешенного метода наименьших квадратов (МНК) в случае разных стандартных отклонений зависимой переменной). Иными словами, предполагается существование лишь ошибок измерения зависимой переменной, распределение которых принимается одинаковым.

На практике, как правило, это предположение не выполняется, особенно при измерениях в достаточно широких диапазонах. Например, в задаче нахождения зависимости коэффициента преломления n прозрачного полимера от длины волны λ погрешности измерения каждого физического параметра в разных точках, вообще говоря, различны [2]. Так, если для измерения длины волны λ используется дифракционная решетка, то постоянной является относительная погрешность определения длины волны $\sigma_{\lambda_i}/\lambda_i \approx const$ и, следовательно, погрешность определения длины волны зависит от самой длины волны. Подобная ситуация фиксированной относительной (а не абсолютной) ошибки является типичной для физических экспериментов.

Таким образом, возникает задача поиска оптимальных коэффициентов регрессионной формулы с учетом различающихся погрешностей измерения в разных экспериментальных точках. Для некоторых частных случаев эта задача решена.

Так, детальный обзор методов решения этой задачи в случае линейной регрессии приведен в [3].

[1]Московский физико-технический институт, 0xd34df00d@gmail.com

В частности, для линейных моделей рассматривается даже более общая задача, когда распределение ошибок не является точно известным. Однако, по крайней мере для ряда методов, априорная информация все равно необходима, как то: значение отношения стандартных отклонений зависимой и независимой переменных в случае регрессии Деминга [4] либо наличие инструментальных переменных при использовании одноименного метода [5]. Отметим, что дополнительная априорная информация необходима для обеспечения возможности однозначного определения параметров модели, иначе модель становится неидентифицируемой. При этом условие идентифицируемости модели для случая многомерной линейной регрессии в общем виде до сих пор неизвестно [6].

Обзор методов решения аналогичной задачи для случая нелинейной регрессии приведен в [7]. Так, например, метод инструментальных переменных обобщается на случай нелинейных моделей, при этом опять же требуется наличие дополнительных наблюдаемых переменных, пропорциональных регрессору с точностью до некоторой аддитивной ошибки. Заметим, что условия идентифицируемости модели при этом неизвестны.

В ряде работ изучаются конкретные нелинейные регрессионные модели, и соответствующие ошибки измерений предполагаются экспертно заданными. Например, в [8] рассматривается модель Басса, описывающая динамику процесса распространения новых потребительских продуктов, для которой вводится предположение о неравной точности измерений в разных экспериментальных точках, что описывается разными весовыми коэффициентами при соответствующих регрессионных остатках. При этом весовые коэффициенты имеют достаточно общий вид и вводятся произвольно в виде экспертно указанных значений.

Другим примером является [9], где рассматривается задача оценки коэффициентов трехпараметрического распределения Вейбулла по неточно измеренным данным. Для этого используется метод латентных переменных: к независимым переменным t_i добавляются «свободные» переменные δ_i, предоставляющие степень свободы в пространстве независимых переменных, и минимизируется функционал вида

$$T(\alpha,\beta,\eta,\boldsymbol{\delta}) = \\ = \sum_{i=0}^{n} w_i \left[f(t_i+\delta_i;\alpha,\beta,\eta)-y_i\right]^2 + \sum_{i=0}^{n} p_i\delta_i^2,$$

где α, β и η — параметры распределения, а w_i и p_i являются некоторыми экспертно заданными весами, соответствующими относительной точности i-го измерения аналогично [8].

В настоящей работе рассмотрена более общая ситуация, в которой не только зависимые, но и независимые переменные определяются неточно и каждая переменная имеет свою погрешность измерения, заданную экспертно. Исследуется случай нелинейной регрессионной зависимости, в отличие, например, от [10], где изучается линейная модель. Предлагается модифицированный функционал качества, учитывающий погрешности как зависимых, так и независимых переменных в виде, достаточном для большинства практических приложений. Весовые коэффициенты при регрессионных остатках в настоящей работе выводятся из базовых предположений о распределении погрешностей измерения и о поведении регрессионной модели в окрестности каждой экспериментальной точки. В частности, оказывается, что весовые коэффициенты зависят не только от самой погрешности измерений в данной точке, но и от производных регрессионной модели в окрестности этой точки.

Предложенный функционал наиболее близок к описанному в [11]. Однако, кроме того, в настоящей работе предлагается вероятностная интерпретация функционала для случая нормально распределенных ошибок.

В разд. 2 настоящей работы формально поставлена задача нахождения оптимальных параметров регрессионной модели с учетом гетероскедастичных погрешностей определения как зависимых, так и независимых переменных. В разд. 3 выводится предлагаемый функционал качества. Затем, в разд. 4, описывается метод использования имеющихся алгоритмов оптимизации, применяемых в подобных задачах (как, например, алгоритм Левенберга–Марквардта [12]), для минимизации предлагаемого функционала. В разд. 5 приводятся результаты вычислительного эксперимента, состоящие из трех частей: во-первых, приводятся результаты анализа экспериментальных данных по измерению параметров усиливающей среды газового лазера; затем сравнивается сходимость оптимальных параметров для предложенного функционала качества к параметрам, минимизирующим классический функционал среднеквадратичной ошибки, в зависимости от параметров распределения ошибок; кроме того, фиксируется некоторый вектор параметров модели, принимаемый «истинным», согласно которому генерируется набор зашумленных обучающих выборок, для которых анализируется сходимость параметров, оптимальных для классического и для предлагаемого функционалов качества, к «истинным» в зависимости от параметров шума. Показано, что в подавляющем большинстве

рассмотренных случаев предложенный функционал дает лучшие приближения.

2 Постановка задачи

Дана обучающая выборка D:

$$D = \{\mathbf{x}_i, y_i\} \,|i \in \{1, \ldots, \ell\}\,, \;\; \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}. \quad (1)$$

Для каждой зависимой переменной y_i известно стандартное отклонение ошибки ее измерения σ_{y_i}, а для соответствующего вектора независимых переменных $\mathbf{x}_i$ аналогично известны стандартные отклонения его компонент $\sigma_{x_{ij}}|j \in \{1, \ldots, m\}$. При этом допускается, что близкие точки могут иметь сколь угодно различные ошибки. Кроме того, различные ошибки измерения независимы.

Для удобства введем вектор ошибок измерений зависимых переменных σ_{y_i}:

$$\boldsymbol{\sigma}_y = \{\sigma_{y_1}, \ldots, \sigma_{y_\ell}\}\,.$$

Аналогично введем матрицу ошибок измерений независимых переменных $\sigma_{x_{ij}}$:

$$\Sigma_x = \|\sigma_{x_{ij}}\| \,|i \in \{1, \ldots, \ell\}\,, \;\; j \in \{1, \ldots, m\}\,.$$

Отметим, что эта матрица не является ковариационной матрицей ошибок каждого конкретного объекта из обучающей выборки, поэтому нельзя утверждать, что она является диагональной (и, более того, квадратной).

Пусть выбрана некоторая регрессионная модель $y = f(\mathbf{x}, \boldsymbol{\omega})$, параметризованная вектором $\boldsymbol{\omega}$. Требуется построить функционал ошибки $\breve{S}(\boldsymbol{\omega})$ вектора параметров $\boldsymbol{\omega}$ модели f, учитывающий ошибки измерений $\boldsymbol{\sigma}_y$ и Σ_x:

$$\breve{S}(\boldsymbol{\omega}) = \breve{S}\left(\boldsymbol{\omega}, \boldsymbol{\sigma}_y, \Sigma_x, D\right)\,, \quad (2)$$

и, кроме того, найти вектор параметров ω, минимизирующий функционал (2):

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \breve{S}(\boldsymbol{\omega})\,.$$

3 Модифицированный функционал качества

Воспользуемся следующим качественным соображением: чем больше погрешность определения переменных (зависимых или независимых) для некоторой экспериментальной точки, тем в меньшей степени соответствующий регрессионный остаток должен учитываться при оптимизации параметров

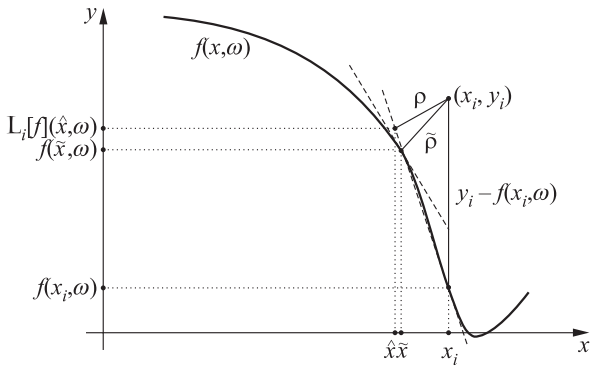

Рис. 1 Различные способы определения расстояния от точки до прямой: $\tilde{\rho}$ — истинное расстояние как минимум расстояния от точки (x_i, y_i) до какой-либо точки на прямой; $y_i - f(x_i, \boldsymbol{\omega})$ — расстояние в классическом функционале среднеквадратичной ошибки в предположении об отсутствии ошибок измерения независимой переменной x; ρ — предлагаемое расстояние

модели. Кроме того, с физической точки зрения складывать можно только величины, имеющие одинаковую размерность, либо безразмерные величины, поэтому необходима соответствующая нормировка невязок по каждой из переменных.

Для упрощения изложения рассмотрим случай одной независимой переменной: $x \in \mathbb{R}$. С учетом приведенных выше соображений введем следующее определение расстояния $\rho(x, i)$ от точки (x_i, y_i) до некоторой точки $(x, f(x, \boldsymbol{\omega}))$ на кривой, описываемой регрессионной моделью $y = f(x, \boldsymbol{\omega})$:

$$\tilde{\rho}^2(x, i) = \frac{(x_i - x)^2}{\sigma_{x_i}^2} + \frac{(y_i - f(x, \boldsymbol{\omega}))^2}{\sigma_{y_i}^2}\,. \quad (3)$$

Непосредственное точное определение расстояния от экспериментальной точки до регрессионной кривой представляется отдельной сложной вычислительной задачей оптимизации (решаемой, например, итерационно), поэтому предлагается рассматривать расстояние от точки не до самой кривой, а до линеаризованной кривой в окрестности этой точки. На рис. 1 показаны различные варианты определения расстояния, при этом в иллюстративных целях размерности и погрешности определения x и y приняты одинаковыми.

Итак, линеаризуем $f(x, \boldsymbol{\omega})$ в окрестности точки $(x_i, f(x_i, \boldsymbol{\omega}))$, обозначив оператор линеаризации в окрестности этой точки $\mathbb{L}_i$:

$$f(x, \boldsymbol{\omega}) \approx \mathbb{L}_i[f](x, \boldsymbol{\omega}) = \\ = f(x_i, \boldsymbol{\omega}) + (x - x_i)\,\frac{\partial f}{\partial x}\,(x_i, \boldsymbol{\omega})\,. \quad (4)$$

Расстояние (3) выражается для линеаризованной функции (4) следующим образом:

$$\rho^2(x,i) = \frac{(x_i - x)^2}{\sigma_{x_i}^2} + \\ + \frac{\left(y_i - f\left(x_i, \boldsymbol{\omega}\right) - \left(\partial f/\partial x\right)\left(x_i, \boldsymbol{\omega}\right)\left(x - x_i\right)\right)^2}{\sigma_{y_i}^2}.$$

Минимизируя это выражение по x:

$$\hat{x} = \arg\min_x \rho^2(x,i),$$

находим расстояние от точки (x_i, y_i) из обучающей выборки до линеаризованной в ее окрестности регрессионной зависимости f при данном векторе параметров $\boldsymbol{\omega}$:

$$\rho^2(f, \boldsymbol{\omega}, i) = \rho^2\left(\hat{x}, i\right) = \\ = \frac{\left(y_i - f(x_i, \boldsymbol{\omega})\right)^2}{\sigma_{y_i}^2 + \left(\partial f/\partial x\right)\left(x_i, \boldsymbol{\omega}\right)^2 \sigma_{x_i}^2}. \quad (5)$$

Отметим, что решение (5) корректируется при последовательном изменении линеаризации в связи с изменением вектора параметров $\boldsymbol{\omega}$ согласно выбранному итерационному методу решения этой задачи.

Аналогично можно получить выражение для расстояния в случае, когда объекты в обучающей выборке представлены m независимыми переменными ($\mathbf{x} \in \mathbb{R}^m$):

$$\rho^2(f, \boldsymbol{\omega}, i) = \frac{\left(y_i - f(\mathbf{x}_i, \boldsymbol{\omega})\right)^2}{\sigma_{y_i}^2 + \sum_{j=1}^m \left((\partial f/\partial x_j)(\mathbf{x}_i, \boldsymbol{\omega})\right)^2 \sigma_{x_{ij}}^2}.$$

Таким образом, предлагаемый функционал, минимизирующий сумму введенных согласно (3) расстояний с учетом их линеаризации, для достаточно гладких функций выглядит следующим образом:

$$\breve{S}(\boldsymbol{\omega}) = \sum_{i=1}^{\ell} \frac{\left(y_i - f(\mathbf{x}_i, \boldsymbol{\omega})\right)^2}{\sigma_{y_i}^2 + \sum_{j=1}^m \left((\partial f/\partial x_j)(\mathbf{x}_i, \boldsymbol{\omega})\right)^2 \sigma_{x_{ij}}^2}. \quad (6)$$

Отметим следующее:

- функционал (6) соответствует классической сумме квадратов регрессионных остатков при условии нормировки квадрата каждого остатка на сумму квадратов погрешности определения зависимой величины σ_{y_i} и произведения частной производной регрессионной модели по j-й компоненте вектора независимых величин на погрешность определения соответствующей компоненты $\sigma_{x_{ij}}$;
- при прочих равных условиях в выражении для расстояния (5) и, соответственно, в функционале (6) с большим весом учитываются те точки, в которых производная регрессионной модели $\partial f/\partial x_j$ по соответствующей компоненте x_j больше, что соответствует соображениям здравого смысла: чем меньше наклон регрессионной зависимости в окрестности данной точки, тем меньше влияние неточного измерения соответствующей независимой переменной на значение регрессионной зависимости в этой точке;
- если все независимые переменные измерены точно, т. е. $\forall i, j : \sigma_{x_{ij}} = 0$, то предложенный функционал переходит в рассмотренный в [8]. Если же, кроме того, все зависимые переменные имеют одну и ту же погрешность σ_y, то предложенный функционал переходит в известную сумму квадратов регрессионных остатков с точностью до некоторого множителя (а именно $1/\sigma_y$), не влияющего на положения минимумов функционала среднеквадратичной ошибки.

Следует отметить возможность вероятностной интерпретации предложенного выражения для расстояния (3). Для случая одной независимой переменной предположим, что вероятность соответствия некоторой точки $(\tilde{x}_i, f(\tilde{x}_i, \boldsymbol{\omega}))$ на регрессионной кривой $y = f(x, \boldsymbol{\omega})$ данной экспериментальной точке (x_i, y_i) описывается двумерным нормальным распределением с центром в этой экспериментальной точке (x_i, y_i) и диагональной ковариационной матрицей

$$\Sigma_i = \begin{Vmatrix} \sigma_{x_i}^2 & 0 \\ 0 & \sigma_{y_i}^2 \end{Vmatrix}$$

(т. е. ошибки измерения каждой координаты независимы):

$$P\left(\tilde{x}_i, f(\tilde{x}_i, \boldsymbol{\omega})\right) \sim \mathcal{N}\left(\left(x_i, y_i\right), \Sigma_i\right) = \frac{1}{2\pi\sqrt{\det \Sigma_i}} \times \\ \times \exp\left\{-\frac{1}{2} \begin{Vmatrix} x_i - \tilde{x}_i \\ y_i - f(\tilde{x}_i, \boldsymbol{\omega}) \end{Vmatrix}^{\mathrm{T}} \Sigma_i^{-1} \begin{Vmatrix} x_i - \tilde{x}_i \\ y_i - f(\tilde{x}_i, \boldsymbol{\omega}) \end{Vmatrix}\right\}.$$

Максимизация логарифма правдоподобия с аналогичной (4) линеаризацией позволяет получить те же выражения (3) и (6). Более подробное рассмотрение такого подхода и следствий из него станет предметом дальнейшей работы.

4 Метод оптимизации предложенного функционала

Для численной оптимизации функционала (6) представим его в виде суммы квадратов регрессионных остатков путем следующего переобозначе-

ния переменных. Вместо выборки (1) рассмотрим выборку

$$\tilde{D} = \{\tilde{\mathbf{x}}_i, \tilde{y}_i\} \,|i \in \{1, \ldots, \ell\},\ \ \tilde{\mathbf{x}}_i \in \mathbb{R}^{m+1}\,,\ \ \tilde{y}_i \in \mathbb{R}\,,$$

где $\tilde{y}_i \equiv 0$, а $\tilde{\mathbf{x}}_i = \{\mathbf{x}_i, y_i\}$ — исходный вектор $\mathbf{x}_i$ с дополнительно приписанным к нему значением y_i. Кроме того, примем

$$\tilde{f}(\tilde{\mathbf{x}}_i, \boldsymbol{\omega}) = \frac{f(\mathbf{x}_i, \boldsymbol{\omega}) - y_i}{\sqrt{\sigma_{y_i}^2 + \sum_{j=1}^{m}((\partial f/\partial x_j)(\mathbf{x}_i, \boldsymbol{\omega}))^2 \sigma_{x_{ij}}^2}}.$$

Тогда минимизация функционала (6) возможна известными методами оптимизации, так как прямой подстановкой можно убедиться, что (6) в этом случае эквивалентен

$$S(\boldsymbol{\omega}) = \sum_{i=1}^{\ell}\left(\tilde{y}_i - \tilde{f}\,(\tilde{\mathbf{x}}_i, \boldsymbol{\omega})\right)^2.$$

Для таким образом преобразованного функционала в качестве базового алгоритма оптимизации может быть использован любой метод решения задачи о наименьших квадратах, как, например, метод градиентного спуска или алгоритм Левенберга–Марквардта [13]. В этом случае при соответствующих условиях гладкости частных производных функции f (что практически всегда выполняется в реальных физических приложениях) сохраняются все свойства исходного алгоритма.

Отметим, что предложенная идея введения весовых коэффициентов, отвечающих разным измерениям и зависящих от точности этих измерений, вообще говоря, применима и для прочих методов решения задачи восстановления регрессии, отличных от символьной регрессии. Подробное рассмотрение этих методов в совокупности с предлагаемым подходом выходит за рамки статьи, однако укажем, что при невозможности выполнить аналитическое дифференцирование функции f предлагается использовать следующий итеративный алгоритм, предназначенный для использования с уже имеющимися реализациями соответствующих методов оптимизации. Предполагается, что реализация «принимает на вход» массив значений y_i, функцию вычисления значения f в точках $\mathbf{x}_i$ с вектором параметров $\boldsymbol{\omega}$.

Алгоритм выглядит следующим образом.

1. Выбирается некоторое начальное приближение вектора параметров $\boldsymbol{\omega}$.
2. Для каждой пары $(\mathbf{x}_i, y_i)$ из обучающей выборки численно или аналитически рассчитывается значение частной производной $\partial f/\partial x$ в точке $(\mathbf{x}_i, \boldsymbol{\omega})$.
3. Каждое значение зависимой переменной y_i и значение функции $f(\mathbf{x}_i, \boldsymbol{\omega})$ нормируется на соответствующую величину
$$\sigma_{y_i}^2 + \sum_{j=1}^{m}\left(\frac{\partial f}{\partial x_j}(\mathbf{x}_i, \boldsymbol{\omega})\right)^2 \sigma_{x_{ij}}^2\,.$$
4. Выполняется итерация классического алгоритма оптимизации для таким образом модифицированных значений функции f и зависимых переменных y_i, получая новое значение вектора $\boldsymbol{\omega}$.
5. Если критерий останова не достигнут, алгоритм продолжает выполнение с п. 2.

Отметим следующее:

– критерием останова могут служить обычные критерии, такие как достижение некоторого числа итераций, порог нормы изменения вектора $\boldsymbol{\omega}$ и т. п.;

– если известно, что производная $\partial f/\partial x$ является достаточно гладкой в окрестности $(\mathbf{x}_i, \boldsymbol{\omega}) \mid i \in \in \{1, \ldots, \ell\}$, на шаге 4 алгоритма представляется разумным выполнить сразу несколько итераций классического алгоритма во избежание потенциально ресурсоемкого пересчета производных и перенормировки значений y_i и f.

5 Вычислительный эксперимент

В вычислительном эксперименте рассматриваются данные, полученные в ходе измерения зависимости интенсивности излучения I лазера от прозрачности его резонатора. Изучался лазер высокого давления ($\approx$ 3 атм He, $\approx$ 60 Topp Ne, $\approx$ 20 Topp Ar) на $3p$–$3s$ переходах неона (основной переход — 585 нм), возбуждаемый электронным пучком [14].

Пусть насыщающая переход интенсивность излучения — I_s, наблюдаемая интенсивность — I_l. В таком случае для безразмерной величины $y = I_l/I_s$ с учетом однородного уширения линии усиления при высоком давлении газа и хорошей однородности возбуждения, обеспечиваемой электронным пучком, можно получить нелинейное уравнение [15]:

$$\alpha_0 L - \frac{1}{2}\ln R_0 = g_0 L \frac{1+\sqrt{R_0}}{1-\sqrt{R_0}}\,\frac{1}{y} \times \\ \times \ln\left(1 + \frac{y(1-\sqrt{R_0})/(1+\sqrt{R_0})}{1 + y(2\sqrt{R_0})/(1-R_0)}\right), \quad (7)$$

где α_0 — распределенные потери (например, на рассеяние света); g_0 — коэффициент усиления слабого

сигнала; R_0 — коэффициент отражения выходного зеркала лазера. Однородность накачки означает, что g_0 и α_0 одинаковы по всему объему с хорошей точностью.

Значение R_0 является независимой переменной, изменяемой экспериментаторами, и в данном разделе также обозначается x сообразно остальной части работы.

Для достаточно больших R_0, близких к единице (фактически для $R_0 \geq 0{,}6 \ldots 0{,}7$), можно упростить (7), заменив $2\sqrt{R_0} \approx 1 + R_0$ и получив хорошо известное выражение [15]:

$$y(R_0) = \gamma \frac{1 - R_0}{1 + R_0} \left(\frac{g_0}{\alpha_0 - (1/(2L)) \ln R_0} - 1 \right), \quad (8)$$

где γ — нормировочный коэффициент.

В рассматриваемом физическом эксперименте длина активной среды L — 150 см, точность определения мощности лазера y имеет относительную погрешность в 2%, точность определения прозрачности R_0 имеет абсолютную погрешность и составляет 0,01 при $R_0 \geq 0{,}6$ и 0,02 при $R_0 < 0{,}6$.

В ходе измерений получены значения $y(R_0)$, приведенные в табл. 1.

Таблица 1 Экспериментальные значения $y(R_0)$

R_0	y
0,48	3,25
0,56	10,2
0,65	16,5
0,73	20,5
0,80	22,5
0,87	23,2
0,94	18,2

Таким образом, решается задача минимизации функционала (6) при

$$\begin{gathered} \boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3) = (\gamma, \alpha_0, g_0) \ ; \\ f(x, \boldsymbol{\omega}) = y\left(R_0, \gamma, \alpha_0, g_0\right) \ ; \\ \left.\begin{aligned} \sigma_{y_i} &= 0{,}02 y_i; \\ \sigma_{x_i} &= \begin{cases} 0{,}01 & | \ x_i \geq 0{,}6\,; \\ 0{,}02 & | \ x_i < 0{,}6\,. \end{cases} \end{aligned} \right\} \end{gathered} \quad (9)$$

5.1 Оптимальные параметры модели

Кроме предложенного в настоящей работе функционала (6) рассмотрен также и классический функционал среднеквадратичной ошибки:

$$S = \sum_{i=1}^{\ell} \left(y_i - f\left(x_i, \boldsymbol{\omega}\right)\right)^2 . \quad (10)$$

Таблица 2 Оптимальные значения параметров модели

Параметры	$\boldsymbol{\omega}$	$\boldsymbol{\omega}^0$	$\frac{\lvert\omega_i - \omega_i^0\rvert}{\omega_i^0}$
g_0	$2{,}93 \cdot 10^{-3}$	$2{,}92 \cdot 10^{-3}$	0,31%
α_0	$2{,}07 \cdot 10^{-4}$	$2{,}22 \cdot 10^{-4}$	6,59%
γ	98,6	101,5	2,9%
$\check{S}$ (6)	0,542	0,645	16%
S (10)	0,328	0,183	80%

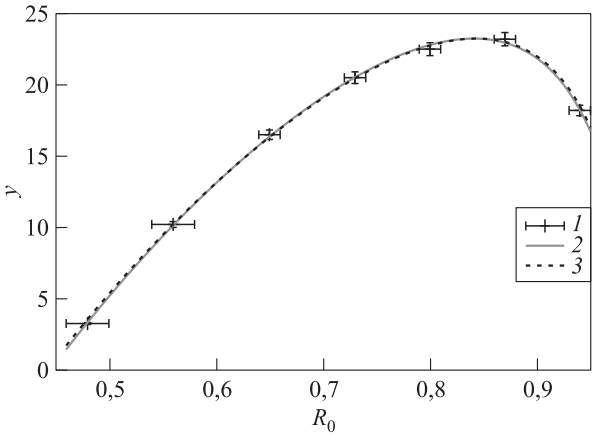

Рис. 2 Графики (8), соответствующие параметрам, минимизирующим (6) и (10): *1* — экпериментальные данные; *2* — ω^0; *3* — ω

В табл. 2 приведены значения параметров $\boldsymbol{\omega}$ и $\boldsymbol{\omega}^0$ функции (8), минимизирующие (6) и (10) соответственно, а также относительные разности их компонент. Кроме того, приведены значения функционалов (6) и (10) для обоих векторов параметров.

Отдельно отметим, что сравнивать непосредственные значения функционалов (6) и (10) не имеет смысла. Вместо этого необходимо сравнивать различные модели по каждому из этих функционалов в отдельности. Так, результаты, приведенные в табл. 2, показывают вполне естественный результат: каждый из двух векторов параметров ($\boldsymbol{\omega}$ и $\boldsymbol{\omega}^0$) является оптимальным лишь для того функционала, который он минимизирует.

Графики регрессионной модели (8), соответствующие $\boldsymbol{\omega}$ и $\boldsymbol{\omega}^0$, приведены на рис. 2.

5.2 Сходимость оптимальных параметров к классическим

Численно исследована зависимость сходимости параметров $\boldsymbol{\omega}$ к параметрам $\boldsymbol{\omega}^0$, получаемым минимизацией функционалов (6) и (10) соответственно, от погрешности σ_y измерения зависимой переменной y.

Следует ожидать, что при увеличении погрешности измерения величины y при фиксированной погрешности измерения R_0 оптимальный вектор $\boldsymbol{\omega}$ будет приближаться к $\boldsymbol{\omega}^0$, так как тем более незначителен вклад ошибки измерения независимой переменной.

Рассматриваются два случая.

1. Погрешность i-го измерения y_i задается как $\sigma_{y_i} = 0{,}02ky_i$, т. е. погрешность зависит от значения самого y_i.
2. Погрешность i-го измерения y_i задается как $\sigma_{y_i} = 0{,}02ky_{\max}$, т. е. погрешность от значения конкретного y_i не зависит. Заметим, что выбор конкретного значения y, определяющего погрешность, является в данном случае достаточно произвольным и соответствует умножению всех погрешностей на некоторую константу (что нивелируется соответствующим изменением выбора диапазона k).

В первом случае ошибки измерения y распределены неодинаково; следовательно, применение стандартного МНК не обосновано. В то же время во втором случае ошибки принадлежат одному и тому же распределению и, кроме того, независимы, поэтому в данном случае МНК-оценка применима (с точностью до ошибки измерения независимой переменной).

Для обоих случаев подробно рассматривалась область $k \in [1; 100]$, значение k изменялось с шагом 0,01. Отметим, что уже при $k \approx 25$ характерная погрешность измерения величины y сопоставима с самой величиной y, а при $k > 50$ превышает ее.

Результаты приведены на рис. 3. На графиках отображены компоненты вектора $\boldsymbol{\omega}$, нормированные на соответствующие значения $\boldsymbol{\omega}^0$, в зависимости от значения k.

В случае фиксированной погрешности σ_{y_i} значения $\boldsymbol{\omega}$ действительно стремятся к $\boldsymbol{\omega}^0$ для разумных значений k, а в случае гетероскедастичных ошибок такой зависимости не наблюдается, хотя значения $\boldsymbol{\omega}$ и оказываются достаточно близки к $\boldsymbol{\omega}^0$. По мнению автора, такое поведение вектора оптимальных параметров является вполне ожидаемым и демонстрирует несостоятельность классического функционала качества в случае неодинаково распределенных ошибок.

5.3 Сходимость параметров к истинным

Численно исследована зависимость сходимости параметров $\boldsymbol{\omega} = \arg\min \breve{S}$ и $\boldsymbol{\omega}^0 = \arg\min S$ к некоторому «истинному» значению вектора параметров $\hat{\boldsymbol{\omega}}$ от числа точек ℓ в обучающей выборке и от погрешности определения независимой переменной.

Для этого вектор параметров $\boldsymbol{\omega}$, полученный минимизацией обучающей выборки из табл. 1, принимается за некоторый «истинный» вектор параметров $\hat{\boldsymbol{\omega}}$ и на каждой j-й итерации генерируется обучающая выборка $D_j(\ell, k)$:

$$D_j(\ell,k) = \\ = \{(x_i + \xi_i^x, y(x_i, \hat{\boldsymbol{\omega}}) + \xi_i^y)\} \mid \xi_i^x \sim \mathcal{N}(0, k\sigma_{x_i}), \\ \xi_i^y \sim \mathcal{N}(0, \sigma_{y_i}),\ \ i \in \{1, \ldots, \ell\},$$

где $y(x, \boldsymbol{\omega})$ задано соотношением (8), а σ_{x_i} и σ_{y_i} определяются соотношениями (9).

Иными словами, генерируется обучающая выборка согласно искомой модели с известным и фиксированным вектором параметров, которая затем зашумляется нормально распределенными случайными величинами. При этом стандартное отклонение шума для зависимой величины совпадает

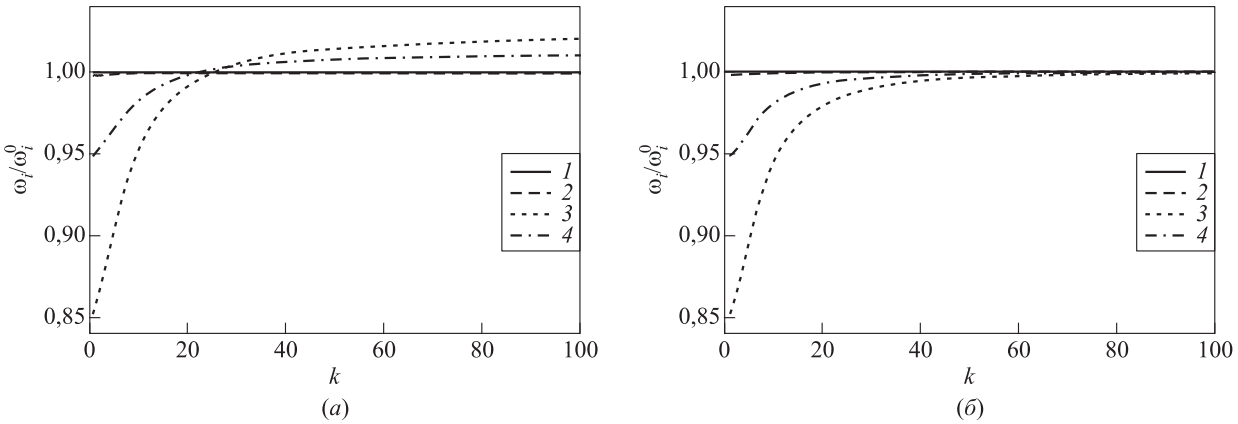

Рис. 3 Зависимость оптимальных параметров от $k \in [1; 100]$: (*а*) $\sigma_{y_i} = 0{,}02ky_i$; (*б*) $\sigma_{y_i} = 0{,}02ky_{\max}$; *1* — классическое значение; *2* — g_0; *3* — α_0; *4* — γ

с экспертно предложенной погрешностью измерений для реального эксперимента, а стандартное отклонение независимой величины отличается от экспертно предложенной погрешности для этой величины в k раз.

После генерации выборки $D_j(\ell, k)$ по ней находятся значения $\boldsymbol{\omega}_j = \arg\min \breve{S}(D_j)$ и $\boldsymbol{\omega}_j^0 = \arg\min S(D_j)$, что повторяется N раз, и $\forall\, i$ рассматриваются значения:

$$\overline{\delta\omega_i} = \frac{\sum_{j=1}^{N}(\omega_{ji} - \hat{\omega}_i)}{N};$$

$$\overline{\delta\omega_i^0} = \frac{\sum_{j=1}^{N}(\omega_{ji}^0 - \hat{\omega}_i)}{N}.$$

Обозначим, кроме того, $\overline{\delta\boldsymbol{\omega}} = \{\overline{\delta\omega_1}, \overline{\delta\omega_2}, \overline{\delta\omega_3}\}$.

Таким образом, при варьировании k и изучении поведения $\overline{\delta\boldsymbol{\omega}}$ и $\overline{\delta\boldsymbol{\omega}^0}$ исследуется влияние

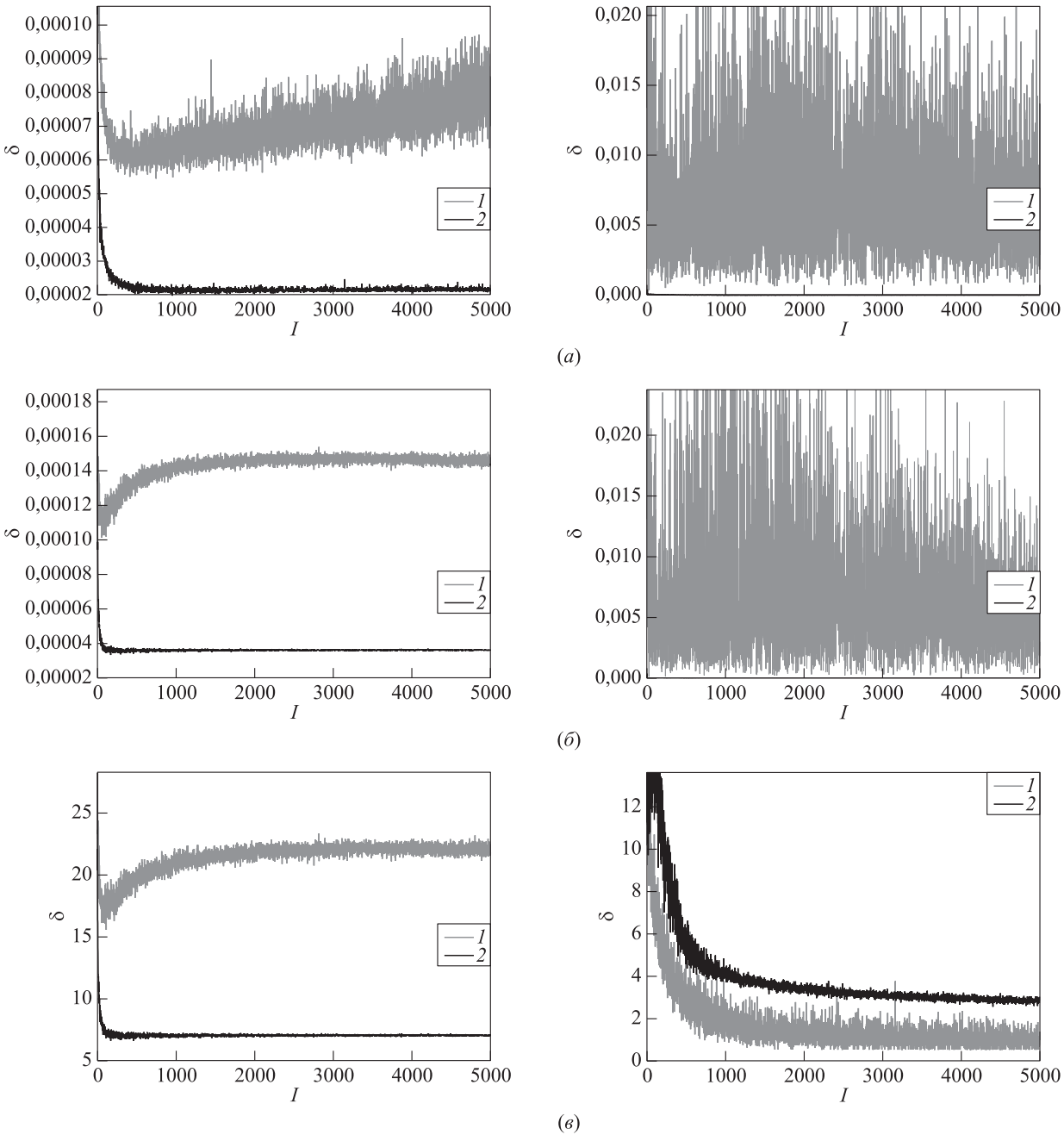

Рис. 4 Сходимость параметров g_0 (*а*), α_0 (*б*) и γ (*в*) к истинным при $k = 0{,}2$ (левый столбец) и 0,65 (правый столбец): *1* — ω^0; *2* — ω

погрешности определения независимой переменной на разность между $\hat{\omega}$ и оптимальными параметрами ω и ω^0 согласно (6) и (10) соответственно.

Так как при уменьшении k монотонно уменьшается погрешность определения независимой переменной, естественно ожидать, что различие между ω и ω^0 будет уменьшаться. Однако вычислительный эксперимент демонстрирует, что это не так.

В проведенном эксперименте $N = 1000$, $\ell \in \in \{10; \ldots; 5000\}$, $k \in \{0{,}2; 0{,}25; \ldots; 1\}$.

Результаты представлены на рис. 4 и 5.

Результаты на рис. 4 (левый столбец) характерны для всех значений $k \in [0{,}2; 0{,}6]$. Однако при $k \geq 0{,}65$ поведение оптимальных параметров резко меняется. Так, на рис. 4 (правый столбец) приведены графики сходимости для $k = 0{,}65$. Видно, что поведение параметров, оптимизированных согласно классическому функционалу качества S, является существенно более хаотическим, что может говорить о меньшей устойчивости [16] модели, оптимизированной согласно S.

Более того, для оценок параметров g_0 и α_0 соответствующее приближение на несколько порядков хуже, чем полученное минимизацией $\check{S}$, вплоть до того, что кривые, соответствующие минимизирующим $\check{S}$ параметрам, практически не видны на графиках (см. рис. 4, *а* и 4, *б*, правый столбец), поскольку в выбранном масштабе они практически совпадают с осью абсцисс. С другой стороны, важно отметить, что оценка параметра γ, полученная минимизацией S, является несколько лучшей для $k = 0{,}65$ (для $k = 0{,}7$ график выглядит аналогично), но минимизация $\check{S}$ дает все лучшие и лучшие приближения с ростом k (см. рис. 5).

Отметим следующее:

– практически во всех случаях (кроме оценки γ для $k = 0{,}8$) предложенный в настоящей работе функционал (6) дает лучшее приближение, в том числе при разумно малом объеме обучающей выборки. Кроме того, в подавляющем большинстве случаев предпочтительность предложенного функционала сохраняется и для большего числа экспериментальных точек;

– для малых $k \leq 0{,}6$ ошибка оценки параметров при помощи классического функционала (10) имеет ярко выраженный минимум в окрестности 60–100 для α_0 и γ и 400 для g_0 экспериментальных точек (см. рис. 4, левый столбец). Дальнейшее увеличение обучающей выборки ведет к ухудшению приближения, получаемого минимизацией (10);

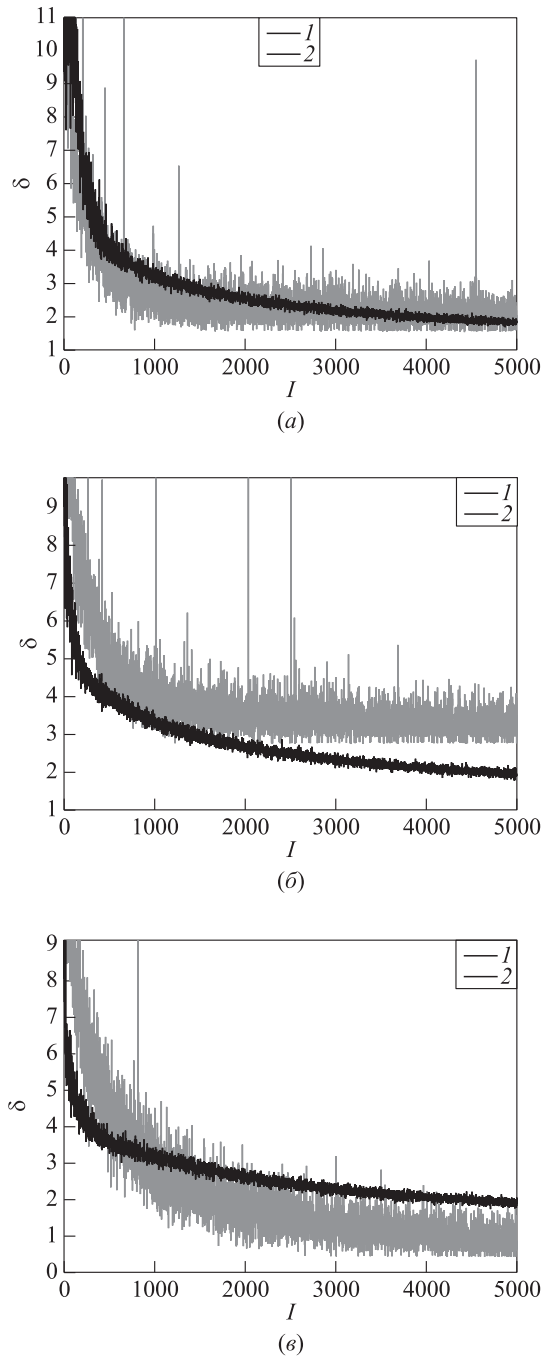

Рис. 5 Сходимость параметра γ при $k = 0{,}8$ (*а*), 0,9 (*б*) и 1,0 (*в*): *1* — ω^0; *2* — ω

– для некоторых k ошибка приближения, получаемого минимизацией предложенного функционала (6), имеет явную горизонтальную асимптоту (см. рис. 4, *а*, 4, *б* и 4, *в* (левый столбец)).

Причины подобного поведения оптимальных параметров являются предметом дальнейших исследований.

6 Заключение

Предложен модифицированный функционал среднеквадратичной ошибки для задач регрессии, применимый в случае наличия ошибок измерения независимых переменных и различных распределений, к которым принадлежат ошибки, в разных точках обучающей выборки. Предложена вероятностная интерпретация этого функционала для случая нормального распределения ошибок измерения.

Показана сходимость предложенного функционала к классическому функционалу среднеквадратичной ошибки для случая гомоскедастичности погрешностей зависимой переменной и пренебрежимо малой погрешности измерения независимых переменных.

Исследовано поведение оптимального вектора параметров для предлагаемого функционала в зависимости от параметров распределений ошибок независимых переменных, в том числе в сравнении с вектором параметров, минимизирующим классический функционал качества.

Представляется разумным использовать предложенный в настоящей работе функционал при оптимизации параметров регрессионных моделей и анализе их устойчивости к погрешностям как зависимых, так и независимых переменных [2, 16].

Литература

1. *Гладун А. В.* Лабораторный практикум по общей физике. — М.: МФТИ, 2004. 316 с.
2. *Рудой Г. И.* О возможности применения методов Монте-Карло в анализе нелинейных регрессионных моделей // Сиб. ж. вычисл. мат., 2015. Т. 4. С. 425—434.
3. *Gillard J. W.* 2006. An historical overview of linear regression with errors in both variables. Cardiff University School of Mathematics. Technical Report.
4. *Deming W. E.* Statistical adjustment of data. — New York, NY, USA: Wiley, 1943. 216 p.
5. *Bowden R. J., Turkington D. A.* Instrumental variables. — Cambridge: Cambridge University Press, 1990. 236 p.
6. *Bekker P. A.* Comment on identification in the linear errors in variables model // Econometrica, 1986. Vol. 54. No. 1. P. 215–217.
7. *Carroll R. J., Ruppert D, Stefanski L. A., Crainiceanu C. M.* Measurement error in nonlinear models: A modern perspective. — London: Chapman and Hall/CRC, 2006. 484 p.
8. *Jukić D.* On nonlinear weighted least squares estimation of bass diffusion model // Appl. Math. Comput., 2013. Vol. 219. No. 14. P. 7891–7900.
9. *Jukić D., Marković D.* On nonlinear weighted errors-in-variables parameter estimation problem in the three-parameter Weibull model // Appl. Math. Comput., 2010. Vol. 215. No. 10. P. 3599–3609.
10. *Kiryati N., Bruckstein A. M.* Heteroscedastic hough transform (HtHT): An efficient method for robust line fitting in the ‘errors in the variables’ problem // Comput. Vis. Image Und., 2000. Vol. 78. No. 1. P. 69–83.
11. *Boggs P. T., Byrd R. H., Schnabel R. B.* A stable and efficient algorithm for nonlinear orthogonal distance regression // SIAM J. Sci. Stat. Comp., 1987. Vol. 8. No. 6. P. 1052–1078.
12. *Marquardt D. W.* An algorithm for least-squares estimation of non-linear parameters // J. Soc. Ind. Appl. Math., 1963. Vol. 11. No. 2. P. 431–441.
13. *King D. E.* Dlib-ml: A machine learning toolkit // J. Mach. Learn. Res., 2009. Vol. 10. P. 1755–1758.
14. *Александров А. Ю., Долгих В. А., Рудой И. Г., Сорока А. М.* Кинетика возбуждаемого электронным пучком лазера высокого давления на «желтой» линии неона // Квантовая электроника, 1991. Т. 18. № 9. С. 1029–1033.
15. *Champagne L. F.* Transient optical absorption in the ultraviolet // *Applied atomic collision physics. Vol. 3: Gas lasers* / Eds. E. W. McDaniel, W. L. Nighan. — Amsterdam, Netherlands: Elsevier, 1982. 349–386.
16. *Rudoy G. I.* Analysis of the stability of nonlinear regression models to errors in measured data // Pattern Recognit. Image Anal., 2016. Vol. 26. No. 3. P. 608–616.

Поступила в редакцию 15.09.16

ON MODIFICATION OF THE MEAN SQUARED ERROR LOSS FUNCTION FOR SOLVING NONLINEAR HETEROSCEDASTIC ERRORS-IN-VARIABLES PROBLEMS

G. I. Rudoy

Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation

Abstract: The paper considers the problem of finding the optimal parameters of a nonlinear regression model accounting for errors in both dependent and independent variables. The errors of different measurements are assumed to belong to different probability distributions with different variances. A modified mean squared error-based loss function is derived and analyzed for this case. In the computational experiment, the measurements of the laser's radiation power as a nonlinear function of the resonator's transparency are used to compare the parameters vectors minimizing the presented loss function and the classical mean squared error. The convergence of the parameters minimizing the presented loss function to the optimal parameters for the classical loss function is studied. In addition, some values of the parameters are considered to be "true" ones and are used to generate synthetic data using the physical model and Gaussian noise, which is then used to study the convergence of the parameters minimizing the presented and the classical loss function, respectively, as the function of the noise parameters.

Keywords: errors-in-variables models; heteroscedastic errors; symbolic regression; nonlinear regression

DOI: 10.14357/19922264170209

References

1. Gladun, A. V. 2004. *Laboratornyy praktikum po obshchey fizike* [Laboratory classes on general physics]. Moscow: MFTI. 316 p.
2. Rudoy, G. I. 2015. Applying Monte Carlo methods to analysis of nonlinear regression models. *Numer. Anal. Appl.* 4:344–350.
3. Gillard, J. W. 2006. An historical overview of linear regression with errors in both variables. Cardiff University School of Mathematics. Technical Report.
4. Deming, W. E. 1943. *Statistical adjustment of data*. New York, NY: Wiley. 216 p.
5. Bowden, R. J., D. A. Turkington. 1990. *Instrumental variables*. Cambridge: Cambridge University Press. 236 p.
6. Bekker, P. A. 1986. Comment on identification in the linear errors in variables model. *Econometrica* 54(1):215–217.
7. Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. *Measurement error in nonlinear models: A modern perspective*. London: Chapman and Hall/CRC. 484 p.
8. Jukić, D. 2013. On nonlinear weighted least squares estimation of bass diffusion model. *Appl. Math. Comput.* 219(14):7891–7900.
9. Jukić, D., and D. Marković. 2010. On nonlinear weighted errors-in-variables parameter estimation problem in the three-parameter Weibull model. *Appl. Math. Comput.* 215(10):3599–3609.
10. Kiryati, N., and A. M. Bruckstein. 2000. Heteroscedastic hough transform (HtHT): An efficient method for robust line fitting in the errors in the variables' problem. *Comput. Vis. Image Und.* 78(1):69–83.
11. Boggs, P. T., R. H. Byrd, and R. B. Schnabel. 1987. A stable and efficient algorithm for nonlinear orthogonal distance regression. *SIAM J. Sci. Stat. Comp.* 8(6):1052–1078.
12. Marquardt, D. W. 1963. An algorithm for least-squares estimation of non-linear parameters. *J. Soc. Ind. Appl. Math.* 11(2):431–441.
13. King, D. E. 2009. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* 10:1755–1758.
14. Aleksandrov, A. Yu., V. A. Dolgikh, I. G. Rudoy, and A. M. Soroka. 1991. Kinetics of a high-pressure electron-beam-excited laser emitting the "yellow" neon line. *Sov. J. Quantum Electronics* 21(9):933–937.
15. Champagne, L. F. 1982. Transient optical absorption in the ultraviolet. *Applied atomic collision physics. Vol. 3: Gas lasers*. Eds. E. W. McDaniel and W. L. Nighan. Amsterdam, Netherlands: Elsevier. 349–386.
16. Rudoy, G. I. 2016. Analysis of the stability of nonlinear regression models to errors in measured data. *Pattern Recognit. Image Anal.* 26(3):608–616.

Received September 15, 2016

Contributor

Rudoy Georg I. (b. 1991) — PhD student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation; 0xd34df00d@gmail.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 85–100

ПЕРСОНАЛЬНАЯ ОТКРЫТАЯ СЕМАНТИЧЕСКАЯ ЦИФРОВАЯ БИБЛИОТЕКА LibMeta. КОНСТРУИРОВАНИЕ КОНТЕНТА. ИНТЕГРАЦИЯ С ИСТОЧНИКАМИ LOD*

О. М. Атаева[1], В. А. Серебряков[2]

Аннотация: Развитие семантических технологий вывело цифровые библиотеки на уровень, на котором на первый план выступила необходимость осмысленного представления контента цифровых библиотек. Одновременно возникает необходимость ограничения его в терминах некоторой предметной области. В работе рассматривается конструирование контента библиотеки для некоторой предметной области в рамках разработанной системы LibMeta. Персональная открытая семантическая цифровая библиотека LibMeta с системой поддержки работы пользователей с цифровыми ресурсами библиотек и их коллекциями для некоторой предметной области, ограниченной терминологически с помощью тезауруса, предоставляет функциональность конструирования контента библиотеки согласно определенным требованиям и требует всего лишь произвести начальную настройку системы под конкретную предметную область, ограниченную терминологически с помощью тезауруса. В качестве примера предметной области в работе используется узкоспециализированный тезаурус обыкновенных дифференциальных уравнений (ОДУ).

Ключевые слова: семантические библиотеки; модель данных; онтологии; источники данных; поиск в LOD

DOI: 10.14357/19922264170210

1 Введение

Взрывное развитие технологий в последние десятилетия повлияло на все аспекты деятельности человека. Накопленные в библиотеках данные стали через сеть доступны широкому кругу пользователей, удовлетворяя информационные потребности которых, разработчики расширяли функциональность цифровых библиотек.

Развитие семантических технологий вывело цифровые библиотеки на новый уровень, на котором на первый план выступила необходимость осмысленного представления контента цифровых библиотек. В решении этих задач ключевую роль стали играть онтологии [1], позволяя представлять концептуальные модели для описания самого контента этих библиотек, основываясь на ранее разработанных форматах описания, таких как MARC[3]. Такие онтологии получили название библиографических, дополняя семантикой эти форматы. Фактически в библиографических онтологиях фиксируются ключевые понятия объектов, составляющих наполнение библиотеки, и связи между ними. Этих понятий достаточно для описания обычной классической цифровой библиотеки для любой предметной области, в которой представлена информация о различных печатных изданиях и, возможно, их электронные версии. Но развитие семантических библиотек [2] способствует расширению модели, определяющей наполнение библиотеки, в которой теперь могут содержаться самые различные типы объектов.

Одновременно с расширением модели библиотечного наполнения возникает необходимость ограничения его в рамках некоторой предметной области. Для этого вводится набор терминов, используемых для описания этой предметной области. Чаще всего эти термины организованы в виде некоторой таксономии с поддержкой разнообразных связей между ними. В дальнейшем будем называть наполнение библиотеки с такой терминологической поддержкой некоторой предметной области контентом семантической цифровой библиотеки, или просто контентом.

Для тематической классификации ресурсов библиотеки используются различные классифи-

*Работа выполнена при финансовой поддержке РФФИ (проект 14-07-00058 А).

[1]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, oli@ultimeta.ru

[2]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, serebr@ultimeta.ru

[3]http://www.loc.gov/marc/unimarctomarc21.html.

каторы, которые отличаются друг от друга охватом предметных областей и степенью гранулярности при классификации этих областей. Для этих целей может использоваться один из широко распространенных классификаторов, таких как УДК[1] (универсальная десятичная классификация), ББК[2] (библиотечно-библиографическая классификация), ГРНТИ[3] (государственный рубрикатор научно-технической информации). Эти классификаторы охватывают почти все области научного знания и перечень понятий, характерных для этих областей. Обычно эти понятия носят довольно общий характер и не отражают всего разнообразия направлений в каждой отдельной области научного знания.

Специализированные по конкретным областям библиотеки используют обычно свои классификаторы для систематизации своих ресурсов. Такой подход обеспечивает более детальный анализ содержания документов и соотношение смысловых понятий в документе с определенным направлением специализированной области знания. К таким классификаторам можно, например, отнести MSC[4] (Mathematics Subject Classification), который используется для классификации разделов математики.

Семантические библиотеки предоставляют своим пользователям большой арсенал возможностей для удовлетворения их информационных потребностей [2]. Это разнообразные средства поиска: атрибутный поиск, полнотекстовый поиск, поиск по коллекциям на основе тематических классификаторов, поиск по разнообразным типам ресурсов, включенных в библиотеку. Для пользователей семантических библиотек, являющихся активными потребителями информации, во многих современных решениях предоставляется возможность создать собственную коллекцию.

Возникает необходимость дать пользователям специфицировать свои предпочтения, развивая возможность определения собственных терминов в рамках некоторого направления научного знания, уточняя и очерчивая круг своих интересов, позволяя организовывать группы пользователей со сходными интересами для возможности отслеживания всей информации по определенным направлениям.

Широкое применение онтологий позволяет интегрировать данные библиотек с данными из различных источников, основываясь на их семантике [3]. Эти источники не обязательно сами являются библиотеками. Множество таких источников подключено к облаку LOD (Linked Open Data) [4]. Основная идея LOD заключается в решении задач интеграции данных, представленных в сети, для чего предлагается представить информацию в формализованном виде с помощью онтологий, что делает ее доступной для машинной обработки. В этих источниках данных провязаны самые различные типы ресурсов, которые представляют интерес для пользователей библиотек с точки зрения обогащения данных как структурно, так и семантически.

На основе модели понятий, описанной в предыдущих работах [5], а также идей Semantic Web и LOD была разработана персональная открытая семантическая цифровая библиотека LibMeta с системой поддержки работы пользователей с цифровыми ресурсами библиотек и их коллекциями для некоторой предметной области, ограниченной терминологически с помощью тезауруса [3, 6].

2 LibMeta — основные идеи

При реализации LibMeta авторы руководствовались набором основных задач, которые должна решать разрабатываемая система:

(1) библиотека должна поддерживать возможность использования медийных объектов или ссылки на них при описании своих объектов, включая текст, аудио-, видеофайлы или любую их комбинацию. Это требование отражается в названии словом «цифровая»;

(2) типы используемых ресурсов и связи между ними должны быть описаны средствами системы в рамках определенных в предыдущей работе понятий, составляющих семантическое описание ресурсов контента библиотеки. При этом согласно принципам LOD при описании ресурсов поддерживается использование классов и свойств ранее используемых онтологий в сообществе, поддерживающем LOD. Эта поддержка выражается либо в непосредственном использовании готовых онтологий при описании ресурсов и связей между ними, либо возможностью ссылок на их элементы, используя связи на уровне описания ресурсов. Это требование отражается в названии словом «семантическая»;

(3) библиотека должна служить интеграционным узлом, предоставляя возможность связывания своих данных с данными из разных источников, которые включены в облако LOD. Должна

[1] http://nlib.sakha.ru/Cataloque/udk/index.shtml.
[2] http://roslavl.library67.ru/files/382/bbk.pdf.
[3] http://www2.viniti.ru/index.php?option=com_content&view=article&id=39:rubrikator-nti.
[4] http://www.ams.org/msc/pdfs/classifications2010.pdf.

также обеспечиваться возможность извлекать данные этой библиотеки в машиночитаемом формате. Это требование отражается в названии словом «открытая»;

(4) пользователи библиотеки должны иметь возможность организовывать свои коллекции по интересующему их научному направлению, добавляя новые термины в предметный тезаурус, уточняя таким образом область своих интересов. Пользователи должны также иметь возможность осуществлять поиск не только среди объектов в рамках системы, но и по источникам данных без необходимости использования специализированного языка для поисковых запросов. Это требование отражается в названии словом «персональная».

Основные требования, предъявляемые при этом к контенту системы, — универсальность, структурированность, адаптируемость — не противоречат этим свойствам и обеспечивают поддержку настраиваемого хранилища метаданных для объектов и расширяемый набор информационных ресурсов. Универсальность обеспечивает описание типов ее ресурсов и объектов независимо от предметной области и области интересов пользователей. Структурированность описания обеспечивает поддержку связей между различными типами ресурсов как внутри системы, так и вне ее, исходя из определений LOD. Адаптируемость описания ресурсов обеспечивает возможность добавления новых свойств и связей в процессе развития системы и обеспечивает настройку пользовательских интерфейсов под эти изменения.

Фактически LibMeta предоставляет функциональность конструирования контента библиотеки согласно этим требованиям, и на начальном этапе при установке системы требуется всего лишь произвести настройку системы под конкретную предметную область, описав ее ресурсы и таксономии, которые будут очерчивать тематически предметную область ее ресурсов и таким образом составлять ее тезаурус.

3 LibMeta — первый пример конструирования

Рассмотрим простой пример реализации библиотеки LibMeta, основанной на данных публикаций из электронной библиотеки «Научное наследие России» [7]. Основных типов ресурсов, которые определены для этих данных, всего два: персоны и публикации. Для тематической классификации этих публикаций используется классификатор ГРНТИ, и каждая публикация снабжена номером УДК.

Авторы не ставили своей целью создание уменьшенной копии «Научного наследия». Основная цель, преследуемая в контексте предлагаемой системы, — это связывание этих данных с данными, опубликованными в LOD, и их публикация для возможности доступа к ним других систем. В качестве источника данных для связывания в этом примере используется DBpedia[1], служащая ядром LOD.

Итак, основная цель при конструировании описания контента заключается в том, чтобы представленное описание по возможности максимально облегчало реализацию процедуры поиска данных в узлах LOD. Жертвой этой идеи становится, возможно, некоторая упрощенность структуры контента, в отличие от выразительности, представляемой средствами языка OWL[2], как будет показано ниже, но при этом получаем гибкость при построении интеграционного узла для различных типов ресурсов, описание которых можно расширять в процессе жизнедеятельности системы.

Фактически понятия *персоны* и *публикации* представляют собой экземпляры класса *информационный ресурс*, определенного как базовая единица контента семантической библиотеки. Так как каждый ресурс обладает набором атрибутов, для каждого из этих экземпляров задается собственный набор из множества атрибутов, предварительно описанных в системе. Множество атрибутов для информационных ресурсов состоит из следующих элементов: *название на языке оригинала, название на русском, фамилия, имя, отчество, электронный адрес, дата рождения, аннотация, идентификатор, автор, деятельность, тип публикации, место рождения, биография, описание, дополнительное заглавие, язык*.

Конкретные персоны — это объекты, представляющие экземпляры класса *информационный объект*, они определяются информационным ресурсом *персона* и представляются значениями атрибутов соответствующего ресурса. Помимо свойств, заданных атрибутами, представленными в наборе атрибутов своего информационного ресурса, каждый объект обладает также свойствами, общими для всех информационных объектов, такими как *теги, описание, дата создания, дата изменения, владелец, уникальный идентификатор*.

На рис. 1 приведена упрощенная схема, сконструированная для этих типов ресурсов. На схе-

[1] http://dbpedia.org.
[2] https://www.w3.org/2001/sw/wiki/OWL.

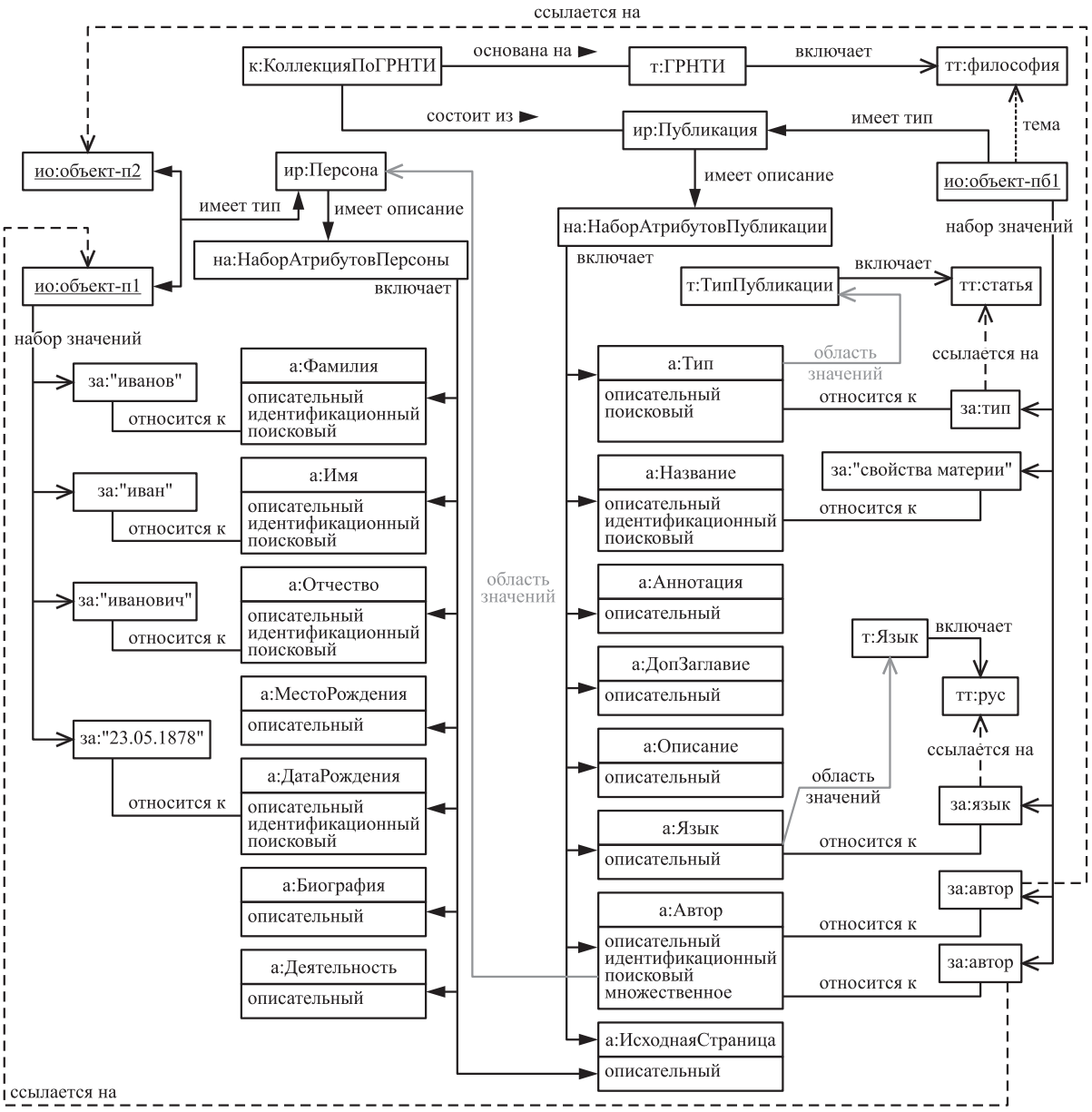

Рис. 1 Конструирование информационных ресурсов

ме проиллюстрированы связи между экземплярами информационных ресурсов *персона* и *публикация* и конкретными экземплярами класса информационного объекта (названия объектов *объект-п1*, *объект-п2*, *объект-пб1* подчеркнуты). Префиксы «ио», «ир», «к», «т», «тт», «на», «а», отделяемые двоеточием, указывают на принадлежность экземпляра соответственно к классам *информационный объект, информационный ресурс, коллекция, таксономия, таксон, набор атрибутов, атрибут*.

Для тематической классификации объектов публикации используется коллекция, основанная на классификаторе ГРНТИ.

Серые стрелки, исходящие из экземпляров атрибутов, указывают на область возможных значений для них. Областью значений остальных атрибутов являются простые типы данных. На схеме значения атрибутов представлены с помощью объектов вспомогательного класса *значение атрибута* с префиксом «за». Объекты этого класса содержат для простых типов атрибутов их значения (например, значения текстовых атрибутов *фамилия, имя, название* представлены на схеме в кавычках).

Для объектного атрибута *автор* его значение содержит ссылку на соответствующий экземпляр информационного объекта с типом персона,

```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:lbm="http://libmeta.ru/">
  ▼<lbm:InformationResource rdf:about="http://libmeta.ru/resource/person">
     <lbm:title>Person</lbm:title>
     <lbm:label>Персона</lbm:label>
     <lbm:description>Ресурс соответствующий персонам</lbm:description>
    ▼<lbm:properties>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#activity"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#bio"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#dateOfBirth"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#dateOfDeath"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#first"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#keywords"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#last"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#middle"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#placeOfBirth"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#seeAlso"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#email"/>
     </lbm:properties>
     <lbm:dateCreated>10-05-2016 23:11</lbm:dateCreated>
     <lbm:dateUpdated>10-05-2016 23:11</lbm:dateUpdated>
  </lbm:InformationResource>
  ▼<lbm:InformationResource rdf:about="http://libmeta.ru/resource/publication">
     <lbm:title>Publication</lbm:title>
     <lbm:label>Публикация</lbm:label>
     <lbm:description>Ресурс соответствующий публикациям</lbm:description>
    ▼<lbm:properties>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#aboutTitle"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#annt"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#auth"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#desc"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#full"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#issueDate"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#issueMonth"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#keywords"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#lang"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#msc"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#media"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#originalTitle"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#riopubtype"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#pubtype"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#russianTitle"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#seeAlso"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#storage"/>
       <lbm:property rdf:resource="http://libmeta.ru/attribute#udc"/>
     </lbm:properties>
     <lbm:dateCreated>10-05-2016 23:11</lbm:dateCreated>
     <lbm:dateUpdated>10-05-2016 23:11</lbm:dateUpdated>
  </lbm:InformationResource>
</rdf:RDF>
```

Рис. 2 Описание ресурсов в формате RDF/XML

что отображено на схеме пунктирной стрелкой. Таксономические атрибуты *тип* и *язык* в качестве области значений указывают на соответствующие таксономии *тип публикации* и *язык*, представляющие собой линейные словари, элементы которых (таксоны) используются в качестве значений атрибутов.

Для каждого атрибута указан его вид: *описательный, идентификационный* или *поисковый*. Атрибут может относиться к нескольким видам одновременно. Поисковые атрибуты используются для динамической генерации формы поиска по объектам определенного типа ресурсов. Описательные атрибуты используются для генерации формы представления информации об объекте для пользователя.

Набор значений идентификационных атрибутов необходим, как понятно из названия, для идентификации объекта. В наборе атрибутов для публикаций атрибут *автор* помечен как *множественный*. Этот атрибут может иметь при описании информационных объектов, соответствующих по типу ресурса *публикациям*, несколько значений, что отражено в качестве примера на схеме.

Описание структуры контента в терминах LibMeta в формате RDF/XML[1] представлено на рис. 2. Задание структуры может осуществляться с помощью пользовательских интерфейсов системы или с помощью загрузки RDF/XML с описанием структуры контента в соответствующем разделе системы пользователем, наделенным соответствующим уровнем прав.

Основные понятия для описания контента библиотеки представлены в работе [5]. Фактически исходная онтология контента LibMeta содержит необходимые понятия, отношения и аксиомы. При описании конкретной предметной области в эту онтологию добавляются отдельные экземпляры определенных в ней понятий, которые и составляют контент создаваемой библиотеки.

[1] http://www.w3.org/RDF/; http://www.w3.org/XML/; http://www.w3.org/TR/rdf-syntax-grammar.

```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:lbm="http://libmeta.ru/">
  ▼<lbm:InformationObject rdf:about="http://libmeta.ru/io/Publication#rio#42025889">
     <lbm:type rdf:resource="http://libmeta.ru/resource/publication"/>
     <lbm:description/>
     <lbm:dateCreated>16-05-2016 22:04</lbm:dateCreated>
     <lbm:dateUpdated>16-05-2016 22:04</lbm:dateUpdated>
    ▼<lbm:properties>
      ▼<lbm:property>
         <lbm:type rdf:resource="http://libmeta.ru/attribute#annt"/>
        ▼<lbm:value>
            В работе предлагается методология вычисления вторых производных функции при условии соблюдения функциональной связи ее переменных.
            Исходной системе уравнений, описывающей связь переменных, ставятся в соответствие две сопряженные линейные системы уравнений,
            решения которых используются для нахождения гессиана функции, причем, основные матрицы этих двух вспомогательных систем являются
            сопряженными друг другу. Этот подход используется при вычислении вторых производных сложной функции в многошаговых процессах,
            получаемых в результате дискретной аппроксимации задачи оптимального управления. Выводятся формулы для вычисления гессиана целевой
            функции в таких задачах, получаемых в результате дискретной аппроксимации задачи оптимального управления по схеме Эйлера,
            модифицированной схеме Эйлера и методом Рунге-Кутта. Приводятся примеры численного решения таких задач методом Ньютона с помощью
            выведенных формул вторых производных.
          </lbm:value>
       </lbm:property>
      ▼<lbm:property>
         <lbm:type rdf:resource="http://libmeta.ru/attribute#auth"/>
         <lbm:value>Евтушенко Юрий Гаврилович</lbm:value>
       </lbm:property>
      ▼<lbm:property>
         <lbm:type rdf:resource="http://libmeta.ru/attribute#auth"/>
         <lbm:value>Зубов Владимир Иванович</lbm:value>
       </lbm:property>
      ▼<lbm:property>
         <lbm:type rdf:resource="http://libmeta.ru/attribute#auth"/>
         <lbm:value>Засухина Елена Семеновна</lbm:value>
       </lbm:property>
      ▼<lbm:property>
         <lbm:type rdf:resource="http://libmeta.ru/attribute#desc"/>
         <lbm:value>519.653:519.658</lbm:value>
       </lbm:property>
      ▼<lbm:property>
         <lbm:type rdf:resource="http://libmeta.ru/attribute#issueDate"/>
         <lbm:value>2004</lbm:value>
       </lbm:property>
      ▼<lbm:property>
         <lbm:type rdf:resource="http://libmeta.ru/attribute#issueMonth"/>
         <lbm:value>4</lbm:value>
       </lbm:property>
```

Рис. 3 Информационный объект в формате RDF/XML

На рис. 3 приведен пример представления экземпляра информационного объекта по заданному набору атрибутов из соответствующего ему экземпляра информационного ресурса.

4 LibMeta — второй пример конструирования

Рассмотрим пример, когда в качестве терминов предметной области используется узкоспециализированный тезаурус ОДУ [8]. Особенность этого тезауруса заключается в том, что он содержит не только сами понятия и термины, но и ссылки на публикации, в которых вводятся/определяются эти понятия, их математическая запись. Был введен новый информационный ресурс *литература* для описания публикаций, ставших основой построения этого тезауруса. В соответствие ему был поставлен тот же набор атрибутов, что и в предыдущем примере. На рис. 4 представлены понятия тезауруса, связанные иерархически, и для каждого понятия отображаются его горизонтальные связи.

С помощью этого тезауруса был размечен набор публикаций со схожей тематикой. Схожесть тематики публикации тезаурусу ОДУ определялась по ее ключевым словам, соответствующим терминам тезауруса.

На рис. 5 представлен пример связи понятия из ОДУ и найденных публикаций. В качестве связанных объектов могут выступать не только *публикации*, но и, например, персоны, в описании деятельности которых могут встречаться соответствующие понятию из ОДУ ключевые слова.

В качестве модели информационных ресурсов было использовано то же описание *персоны* и *публикации*, что и в предыдущем примере. В этом случае один и тот же набор атрибутов использовался как для описания ресурса *литература*, так и для описания ресурса *публикация*. Это позволило отдельно настроить права доступа для всех объектов *литературы*, запретив их модификацию или удаление пользователям, не являющимся редакторами предметной области. Сами публикации извлечены из Единого научного информационного пространства (ЕНИП) РАН — это интегрированное информационное пространство

Рис. 4 Тезаурус ОДУ

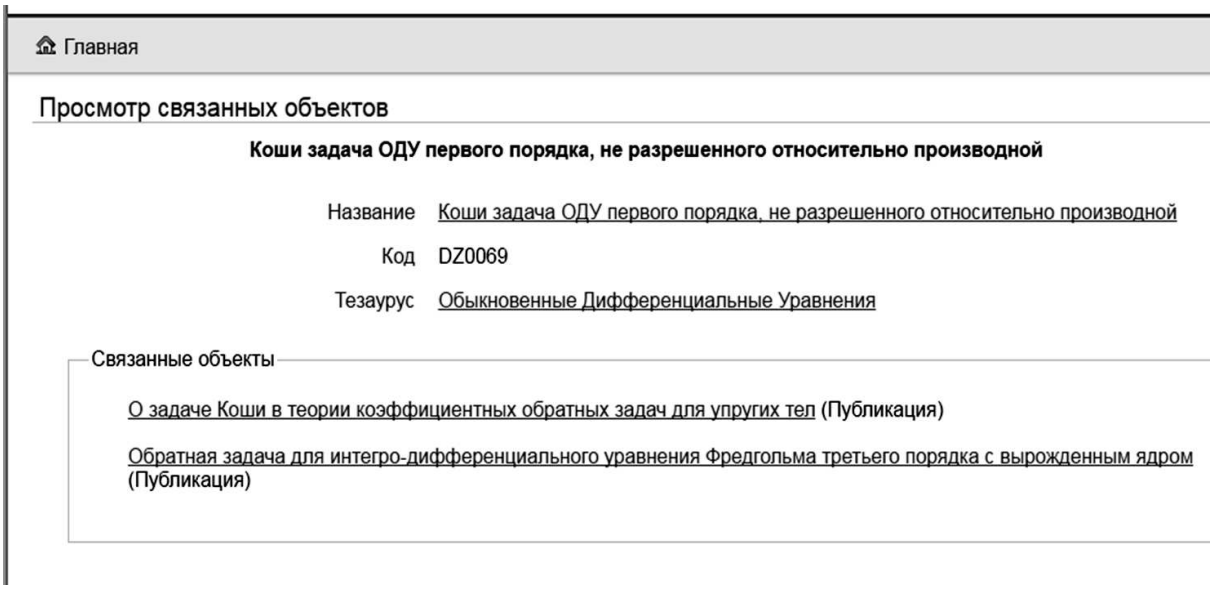

Рис. 5 Информационные объекты и термины ОДУ

распределенных и локальных цифровых (электронных) ресурсов организаций РАН и комплекс программно-технических средств, обеспечивающих использование этих ресурсов и полнофункциональное управление ими [9].

Для извлечения информации о публикациях и авторах использовался протокол OAI-PMH[1]. Данные были представлены в формате Dublin Core[2]. Часть публикаций была размечена ключевыми словами, однако термины не разделены между собой и просто перечислялись через запятую в одном поле. Ключевые слова публикации были преобразованы в набор ключевых слов соответствующего информационного объекта для каждой извлеченной публикации из ЕНИП.

В коллекцию публикаций тезауруса ОДУ добавлялись те объекты, в наборе ключевых слов которых находились термины ОДУ.

[1] https://www.openarchives.org/pmh.
[2] http://dublincore.org.

Размещение публикации в ту или иную ветвь коллекции может осуществлять как сам пользователь, так и соответствующий модуль автоматической разметки информационных объектов по тезаурусу, в котором задаются простые правила разметки в рамках описания тезауруса.

5 LibMeta — взаимодействие с источниками LOD

Основная проблема приложений, разрабатываемых для работы с данными из источников, интегрированных в LOD, состоит в том, что данные в этих источниках очень слабо провязаны с данными других источников. Большинство имеющихся связей расположены на уровне самих данных, при этом на уровне схем такие связи практически отсутствуют. Для решения этой проблемы предлагаются разные подходы, которые основаны на методах сравнения онтологий [1, 4]. Некоторые из них используют для сравнения онтологий данные, доступные в сети. В частности, используется Wikipedia и ее иерархический рубрикатор[1].

В отличие от других работ, построение иерархий классов при трансляции ресурсов на источник данных в данной ситуации не представляет интереса, поэтому для проставления связей используется связь, которая указывает, что два разных класса *могут* иметь одинаковых представителей. Эта связь может указывать на класс в источнике данных LOD, который является источником дополнительной информации о ресурсе-субъекте, или на эквивалентный ему класс, возможно, с разной степенью детализации описания объектов. Фактически предполагается, что онтология источника данных *частично* совместима со структурой ресурсов описываемой библиотеки. Это означает, что хотя бы один ресурс ее онтологии может быть транслирован в некоторый класс в онтологии источника данных. Требуется лишь минимальное частичное соответствие ресурсу LibMeta. Это означает, что для однозначной идентификации экземпляров соответствующего класса из источника данных отображаться должны как минимум идентифицирующие атрибуты.

В связи с гибкостью схемы LibMeta предполагается возможным сценарий создания дополнительных типов ресурсов для подключаемых источников, информацию из которых можно использовать как значения некоторых атрибутов основных ресурсов.

Для трансляции атрибутов ресурса в свойства выбранного класса источника данных будет использоваться связь, которая указывает, что значения атрибута и свойства полностью или частично совпадают в рамках установленного соответствия на уровне ресурса библиотеки и класса онтологии. При совпадении значений все очевидно, проблема возникает при разной детализации данных, когда возможно отображение значения свойства класса в несколько атрибутов и наоборот. В связи с гибкостью модели данных может быть принято решение о расширении схемы ресурса. В другом случае пользователь может использовать набор вспомогательных функций, например для расщепления или слияния данных. Простой пример такого рода преобразований связан с трансформацией имени персоны. В первом случае, когда имя персоны описывается одним значением в источнике данных, а транслируется в три отдельных атрибута, используется функция расщепления (*ФИО* → *Ф, И, О*). Во втором случае значения отдельных свойств преобразуются в значения одного атрибута (*Ф, И, О* → *ФИО*), отображение свойств класса производится в один атрибут и тогда данные будут склеиваться как одно значение этого атрибута именно в том порядке, в котором они были перечислены при описании трансляции.

Для преобразования данных в соответствующие типы значений, которые указаны при описании атрибута, использованы встроенные функции преобразования, которые нет нужды настраивать пользователю. В случае если такое преобразование заканчивается неудачно, то информация об этом будет сохранена в соответствующем административном атрибуте объекта для возможности дальнейшей обработки и исправления ошибок.

Поиск эквивалентных классов в источниках данных пользователь может выполнить:

(1) вручную, выбирая из списка доступных классов в указанном источнике данных;

(2) полуавтоматически, используя имеющиеся описания связей с другими классами внешних онтологий, заранее определенными при описании структуры ресурсов;

(3) автоматически.

В первых двух случаях пользователь на первом шаге предварительно указывает, с каким типом ресурсов он предполагает работать, привязывая тот или иной источник данных. На втором шаге он определяет соответствие атрибутов и свойств. В третьем варианте он получает возможность получить общую оценку соответствия схемы ресурсов библиотеки некоторой онтологии и на основе этой оценки принимать решение о трансляции ресурсов

[1] https://en.wikipedia.org/wiki/Special:Categories.

на тот или иной источник данных, использующий эту онтологию.

Назовем проекцию понятия библиотеки IR на понятие C источника данных из LOD *допустимой*, если возможно установить между ними хотя бы одно отношение из $\{R_1, R_2, R_3, R_4\}$:

- $R_1(C, \mathrm{IR})$ означает, что понятие C включает в себя IR;
- $R_2(C, \mathrm{IR})$ означает, что понятие IR включает в себя C;
- $R_3(C, \mathrm{IR})$ означает, что понятие IR связано отношением эквивалентности с C;
- $R_4(C, \mathrm{IR})$ означает, что понятие IR связано отношением частичной эквивалентности с C.

Все эти четыре отношения говорят о том, что понятия IR и C могут иметь одинаковых представителей. По семантике R_1 соответствует skos:broader (например, IR = *Человек*, C = *Студент*); R_2 соответствует skos:narrower (например, IR = *Студент*, C = *Человек*); R_3 соответствует skos:exactMatch (например, IR = *Студент*, C = *Студент*); R_4 соответствует skos:closeMatch (например, IR = *Студент*, C = *Учащийся*).

Отображение IR на C *возможно*, если трансляция набора атрибутов IR на свойства класса C выполнена хотя бы для идентифицирующих атрибутов, т. е. любой идентифицирующий атрибут из $a_1, \ldots, a_k$, принадлежащих набору атрибутов IR, где $k < n$, n — число атрибутов соответствующего набора, транслируется хотя бы на одно свойство $c_1, \ldots, c_m$ класса C.

Трансляция атрибута на свойство источника $t(a_i, c_j)$ для искомых или связываемых объектов может быть:

- *прямой*, когда атрибут отображается на свойство: $t_1 = \{t(a_i, c_j) : \ a_i = c_j\}$, т. е. значения должны быть эквивалентны;
- *неполной*, когда атрибут отображается на свойство лишь частично: $t_2 = \{t(a_i, c_j) : \ a_i \subset c_j\}$, т. е. значение атрибута включается в значение свойства;
- *избыточной*, когда атрибут шире свойства: $t_3 = \{t(a_i, c_j) : a_i \supset c_j\}$, т. е. значение атрибута содержит больше информации, чем значение свойства.

Из определения проекции некоторого понятия IR и трансляции его атрибутов, задающих отображение понятия на некоторый набор данных источника, следует, что это отображение сюръективно и неинъективно для наборов его атрибутов.

При использовании трансляции для поиска связанных объектов в случае *полной* трансляции идентифицирующего атрибута все ограничивается выбором соответствующего свойства и значения могут сравниваться. В случае *неполной* и *избыточной* трансляции идентифицирующего атрибута явного сравнения значений недостаточно. В любом случае возможно использование функций предобработки данных: преобразование форматов, извлечение подстрок и т. д.

В общем виде извлечение объектов из источника для сохранения в качестве информационных объектов библиотеки задается функцией

$$z\left(R_l \cap f(t)\right) = \\ = \{o \in \mathrm{IR} | R_l(\mathrm{IR}, C) : \ \forall\, a_i \exists\, f(t(a_i, c_j))\} \ ,$$

где функция f зависит от типа трансляции атрибутов:

$$f(t_1) = \begin{cases} \text{true}, & a_i = c_j \, ; \\ \text{false}, & a_i \neq c_j \, ; \end{cases}$$

$$f(t_2) = \begin{cases} \text{true}, & a_i \subset c_j \, ; \\ \text{false}, & a_i \not\subset c_j \, ; \end{cases}$$

$$f(t_3) = \begin{cases} \text{true}, & c_j \subset a_i \, ; \\ \text{false}, & c_j \not\subset a_i \, . \end{cases}$$

Для того чтобы выполнять поисковые запросы по источникам данных, необходимо, чтобы отображение понятий библиотеки на понятия источника было и *возможным*, и *допустимым*.

Если оно возможно, то для $R_1(C, \mathrm{IR})$ это означает, что все характерные признаки IR наследуются C, при этом набор признаков C шире набора IR, так как IR является более объемлющим понятием и, следовательно, класс C всегда включает признаки, которые являются идентифицирующими для более широкого понятия IR. Если оно возможно для $R_2(C, \mathrm{IR})$, это означает, что все признаки класса C наследуются IR, при этом набор признаков IR может быть шире набора класса C, так как IR является более узким понятием. Набор идентифицирующих признаков класса C, необходимых для идентификации объектов в источнике данных, включается в набор признаков IR, что достаточно для идентификации эквивалентных объектов в источнике. Если набор необходимых идентифицирующих признаков IR шире, то это означает, что в контексте источника этот набор избыточен и можно его переопределить (сузить) для построения допустимой трансляции. Если оно возможно для $R_3(C, \mathrm{IR})$, это означает, что все характерные признаки совпадают, в том числе и идентифицирующие. Если трансляция возможна, то для понятий, связанных отношением $R_4(C, \mathrm{IR})$, это означает, что

хотя бы идентифицирующие характеристики у них совпадают.

Исходя из сказанного, для любых понятий, связанных одним из этих отношений, возможно построение допустимой трансляции и можно строить поисковые запросы, результатом которых являются интерпретируемые в терминах ресурсов библиотеки объекты.

Если же отображение недопустимо, то, значит, для всех четырех вариантов нарушается отображение идентифицирующих атрибутов, т. е. невозможно извлечь интерпретируемые данные (пример: достанем всех персон по имени Лев, но не сможем понять, кто из них Толстой, если не отобразим идентифицирующие атрибуты, которые определены в библиотеке, например, как Ф, И, О, ДР).

Если отображение невозможно, но допустимо, т. е. некоторый набор атрибутов можно отобразить на некоторые свойства, то, так как один набор атрибутов может соответствовать нескольким понятиям/типам ресурсов в библиотеке, невозможно будет идентифицировать тип ресурса извлекаемого объекта (например, если имеются понятия *Мужчина* и *Женщина*, набор атрибутов которых одинаков, то, извлекая персону с именем «*Джойс Кэрол Оутс*» с датой рождения «16.06.1938», нельзя определить ее принадлежность к понятию). Поэтому при построении отображения надо последовательно проходить этап построения возможных проекций ресурсов, а затем этап построения допустимых трансляций атрибутов этих ресурсов.

Авторы не претендуют на идеальную модель отображения в любой источник, но, по крайней мере, имея адаптивную модель данных, всегда можно выполнить настройку таким образом, чтобы иметь возможность извлечь интересующие данные для определенного круга задач.

Введя функцию интерпретации I_z, которая по построенным отображениям T_i для источников данных D_i и информационного ресурса IR сопоставляет *информационным объектам*, соответствующим этому ресурсу, объекты источников данных, можно построить множество связей этих объектов. Функция I_z называется моделью связанных данных источников данных и LibMeta.

Если не удается построить функцию интерпретации некоторого источника для хотя бы одного ресурса, то надо либо расширять модель LibMeta, либо источник отбрасывается как источник с данными, не интерпретируемыми в данной библиотеке. Таким образом, можно выявить скрытые связи между источниками данных как на уровне схем, так и на уровне данных через ресурсы библиотеки.

6 LibMeta — пример использования внешних онтологий при описании ресурсов и подключении к источникам

Существуют два пути настройки системы под конкретные понятия предметной области: (1) воспользоваться формами создания и редактирования ресурсов, что удобно, когда экземпляров ресурсов мало; (2) воспользоваться имеющимися онтологиями, описывающими понятия предметной области. Если в первом случае все достаточно тривиально, то на втором случае следует остановиться подробней.

В системе предусмотрена подсистема загрузки онтологии, представленной на языке описания онтологий OWL, для автоматического создания экземпляров информационных ресурсов. При загрузке онтологии можно указать, какие классы онтологии извлекаются из нее, при этом они становятся экземплярами класса *информационный ресурс* в LibMeta, а их свойства преобразуются в экземпляры атрибутов и включаются в один набор атрибутов для конкретного ресурса. Естественно, LibMeta не поддерживает всю семантику отношений и ограничений, накладываемых на свойства и классы в OWL, но отображает основные свойства и ограничения на свою схему. Этого достаточно для выполнения задач интеграции и публикации данных в облаке LOD. Так, в наборе данных, полученном из «Научного наследия России» с помощью подсистемы харвестинга по протоколу OAI-PMH, вся исходная информация о том, где находится исходный объект со всем своим описанием, сохраняется. С помощью LibMeta выполняется только провязывание данных с данными из LOD и пользователям предоставляется возможность организовывать их в собственные коллекции, возможно, дополняя описания данных на свое усмотрение, добавляя новые теги или определяя более точно тематическую направленность, используя в качестве базовых расширяемых понятий элементы из ГРНТИ. Перечислим правила отображения внешней онтологии в описание контента библиотеки:

- классы онтологии становятся экземплярами класса *информационный ресурс*;
- свойства класса онтологии становятся экземплярами класса *атрибут*;
- все свойства, относящиеся к одному классу, группируются в *наборы атрибутов*;

– для простых свойств в качестве области значений атрибута указывается соответствующий тип (строка, дата и т. д.);
– по умолчанию все атрибуты относятся к виду *описательный* и *поисковый*, соответствующие настройки можно изменить в дальнейшем через интерфейсы системы;
– для сложных свойств, областью значений которых являются экземпляры некоторого класса, выбирается соответствующий ресурс; если ресурс не был загружен в систему, то дальнейшие решения принимает пользователь, отвечающий за создание структуры контента библиотеки;
– все однозначные свойства становятся однозначными атрибутами, для атрибутов многозначных свойств ставится пометка о возможности подключения нескольких значений;
– иерархические связи между классами не сохраняются в явном виде, но для ресурсов создается атрибут с соответствующим назнанием «*вышестоящий объект*» и «*нижестоящий объект*», который позволяет устанавливать подобные связи на уровне объектов (практически идентичны по смыслу связям из онтологии SKOS[1] skos:broader, skos:narrower).

Инверсивность, транзитивность и симметричность свойств не отображаются явно в описании ресурсов LibMeta, но для каждого информационного объекта можно всегда получить список ссылающихся на него объектов и информацию о том, посредством какого атрибута это делается. Фактически онтологическое описание классов сводится к набору *поисковых* и *описательных* атрибутов выделением среди них *идентифицирующих*, используемых, например, в задачах выявления дубликатов.

Важно отметить, что все идентификаторы классов и свойств онтологии сохраняются в описании соответствующих экземпляров ресурсов и атрибутов с помощью использования связи, определяющей их эквивалентность. Это позволяет в дальнейшем при настройке отображения ресурса на некоторый источник в LOD, который в своей схеме использует эти классы и свойства, выполнять эту процедуру почти полностью автоматически. При конструировании структуры контента через интерфейсы также имеется возможность для каждого ресурса и его атрибута указать соответствующие им URI[2] из общеиспользуемых онтологий.

Рассмотрим в качестве примера онтологии, которые широко распространены для описания основных типов ресурсов рассматриваемых понятий *персона* и *публикация* в сообществе LOD, и оценим их описания на соответствие имеющимся в наличии метаданным. Чаще всего эти онтологии содержат десятки классов и свойств и являются избыточными для описания нужных объектов, выделяя подмножество необходимых классов и свойств, используя для отображения лишь малую часть их свойств, необходимых для подключения к источникам данных, которые они охватывают.

AKT

Онтология AKT Reference Ontology, или кратко AKT[3] (доступна по адресу http://swl.slis.indiana.edu/repository/owl/aktportal.owl), разработана в целях унификации доступа к библиографической информации в 2003 г. И хотя проект был закрыт, данные AKT на сегодняшний момент представлены более чем в 200 источниках, таких как DBLP, Citeseer, CORDIS, NSF, EPSRC, ACM, IEEE и др.

Объединяет несколько онтологий; из них интерес представляет основная онтология Portal Ontology, которая содержит понятия для описания *персон* и *публикаций*. Данные разнородны и опираются на очень узкие подмножества этой онтологии. Многие поля, имеющиеся в этой богатой онтологии, остаются незаполненными при описании реальных данных.

Dublin Core

Исторически Dublin Core представляет собой набор понятий, используемых для описания разнообразных типов ресурсов, из которых 15 являются обязательными для описания. Практически можно описать метаданные о *персонах* и *публикациях* из рассматриваемого примера в терминах этих понятий. Элементы Dublin Core часто повторно используются, дополняются и конкретизируются в других онтологиях. Охватывает огромное число источников, включая DBpedia.

FOAF

Онтология FOAF[4] (Friend-of-a-Friend) уже является практически стандартом для описания людей и их отношений с другими ресурсами. Используется в разнообразных контекстах и может использоваться для описания в любых сценариях с участием персон. Часто также включается и конкретизируется в других онтологиях.

[1] https://www.w3.org/2004/02/skos.
[2] https://tools.ietf.org/html/rfc3986.
[3] http://projects.kmi.open.ac.uk/akt/ref-onto.
[4] http://xmlns.com/foaf/spec.

BIBO

Онтология BIBO[1] (Bibliographic Ontology) предназначена для описания библиографических данных, включает в себя понятия из других онтологий, таких как Dublin Core и FOAF, расширяя и конкретизируя их понятия, которые используются при описании ее классов. Содержит 38 видов документов, включает понятия, необходимые для описания *персон* и *публикаций*. Можно представить описание собственных ресурсов в терминах этой онтологии, ограничившись лишь частью ее терминов. Охватывает такие источники, как Британская национальная библиотека[2], DBpedia и т. д.

DBpedia

Онтология Dbpedia, разработанная в рамках проекта Dbpedia, содержит большое число классов для описания самых разнообразных объектов, включая понятия *публикация* и *персона*. Она также включает в себя понятия из других онтологий, которые используются при описании ее классов. DBpedia является центральным узлом LOD и связывает информацию из самых разных источников, которые ссылаются на нее.

Таблица 1 показывает отображение информационного ресурса «Публикация» в термины источников данных, схема данных которых опирается на перечисленные онтологии. Если для указанных онтологий нет соответствующего класса, то название класса не указывается. Это всего лишь означает, что элементы этой онтологии могут использоваться в другой онтологии, где они конкретизируются в рамках используемого класса. Если не указывается свойство, значит, в терминах этой онтологии нет такого свойства или близкого ему по смыслу. В случае с BIBO один из перечисленных классов определяет тип публикации. Поиск публикаций в DBpedia представляется бессмысленным в рассматриваемых примерах, поэтому в табл. 1 информация из этой онтологии не включалась

В табл. 2 представлено отображение информационного ресурса «Персона» в термины источников данных, схема данных которых опирается на перечисленные онтологии. Например, для персон из DBpedia представлены элементы из собственного пространства имен, но DBpedia также включает и FOAF-онтологию, поэтому можно было отобразить значения и на пространство имен FOAF в рамках источника данных DBpedia.

Эта информация об отображении атрибутов LibMeta на свойства других онтологий также может быть включена в описание каждого атрибута с помощью использования соответствующей связи. Эта информация позволит быстро подключаться к нужным источникам данных и формировать описания объектов в терминах нужной онтологии. В рассматриваемом примере были подключены два источника данных — это Dbpedia и данные о персонах из системы MathNet[3], выгруженные предварительно в отдельное хранилище в виде RDF-троек в формате Dublin Core.

Таблица 1 Элементы описания ресурса «Публикация»

Libmeta	AKT	Dublin Core	FOAF	BIBO
	Класс Akt:Publication-Reference	Класс Dc:bibliographicresource	Класс	Класс Bibo:Article, bibo:academicarticle, bibo:Proceedings
Название	akt:has-title	dc:title	foaf:title	dc:title
Аннотация	akt:has-abstract	dc:description		bibo:abstract
Дополнительное заглавие				bibo:shortTitle
Тип	akt:article-of-journal	dc:type		
Язык		dc:contributor		dc:contributor
Автор	akt:has-author	dc:language		dc:language
Исходная страница	akt:has-web-address	dc:source	foaf:homepage	foaf:homepage
Описание	akt:addresses-generic-area-of-interest			bibo:shortDescription

[1] http://bibliontology.com.
[2] http://www.bl.uk.
[3] http://www.mathnet.ru.

Таблица 2 Элементы описания ресурса «Персона»

LibMeta	AKT	Dublin Core	FOAF	BIBO	DBpedia
	Класс	Класс Dc:agent	Класс Foaf:agent, foaf:person	Класс Foaf:agent, foaf:person	Класс Dbo:person
Фамилия	akt:full-name, akt:family-name	dc:title	foaf:name, foaf:lastname, foaf:family_name	foaf:family_name	dbo:birthName
Имя	akt:full-name, akt:given-name	dc:title	foaf:name, foaf:given_name	foaf:given_name	dbo:birthName
Отчество	akt:full-name	dc:title	foaf:name, foaf:surname		dbo:birthName
Дата рождения		dc:date	foaf:birthData		dbo:birthDate
Место рождения					dbo:birthPlace
Биография		dc:description			dbo:abstract
Деятельность		dc:description			dbo:occupation

7 LibMeta — роль пользователей в системе

Несомненно, самыми главными действующими лицами в любой библиотеке являются ее пользователи, и LibMeta не исключение. Пользователи LibMeta делятся на несколько категорий в соответствии со своими ролями:

(1) администраторы контента библиотеки;

(2) редакторы предметной области;

(3) администраторы источников данных;

(4) редакторы информационных объектов;

(5) простые пользователи.

Администраторы контента библиотеки отвечают за создание информационных ресурсов, атрибутов и их наборов и получают доступ ко всей функциональности системы. Редактор предметной области имеет право на редактирование тезауруса и определение основных коллекций системы. Администраторы источников данных отвечают за их подключение и настройку отображения. Редактор информационных объектов может создавать, редактировать и удалять любой информационный объект. В отличие от всех остальных ролей, для редактора информационных объектов роль определяет доступ к конкретным функциям подсистемы работы с объектами (редактирование, создание, и удаление), а не просто доступ к функциональности для сопровождения отдельных объектов. Зарегистрированные простые пользователи могут описывать свою область интересов на основе терминов тезауруса предметной области, очерчивая тем самым интересующий их круг информационных объектов. Эти пользователи также могут добавлять в тезаурус уточняющие область их интересов термины, которые не отображаются для остальных пользователей, добавлять собственные объекты в свою коллекцию в рамках своей области интересов, сохранять свои результаты поиска по источникам данных. Задание области интересов пользователя позволяет группировать пользователей со сходными интересами, строить связи между ними, анализируя круг используемых терминов, добавленных объектов, и давать рекомендации на основе этих связей. Таким образом, можно отслеживать и выделять взаимосвязи между разными областями интересов. Понятно, что один и тот же пользователь может обладать несколькими ролями и выступать, например, в качестве редактора предметной области и администратора источников данных. Но при этом любой пользователь системы независимо от роли получает доступ к функциям поиска и навигации по объектам системы.

На рис. 6 отображены роли пользователей и в виде четырехугольников очерчивается их область влияния для рассмотренного выше примера сконструированного контента.

8 Заключение

Разрабатывая модель информационной системы LibMeta, авторы хотели получить гибкую систему интеграции различных типов ресурсов с возможностью интеграции с внешними системами. Одним из определяющих условий было отсутствие требования специальной подготовки у простых пользователей системы. Основные идеи при разработке системы были позаимствованы из концепции адаптивных моделей данных, разработанной

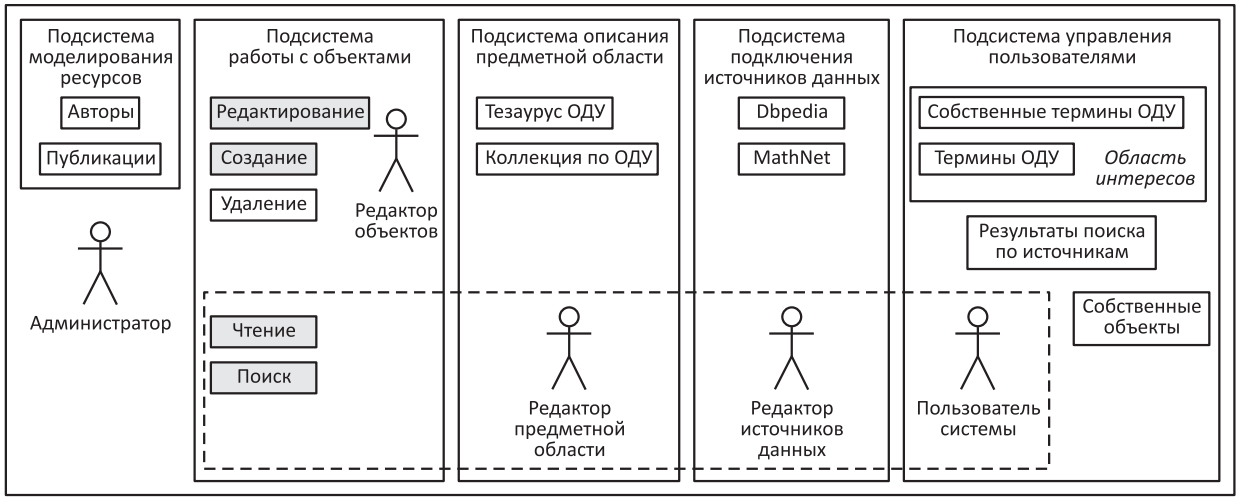

Рис. 6 Пользователи

в 1990-х гг. [10, 11]. Эта модель данных подходит для определенного круга задач, для решения которых нужно разрабатывать довольно сложные частные модели. К таким задачам относится ставшая классической задача интеграции данных из источников с разными моделями. Внесение изменений в интеграционную модель чаще всего выполняется на программном уровне, требуя значительных усилий от разработчиков для внедрения изменений. Применение адаптивной модели позволяет понизить сложность как самой модели данных, так и разрабатываемых на их основе систем, в которых попутно решается задача создания динамических (адаптивных) пользовательских интерфейсов.

Применение этой модели данных делает возможной динамическую трансформацию и интерпретацию модели данных в приложении, решающем задачу интеграции данных, позволяя настраивать используемые решения под определенную предметную область. В таких задачах часто меняются требования к модели данных, меняется детализация схемы, детализируя или, наоборот, обобщая ее описание. Реализация приложений для конечных пользователей под эти требования занимает больше времени, чем хотелось бы. Поиск решений для такого рода задач на более высоком уровне абстракции привел к появлению концепции адаптивной модели данных, в которой декларируется лишь стиль моделирования данных для приложений. Получаемые модели более абстрактны, состоят из меньшего числа понятий с более простыми связями и не привязаны к определенным предметным областям. Поведенческие моменты определяются операциями создать, сохранить и т. д., ролями пользователя, которые отражают условия доступа к этим операциям. Пользовательский интерфейс представляется адаптивным в соответствии с моделью данных и использует для своего воспроизведения/построения заранее определенные паттерны уровня представления. В результате применения этой модели, когда меняется ее структура, система немедленно подстраивается под эти изменения.

В статье был приведен пример конструирования контента семантической библиотеки для авторов и их публикаций в рамках библиотеки LibMeta на основе данных из системы «Научное наследие России». На имеющемся наборе данных, представленном примерно 7000 публикациями и их авторами (около 1300 персон), было проведено тестовое исследование. Из 1300 авторов оказалась представлена в DBPedia примерно треть, и около половины из них были представлены в VIAF[1] (Virtual International Authority File). Часть ссылок на данные VIAF была получена из набора данных DBPedia, другая часть была непосредственно извлечена из самого VIAF, подключенного как источник данных.

Во втором примере, сконструированном для тезауруса ОДУ из 10 000 публикаций базы ЕНИП, было загружено 100 подходящих по тематике. Загруженные публикации были размечены ключевыми словами тезауруса, всего было получено 789 элементов разметки.

Провести связывание с данными из источников, охваченных онтологиями AKT, оказалось невозможным из-за специфики данных, но эти источники были подключены к LibMeta в качестве тестовых для использования их в качестве источников для поиска. В статью не вошли задачи поиска по

[1] http://viaf.org.

семантическим тегам или наборам ключевых слов не только по строго заданным правилам отображения ресурсов, а также поиск по формулам в математических данных. Предполагается осветить это направление работ в рамках предложенной системы в дальнейших работах.

Литература

1. *Gruber T. R.* A translation approach to portable ontologies // Knowl. Acquis., 1993. Vol. 5. No. 2. P. 199–220.
2. Semantic digital libraries / Eds. S. R. Kruk, B. McDaniel. — Berlin–Heidelberg: Springer-Verlag, 2009. 245 p.
3. *Антопольский А. Б., Каленкова А. А., Каленов Н. Е., Серебряков В. А., Сотников А. Н.* Принципы разработки интегрированной системы для научных библиотек, архивов и музеев // Информационные ресурсы России, 2012. № 1. С. 2–6.
4. *Bizer C., Heath T., Berners-Lee T.* Linked data — the story so far // Int. J. Semant. Web Inf. Syst., 2009. Vol. 5. No. 3. P. 1–22.
5. *Серебряков В. А., Атаева О. М.* Основные понятия формальной модели семантических библиотек и формализация процессов интеграции в ней // Программные продукты и системы, 2015. № 4. С. 180–187.
6. *Серебряков В. А., Атаева О. М.* Персональная цифровая библиотека LibMeta как среда интеграции связанных открытых данных // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XVI Всеросс. научной конф. RCDL'2014. — Дубна: ОИЯИ, 2014. С. 66–71.
7. *Каленов Н. Е., Савин Г. И., Сотников А. Н.* Электронная библиотека «Научное наследие России» // Информационные ресурсы России, 2009. № 2(108). С. 19–20.
8. *Моисеев Е. И., Муромский А. А., Тучкова Н. П.* Тезаурус информационно-поисковый по предметной области «обыкновенные дифференциальные уравнения». — М.: МАКС Пресс, 2005. 116 с.
9. *Бездушный А. Н., Бездушный А. А., Серебряков В. А., Филиппов В. И.* Интеграция метаданных Единого Научного Информационного Пространства РАН. — М.: ВЦ РАН, 2006. 238 с.
10. *Yoder J. W., Balaguer F., Johnson R.* Architecture and design of adaptive object-model // Adaptive Object Model, 2000. 11 p. http://www.adaptiveobjectmodel.com/OOPSLA2001/AOMIntriguingTechPaper.pdf.
11. *Welick L., Yode J. W., Wirfs-Broc R.* Adaptive object-model builder // Adaptive Object Model, 2009. 8 p. http://joeyoder.com/PDFs/04welicki.pdf.

Поступила в редакцию 01.12.16

PERSONAL SEMANTIC OPEN DIGITAL LIBRARY LibMeta. CONSTRUCTION OF THE CONTENT. INTEGRATION WITH LOD SOURCES

O. M. Ataeva and V. A. Serebryakov

A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Semantic technologies development has brought digital libraries to the level where a meaningful representation of the content of digital libraries came to the forefront. At the same time, it is necessary to limit it in terms of a certain subject area. The paper describes the libraries content construction with a thesaurus supporting the domain terminology within the developed system LibMeta. The personal semantic open digital library LibMeta provides the functionality of the construction of the library content in accordance with the specific requirements. LibMeta supports users working with resources of digital libraries and their collections in a certain subject area. One needs just to make the initial setup of the system for a specific subject area. For the description of a subject area, the system uses its limited terminology collected in a thesaurus. The domain used as an example is a highly specialized thesaurus of ordinary differential equations.

Keywords: semantic library; data model; ontology; data sources; search in LOD

DOI: 10.14357/19922264170210

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (project 14-07-00058 A).

References

1. Gruber, T. R. 1993. A translation approach to portable ontologies. *Knowl. Acquis.* 5(2):199–220.
2. Kruk, S. R., and B. McDaniel, eds. 2009. *Semantic digital libraries*. Berlin–Heidelberg: Springer-Verlag. 245 p.
3. Antopolsky, A. B., A. A. Kalenkova, N. E. Kalenov, V. A. Serebryakov, and A. Sotnikov. 2012. Printsipy razrabotki integrirovannoy sistemy dlya nauchnykh bibliotek, arkhivov i muzeev [Principles for the development of an integrated system for academic libraries, archives and museums]. *Informatsionnye resursy Rossii* [Information Resources of Russia] 1:2–6.
4. Bizer, C., T. Heath, and T. Berners-Lee. 2009. Linked data — the story so far. *Int. J. Semant. Web Inf. Syst.* 5(3):1–2.
5. Serebryakov, V. A., and O. M. Ataeva. 2015. Osnovnye ponyatiya dlya postroeniya formal'noy modeli semanticheskikh bibliotek i opisaniya protsessov integratsii v ney [The basic concepts for building a formal model of semantic libraries and description of the integration processes in it]. *Programmnye produkty i sistemy* [Software and Systems] 4:180–187.
6. Serebryakov, V. A., and O. M. Ataeva. 2014. Personal'naya tsifrovaya biblioteka LibMeta kak sreda integratsii svyazannykh otkrytykh dannykh [Personal digital library LibMeta as an integration environment of linked data]. *RCDL Proceedings*. 66–71.
7. Kalyonov, N. E., G. I. Savin, and A. N. Sotnikov. 2009. Elektronnaya biblioteka "Nauchnoe nasledie Rossii" [Electronic library "The scientific heritage of Russia"]. *Informacionnye resursy Rossii* [Information Resources of Russia] 2:19–20.
8. Moiseev, E. I., A. A. Muromskij, and N. P. Tuchkova. 2005. *Tezaurus informatsionno-poiskovyy po predmetnoy oblasti "obyknovennye differentsial'nye uravneniya"* [Information search thesaurus of subject area "ordinary differential equations"]. Moscow: MAKS Press. 116 p.
9. Bezdushnyj, A. N., A. A. Bezdushnyj, V. A. Serebryakov, and V. I. Filippov. 2006. *Integratsiya metadannykh Edinogo Nauchnogo Informatsionnogo Prostranstva RAN* [The integration of metadata for common scientific information space of RAS]. — Moscow: Computing Centre of the Russian Academy of Sciences. 238 p.
10. Yoder, J. W., F. Balaguer, and R. Johnson. 2000. Architecture and design of adaptive object-model. *Adaptive Object Model*. Available at: http://www.adaptiveobjectmodel.com/OOPSLA2001/AOMIntriguingTechPaper.pdf (accessed April 17, 2017).
11. Welick, L., J. W. Yode, and R. Wirfs-Broc. 2009. Adaptive object-model builder. *Adaptive Object Model*. Available at: http://joeyoder.com/PDFs/04welicki.pdf (accessed April 17, 2017).

Received December 1, 2016

Contributors

Ataeva Olga M. (b. 1978) — junior scientist, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; oli@ultimeta.ru

Serebryakov Vladimir A. (b. 1946) — Doctor of Science in physics and mathematics, professor, Head of Department, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; serebr@ultimeta.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 101–111

МОДИФИЦИРОВАННЫЕ ЭЛЛИПСОИДАЛЬНЫЕ УСЛОВНО-ОПТИМАЛЬНЫЕ ФИЛЬТРЫ ДЛЯ НЕЛИНЕЙНЫХ СТОХАСТИЧЕСКИХ СИСТЕМ НА МНОГООБРАЗИЯХ*

И. Н. Синицын[1], В. И. Синицын[2], Э. Р. Корепанов[3]

Аннотация: Представлена теория аналитического синтеза по критерию минимума средней квадратической ошибки (с.к.) модифицированных эллипсоидальных условно-оптимальных фильтров (МЭУОФ) для нелинейных дифференциальных стохастических систем на гладких многообразиях (МСтС) на основе эллипсоидальной аппроксимации ненормированных апостериорных распределений. Рассмотрены случаи гауссовских и негауссовских МСтС. Алгоритмы МЭУОФ по сравнению с алгоритмами ЭУОФ обладают достаточной простотой. Алгоритмы МЭУОФ положены в основу модуля инструментального программного обеспечения StS-Filter (version 2017).

Ключевые слова: винеровский шум; метод эллипсоидальной аппроксимации (МЭА); метод эллипсоидальной линеаризации (МЭЛ); модифицированный эллипсоидальный СОФ (МЭСОФ); ненормированная одномерная апостериорная характеристическая функция; пуассоновский шум; субоптимальный фильтр (СОФ); условно-оптимальный фильтр (УОФ); уравнения точности и чувствительности; эллипсоидальный УОФ (ЭУОФ)

DOI: 10.14357/19922264170211

1 Введение

В [1, 2] представлена теория УОФ на базе методов нормальной аппроксимации (МНА), статистической линеаризации (МСЛ) и ортогональных разложений (МОР) для МСтС с винеровскими шумами в уравнениях наблюдения и винеровскими и пуассоновскими шумами в уравнениях состояния. В основу теории УОФ были положены точные нелинейные уравнения для апостериорного одномерного распределения.

В [3] рассмотрено развитие [1, 2] на случай, когда апостериорное одномерное распределение ошибки фильтрации допускает эллипсоидальную аппроксимацию [4]. Получены точные фильтрационные уравнения, а также уравнения точности и чувствительности на основе МОР, даны элементы эллипсоидального анализа распределений, выведены уравнения ЭУОФ по методам эллипсоидальной аппроксимации (МЭА) и эллипсоидальной линеаризации (МЭЛ). В [5, 6] разработана теория аналитического синтеза МЭУОФ на основе приближенного решения по МЭА (МЭЛ) фильтрационного уравнения для ненормированной апостериорной характеристической функции.

Рассмотрим обобщение [4, 6] на случай ЭУОФ. В основу рассмотрения положим соответствующие уравнения для ненормированных апостериорных распределений.

2 Точные фильтрационные уравнения для апостериорного распределения

Следуя [5–7], будем рассматривать задачу фильтрации состояния систем, моделями которых могут служить стохастические дифференциальные уравнения, понимаемые в смысле Ито. При этом стохастические дифференциальные уравнения модели изучаемой системы могут иметь неизвестные параметры и, как правило, всегда содержат параметры, известные с ограниченной точностью. Поэтому возникает задача непрерывного оценивания неизвестных параметров системы (точнее, ее модели) по результатам непрерывных наблюдений. Предположим, что правые части уравнений зависят от конечного множества неизвестных парамет-

*Работа выполнена при поддержке РФФИ (проект 15-07-02244).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sinitsin@dol.ru

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, vsinitsin@ipiran.ru

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ekorepanov@ipiran.ru

ров, которые будем рассматривать как компоненты вектора параметров θ. Одним из возможных подходов в таких случаях является следующий прием: неизвестный векторный параметр θ считают стохастическим процессом $\Theta = \Theta_t$, который определяется дифференциальным уравнением $\dot{\Theta}_t = 0$, и включают компоненты этого векторного процесса в вектор состояния системы («расширяют» вектор состояния путем включения в него неизвестных параметров в качестве дополнительных компонент). Таким образом, задача непрерывного оценивания неизвестных параметров модели системы сводится к задаче непрерывного оценивания состояния системы с расширенным вектором состояния. От неизвестных параметров могут зависеть и уравнения наблюдения. Эти параметры следует включить в вектор Θ и, следовательно, в расширенный вектор состояния.

Итак, пусть векторный стохастический процесс (СтП) $\left[X_t^{\mathrm{T}} Y_t^{\mathrm{T}}\right]^{\mathrm{T}}$ определяется системой векторных стохастических дифференциальных уравнений Ито [1–3]:

$$dX_t = \varphi\left(X_t, Y_t, \Theta, t\right) dt + \psi'\left(X_t, Y_t, \Theta, t\right) dW_0 + \int\limits_{R_0^q} \psi''\left(X_t, Y_t, \Theta, t, v\right) P^0(dt, dv)\,,$$

$$X(t_0) = X_0\,; \quad (1)$$

$$dY_t = \varphi_1\left(X_t, Y_t, \Theta, t\right) dt + \psi_1'\left(X_t, Y_t, \Theta, t\right) dW_0 + \int\limits_{R_0^q} \psi_1''(X_t, Y_t, \Theta, t, v) P^0(dt, dv)\,,$$

$$Y(t_0) = Y_0\,. \quad (2)$$

Здесь $Y_t = Y(t)$ — n_y-мерный наблюдаемый СтП, $Y_t \in \Delta^y$ (Δ^y — гладкое многообразие наблюдений); $X_t = X(t)$ — n_x-мерный ненаблюдаемый СтП (вектор состояния), $X_t \in \Delta^x$ (Δ^x — гладкое многообразие состояний); $W_0 = W_0(t)$ — n_w-мерный винеровский СтП ($n_w \geq n_y$) интенсивности $\nu_0 = \nu_0(\Theta, t)$; $P^0(\Delta, A) = P(\Delta, A) - \mu_P(\Delta, A)$, $P(\Delta, A)$ представляет собой для любого множества A простой пуассоновский СтП, а $\mu_P(\Delta, A)$ — его математическое ожидание, причем

$$\mu_P(\Delta, A) = \mathsf{M} P(\Delta, A) = \int\limits_{\Delta} \nu_P(\tau, A)\, d\tau\,;$$

$\nu_P(\Delta, A)$ — интенсивность соответствующего пуассоновского потока событий, $\Delta = (t_1, t_2]$; интегрирование по v распространяется на все пространство R^q с выколотым началом координат; Θ — вектор случайных параметров размерности n_Θ; $\varphi = \varphi(X_t, Y_t, \Theta, t)$, $\varphi_1 = \varphi_1(X_t, Y_t, \Theta, t)$, $\psi' = \psi'(X_t, Y_t, \Theta, t)$ и $\psi_1' = \psi_1'(X_t, Y_t, \Theta, t)$ — известные функции, отображающие $R^{n_x} \times R^{n_y} \times R$ соответственно в R^{n_x}, R^{n_y}, $R^{n_x n_w}$ и $R^{n_y n_w}$; $\psi'' = \psi''(X_t, Y_t, \Theta, t, v)$ и $\psi_1''(X_t, Y_t, \Theta, t, v)$ — известные функции, отображающие $R^{n_x} \times R^{n_y} \times R^q$ в R^{n_x} и R^{n_y}. Требуется найти оценку $\hat{X}_t$ СтП X_t в каждый момент времени t по результатам наблюдения СтП $Y(\tau)$ до момента t, $Y_{t_0}^t = \{Y(\tau) : t_0 \leq \tau < t\}$.

Предположим, что выполнены условия [1–3]:

- уравнение состояния имеет вид (1);
- уравнение наблюдения (2), во-первых, не содержит пуассоновского шума ($\psi_1'' \equiv 0$), а во-вторых, коэффициент при винеровском шуме ψ_1' в уравнениях наблюдения не зависит от состояния ($\psi_1'(X_t, Y_t, \Theta, t) = \psi_1'(Y_t, \Theta, t)$).

В этом случае уравнения задачи нелинейной фильтрации имеют следующий вид:

$$dX_t = \varphi\left(X_t, Y_t, \Theta, t\right) dt + \psi'\left(X_t, Y_t, \Theta, t\right) dW_0 + \int\limits_{R_0^q} \psi''\left(X_t, Y_t, \Theta, t, v\right) P^0(dt, dv)\,,$$

$$X\left(t_0\right) = X_0\,; \quad (3)$$

$$dY_t = \varphi_1\left(X_t, Y_t, \Theta, t\right) dt + \psi_1\left(Y_t, \Theta, t\right) dW_0\,,$$

$$Y\left(t_0\right) = Y_0\,. \quad (4)$$

Предположим, что выполнены условия существования и единственности СтП, определяемого (3) и (4) при соответствующих начальных условиях [7–9].

Как известно [5, 9], для любых СтП X_t и Y_t оптимальная оценка $\hat{X}^t$, минимизирующая средний квадрат ошибки в каждый момент времени t, представляет собой апостериорное математическое ожидание СтП X_t: $\hat{X}_t = \mathsf{M}\left[X_t \mid Y_{t_0}^t\right]$. Чтобы найти это условное математическое ожидание, необходимо знать $p_t = p_t(x)$ и $g_t = g_t(\lambda)$ — апостериорную одномерную плотность и характеристическую функцию распределения СтП X_t. Соответствующие точные уравнения линейной фильтрации приведены в [3].

Введем ненормированные одномерные апостериорные плотность $\tilde{p}_t(x, \Theta)$ и характеристическую функцию $\tilde{g}_t(\lambda, \Theta)$ согласно формулам:

$$\tilde{p}_t(x, \Theta) = \mu_t p_t(x, \Theta)\,;$$

$$\tilde{g}_t(\lambda, \Theta) = \mathsf{M}_{\Delta^x}^{p_t}\left[e^{i\lambda^{\mathrm{T}} X_t} \mu_t\right] = \mu_t g_t(\lambda, \Theta)\,.$$

Тогда, обобщая [5] на случай уравнений (3) и (4), получим следующее точное уравнение с.к. оптимальной нелинейной фильтрации:

$$d\tilde{g}_t(\lambda,\Theta) = \mathsf{M}^{\tilde{p}_t}_{\Delta^x}\Big\{\Big[i\lambda^{\mathrm{T}}\varphi\,(X,Y_t,\Theta,t) - \\ - \frac{1}{2}\left(\psi'\nu_0\psi'^{\mathrm{T}}\right)(X,Y_t,\Theta,t) + \\ + \int\limits_{R_0^q}\Big[e^{i\lambda^{\mathrm{T}}\psi''(X,Y_t,\Theta,t,v)} - 1 - \\ - i\lambda^{\mathrm{T}}\psi''\,(X,Y_t,\Theta,t,v)\Big]\nu_P(\Theta,t,dv)\Big]e^{i\lambda^{\mathrm{T}}X}\Big\}\,dt + \\ + \mathsf{M}^{\tilde{p}_t}_{\Delta^x}\Big\{\Big[\varphi_1\,(X,Y_t,\Theta,t)^{\mathrm{T}} + \\ + i\lambda^{\mathrm{T}}\left(\psi'\nu_0\psi'^{\mathrm{T}}\right)(X,Y_t,\Theta,t)\Big]e^{i\lambda^{\mathrm{T}}X}\Big\}\times \\ \times\left(\psi'\nu_0\psi'^{\mathrm{T}}\right)^{-1}(Y_t,\Theta,t)\;dY_t\,. \quad (5)$$

Если, следуя [8], функция ψ'' в (3) допускает представление

$$\psi'' = \psi'\omega(\Theta,v)\,,$$

где $P^0(\Delta,A) = P^0((0,t],dv)$, то уравнения (3) и (4) примут следующий вид:

$$\dot{X}_t = \varphi\,(X_t,Y_t,\Theta,t) + \psi'\,(X_t,Y_t,\Theta,t)\,V(\Theta,t)\,, \\ X\,(t_0) = X_0\,; \quad (6)$$

$$\dot{Y}_t = \varphi\,(X_t,Y_t,\Theta,t) + \psi_1\,(Y_t,\Theta,t)\,V_0(\Theta,t)\,, \\ Y\,(t_0) = Y_0\,. \quad (7)$$

Здесь $V_0(\Theta,t) = \dot{W}_0(\Theta,t)$; $V(\Theta,t) = \dot{\bar{W}}(\Theta,t)$,

$$\bar{W}(\Theta,t) = W_0(\Theta,t) + \int\limits_{R_0^q}\omega(\Theta,v)P^0((0,t],dv)\,,$$

где $\nu_P(\Theta,t,v)dv = [\partial\mu(\Theta,t,v)/\partial t]\,dv$ — интенсивность пуассоновского потока скачков, равных $\omega(\Theta,t)$. При этом логарифмические производные от одномерных характеристических функций определяются известными формулами:

$$\chi^{W_0}(\rho;\Theta,t) = -\frac{1}{2}\,\rho^{\mathrm{T}}\nu_0(\Theta,t)\rho\,,$$

$$\chi^{\bar{W}}(\rho;\Theta,t) = -\frac{1}{2}\,\rho^{\mathrm{T}}(\Theta,t)\rho^{\mathrm{T}} + \\ + \int\limits_{R_0^q}\left[e^{i\rho^{\mathrm{T}}\omega(\Theta,v)} - 1 - i\rho^{\mathrm{T}}\omega(\Theta,v)\right]\nu_P(\Theta,t,v)\,dv\,.$$

В таком случае интегральный член в (5) допускает следующую запись:

$$\gamma = \int\limits_{R_0^q}\Big[e^{i\lambda^{\mathrm{T}}\psi''(X_t,Y_t,\Theta,t)\omega(\Theta,v)} - 1 - \\ - i\lambda^{\mathrm{T}}\psi''\,(X_t,Y_t,\Theta,t)\,\omega(\Theta,v)\Big]\nu_P(\Theta,t,v)\,dv\,. \quad (8)$$

Очевидно, что для гауссовской МСтС $\gamma \equiv 0$. Тогда приходим к известным утверждениям [5, 8, 10].

Теорема 1. *Пусть для негауссовской МСтС* (3), (4) *выполнены условия существования и единственности. Тогда уравнение с.к. оптимальной нелинейной фильтрации для ненормированной характеристической функции* $\tilde{g}_t(\lambda,\Theta)$ *имеет вид* (5).

Теорема 2. *Пусть для гауссовской МСтС* (6), (7) *выполнены условия существования и единственности. Тогда уравнение с.к. оптимальной нелинейной фильтрации для ненормированной характеристической функции имеет вид* (5) *при условии* (8).

Как известно [5], необходимость обработки результатов наблюдений в реальном масштабе времени непосредственно в процессе эксперимента привела к появлению ряда приближенных методов оптимальной нелинейной фильтрации, называемых обычно методами условно-оптимальной фильтрации. В этом случае для приближенного решения уравнения для апостериорной одномерной характеристической функции $g_1(\lambda,\Theta)$ вектора X_t можно использовать методы аналитического моделирования, основанные на параметризации одномерных распределений СтП, определяемого стохастическим дифференциальным уравнением. Эти методы позволяют изучить стохастические дифференциальные уравнения для параметров апостериорного распределения. Простейшим таким методом является МНА апостериорного распределения. Исключительно важное практическое значение имеют квазилинейные фильтры, получаемые с помощью методов эквивалентной линеаризации [5]. Эллипсоидальные УОФ, основанные на приближенном решении уравнений для нормированных апостериорных распределений, рассмотрены в [3], а для ненормированных — в [6].

Наряду с методами субоптимальной фильтрации широкое применение нашли методы, средства и информационные технологии, основанные на принципах УОФ В. С. Пугачёва [5].

3 Уравнения субоптимальной фильтрации

Следуя [6], аппроксимируем ненормированную апостериорную плотность вероятности формулой

$$\tilde{p}_t(x) \approx p_t^*(u) = w(u)\left[\mu + \sum_{\nu=1}^{N} c_\nu p_\nu(u)\right]\,. \quad (9)$$

Имеем

$$c_\nu = \mu_t\mathsf{M}^{\Im\mathrm{A}}\left[q_\nu(U_t)\right] = \left[q_\nu(U_\lambda)\tilde{g}_t^{\Im\mathrm{A}}(\lambda)\right]_{\lambda=0}\,, \quad (10)$$

причем

$$u = \left(x^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right) C_t \left(x - \hat{X}_t\right);$$
$$U_t = \left(X_t^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right) C_t \left(X_t - \hat{X}_t\right);$$
$$U_\lambda = \left(\frac{\partial^{\mathrm{T}}}{i\partial\lambda} - \hat{X}_t\right) C_t \left(\frac{\partial}{i\partial\lambda} - \hat{X}_t\right).$$

Как известно [3, 6], для того чтобы составить стохастические дифференциальные уравнения для коэффициентов c_ν, надо найти стохастический дифференциал Ито произведения $q_\chi(u)\tilde{g}_t(\lambda)$, имея в виду, что u зависит от $\hat{X}_t = m_t/\mu_t$ и что m_t и μ_t определяются стохастическими дифференциальными уравнениями. Потом следует заменить переменные x и u операторами $\partial/(i\partial\lambda)$ и U_λ, выполнить дифференцирование и после этого положить $\lambda = 0$.

Повторяя [6], сначала получим уравнения для m_t и μ_t с функцией $\hat{\varphi}_1$, определяемой формулой

$$\hat{\varphi}_1 = \mathsf{M}^{p_t}_{\Delta^x}[\varphi_1];$$

с учетом обозначений

$$\sigma_0 = \psi\nu_0\psi^{\mathrm{T}};\ \ \sigma_1 = \psi\nu_0\psi_1^{\mathrm{T}};\ \ \sigma_2 = \psi_1\nu_0\psi_1^{\mathrm{T}}$$

их можно представить в виде:

$$dm_t = \left[\mu_t f_0\left(Y_t, \hat{X}_t, \Theta, t\right) + \right.$$
$$\left. + \sum_{\nu=1}^{N} c_\nu f_\nu\left(Y_t, \hat{X}_t, \Theta, t\right)\right] dt + \left[\mu_t h_0\left(Y_t, \hat{X}_t, \Theta, t\right) + \right.$$
$$\left. + \sum_{\nu=1}^{N} c_\nu h_\nu\left(Y_t, \hat{X}_t, \Theta, t\right)\right] dY_t; \quad (11)$$

$$d\mu_t = \left[\mu_t b_0\left(Y_t, \hat{X}_t, t\right) + \right.$$
$$\left. + \sum_{\nu=1}^{N} c_\nu b_\nu\left(Y_t, \hat{X}_t, \Theta, t\right)\right] dY_t, \quad (12)$$

где

$$\left.\begin{aligned} f_0 &= f_0\left(Y_t, \hat{X}_t, , \Theta, t\right) = \mathsf{M}^w_{\Delta^x}[\varphi];\\ f_\nu &= f_\nu\left(Y_t, \hat{X}_t, , \Theta, t\right) = \mathsf{M}^{wp_\nu}_{\Delta^x}[\varphi]; \end{aligned}\right\} \quad (13)$$

$$\left.\begin{aligned} h_0 &= h_0\left(Y_t, \hat{X}_t, \Theta, t\right) =\\ &= \mathsf{M}^w_{\Delta^U}\left[\sigma_1(Y_t, \Theta, t) + X\varphi_1(X, Y_t, \Theta, t)^{\mathrm{T}}\right] \times\\ &\qquad \times \sigma_2(Y_t, \Theta, t)^{-1};\\ h_\nu &= h_\nu\left(Y_t, \hat{X}_t, \Theta, t\right) =\\ &= \mathsf{M}^{wp_\nu}_{\Delta^U}\left[\sigma_1(X, Y_t, \Theta, t) + X\varphi_1(X, Y_t, \Theta, t)^{\mathrm{T}}\right] \times\\ &\qquad \times \sigma_2(Y_t, \Theta, t)^{-1}; \end{aligned}\right\} \quad (14)$$

$$\left.\begin{aligned} b_0 &= b_0\left(Y_t, \hat{X}_t, \Theta, t\right) =\\ &= \mathsf{M}^w_{\Delta^U}\left[\varphi_1(X, Y_t, \Theta, t)^{\mathrm{T}}\right] \sigma_2(Y_t, \Theta, t)^{-1};\\ b_\nu &= b_\nu\left(Y_t, \hat{X}_t, \Theta, t\right) =\\ &= \mathsf{M}^{wp_\nu}_{\Delta^U}\left[\varphi_1(X, Y_t, \Theta, t)\right] \sigma_2(Y_t, \Theta, t)^{-1}. \end{aligned}\right\} \quad (15)$$

Далее запишем уравнения (11) и уравнение для одномерной ненормированной характеристической функции в виде:

$$dm_t = f\,dt + h\,dY_t; \quad d\mu_t = b\,dY_t;$$
$$d\tilde{g}_t = A\,dt + B\,dY_t.$$

Здесь обозначено:

$$f = \mu_t f_0 + \sum_{\nu=1}^{N} c_\nu f_\nu; \quad h = \mu_t h_0 + \sum_{\nu=1}^{N} c_\nu h_\nu;$$
$$b = \mu_t b_0 + \sum_{\nu=1}^{N} c_\nu b_\nu;$$

$$A = \mathsf{M}^{\tilde{p}_t}_{\Delta^x}\left\{i\lambda^{\mathrm{T}}\varphi(X, Y_t, \Theta, t) - \right.$$
$$- \frac{1}{2}\lambda^{\mathrm{T}}\left(\psi'\nu_0\psi'^{\mathrm{T}}\right)(X, Y_t, \Theta, t)\,\lambda \times$$
$$\times \int_{R_0^q}\left[e^{i\lambda^{\mathrm{T}}\psi''(X, Y_t, \Theta, t, v)} - 1 - \right.$$
$$\left.\left. - i\lambda^{\mathrm{T}}\psi''(X, Y_t, \Theta, t, v)\right]\nu_P(t, dv)e^{i\lambda^{\mathrm{T}}X}\right\};$$

$$B = \mathsf{M}^{\tilde{p}_t}_{\Delta^x}\left[\varphi_1(X, Y_t, \Theta, t)^{\mathrm{T}} + \right.$$
$$\left. + i\lambda^{\mathrm{T}}\left(\psi'\nu_0\psi_1'^{\mathrm{T}}\right)(X, Y_t, \Theta, t)\right] \times$$
$$\times e^{i\lambda^{\mathrm{T}}X}\left(\psi_1'\nu_0\psi_1'^{\mathrm{T}}\right)^{-1}(X, Y_t, \Theta, t).$$

Дифференциальные уравнения для коэффициентов МОР в (9) и (10) в силу [6] имеют следующий вид:

$$dc_\chi = \mathsf{M}^{p^*}_{\Delta^x}\left\{q'_\chi(u)\left(2\varphi^{\mathrm{T}}C_t\left(X - \hat{X}_t\right) + \mathrm{tr}\left[C_t\sigma_0\right]\right) + \right.$$
$$+ 2q''_\chi(u)\left(X^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right)C_t\sigma_0C_t\left(X - \hat{X}_t\right) +$$
$$+ \int_{R_0^q}\left[q_\chi(\bar{u}) - q_\chi(u) - 2q'_\chi(u)\left(X^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right)C_t\psi''\right] \times$$
$$\times \nu_P(t, dv) - q'_\chi(u)\left(X^{\mathrm{T}} - \hat{X}_t^{\mathrm{T}}\right)C_t\left(h + \hat{X}_t b\right)\frac{\varphi_1}{\mu_t} +$$
$$+ q'_\chi(u)\frac{\mathrm{tr}\left[\left(h + \hat{X}_t b\right)\sigma_1^{\mathrm{T}}C_t\right]}{\mu_t} + 2q''_\chi(u) \times$$

$$\times \frac{\mathrm{tr}\left[\left(h+\hat{X}_t b\right)\sigma_1^{\mathrm{T}} C_t\left(X-\hat{X}_t\right)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)C_t\right]}{\mu_t}\Bigg\}dt+$$

$$+\left\{\frac{1}{2n}\left(c_{\chi-1}+2\chi c_\chi\right)\mathrm{tr}\left[\dot{C}_t K_t\right]+\frac{c_{\chi-1}}{2n}\times\right.$$

$$\left.\times\frac{\mathrm{tr}\left[C_t h\sigma_2 h^{\mathrm{T}}\right]-2\hat{X}_t^{\mathrm{T}}C_t h\sigma_2 b^{\mathrm{T}}+\hat{X}_t^{\mathrm{T}}C_t\hat{X}_t b\sigma_2 b^{\mathrm{T}}}{\mu_t^2}\right\}dt+$$

$$+\mathsf{M}_{\Delta^x}^{p^*}\left\{\left[q_\chi(u)\varphi_1^{\mathrm{T}}+\right.\right.$$

$$\left.\left.+q'_\chi(u)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)C_t\sigma_1\right]\sigma_2^{-1}\right\}dY_t\,. \quad (16)$$

Примем в дополнение к обозначениям (13)–(15) следующие:

$$\gamma_{\chi 0}=\gamma_{\chi 0}\left(Y_t,\hat{X}_t,\Theta,t\right)=$$

$$=\mathsf{M}_{\Delta^x}^{w}\left\{q'_\chi(u)\left(2\varphi\left(X,Y_t,\Theta,t\right)^{\mathrm{T}}C_t\left(X-\hat{X}_t\right)+\right.\right.$$

$$\left.\left.+\mathrm{tr}\left[C_t\sigma_0\left(X,Y_t,\Theta,t\right)\right]\right)+\right.$$

$$+2q''_\chi(u)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)C_t\sigma_0\left(X,Y_t,\Theta,t\right)C_t\left(X-\hat{X}_t\right)+$$

$$+\int\limits_{R_0^q}\left[q_\chi\left(\bar{u}\right)-q_\chi(u)-2q'_\chi(u)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)\times\right.$$

$$\left.\left.\times C_t\psi''\left(X,Y_t,\Theta,t,v\right)\right]\nu_P(t,dv)\right\};$$

$$\gamma_{\chi\nu}=\gamma_{\chi\nu}\left(Y_t,\hat{X}_t,\Theta,t\right)=$$

$$=\mathsf{M}_{\Delta^x}^{wp_\nu}\left\{q'_\chi(u)\left(2\varphi\left(X,Y_t,\Theta,t\right)^{\mathrm{T}}C_t\left(X-\hat{X}_t\right)+\right.\right.$$

$$\left.\left.+\mathrm{tr}\left[C_t\sigma_0\left(X,Y_t,\Theta,t\right)\right]\right)+\right.$$

$$+2q''_\chi(u)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)C_t\sigma_0(X,Y_t,\Theta,t)\,C_t\left(X-\hat{X}_t\right)+$$

$$+\int\limits_{R_0^q}\left[q_\chi\left(\bar{u}\right)-q_\chi(u)-2q'_\chi(u)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)\times\right.$$

$$\left.\left.\times C_t\psi''\left(X,Y_t,\Theta,t,v\right)\right]\nu_P(t,dv)\right\};$$

$$\varepsilon_{\chi 0}=\varepsilon_{\chi 0}\left(Y_t,\hat{X}_t,\Theta,t\right)=$$

$$=\mathsf{M}_{\Delta^x}^{w}\left\{q'_\chi(u)\left[\sigma_1\left(X,Y_t,\Theta,t\right)^{\mathrm{T}}-\right.\right.$$

$$\left.\left.-\varphi_1\left(X,Y_t,\Theta,t\right)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)\right]+\right.$$

$$\left.+2q''_\chi(u)\sigma_1\left(X,Y_t,\Theta,t\right)^{\mathrm{T}}C_t\left(X-\hat{X}_t\right)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)\right\};$$

$$\varepsilon_{\chi\nu}=\varepsilon_{\chi\nu}\left(Y_t,\hat{X}_t,\Theta,t\right)=$$

$$=\mathsf{M}_{\Delta^x}^{wp_\nu}\left\{q'_\chi(u)\left[\sigma_1\left(X,Y_t,\Theta,t\right)^{\mathrm{T}}-\right.\right.$$

$$\left.\left.-\varphi_1\left(X,Y_t,\Theta,t\right)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)\right]+\right.$$

$$\left.+2q''_\chi(u)\sigma_1(X,Y_t,\Theta,t)^{\mathrm{T}}C_t\left(X-\hat{X}_t\right)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)\right\};$$

$$\eta_{\chi 0}=\eta_{\chi 0}\left(Y_t,\hat{X}_t,\Theta,t\right)=$$

$$=\mathsf{M}_{\Delta^x}^{w}\left\{q_\chi(u)\varphi_1\left(X,Y_t,\Theta,t\right)^{\mathrm{T}}+q'_\chi(u)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)\times\right.$$

$$\left.\times C_t\sigma_1\left(X,Y_t,\Theta,t\right)\right\}\sigma_2\left(Y_t,\Theta,t\right)^{-1};$$

$$\eta_{\chi\nu}=\eta_{\chi\nu}\left(Y_t,\hat{X}_t,\Theta,t\right)=$$

$$=\mathsf{M}_{\Delta^x}^{wp_\nu}\left\{q_\chi(u)\varphi_1\left(X,Y_t,\Theta,t\right)^{\mathrm{T}}+q'_\chi(u)\left(X^{\mathrm{T}}-\hat{X}_t^{\mathrm{T}}\right)\times\right.$$

$$\left.\times C_t\sigma_1\left(X,Y_t,\Theta,t\right)\right\}\sigma_2\left(Y_t,\Theta,t\right)^{-1}.$$

Тогда можем переписать уравнения (16) в виде:

$$dc_\chi=\left\{\mu_t\gamma_{\chi 0}\left(Y_t,\hat{X}_t,\Theta,t\right)+\right.$$

$$\left.+\sum_{\nu=1}^{N}c_\nu\gamma_{\chi\nu}\left(Y_t,\hat{X}_t,\Theta,t\right)+\right.$$

$$+\mathrm{tr}\left[\mu_t\left(h_0\left(Y_t,\hat{X}_t,\Theta,t\right)+\hat{X}_t b_0\left(Y_t,\hat{X}_t,\Theta,t\right)\right)+\right.$$

$$+\sum_{\nu=1}^{N}c_\nu\left(h_\nu\left(Y_t,\hat{X}_t,\Theta,t\right)+\hat{X}_t b_\nu\left(Y_t,\hat{X}_t,\Theta,t\right)\right)\times$$

$$\times\left\{\varepsilon_{\chi 0}\left(Y_t,\hat{X}_t,\Theta,t\right)+\right.$$

$$\left.\left.+\sum_{\nu=1}^{N}\frac{c_\nu\varepsilon_{\chi\nu}\left(Y_t,\hat{X}_t,\Theta,t\right)}{\mu_t}\right\}C_t\right]+$$

$$+\frac{1}{2n}\left(c_{\chi-1}+2\chi c_\chi\right)\mathrm{tr}\left[\dot{C}_t K_t\right]+$$

$$+\frac{c_{\chi-1}}{2n}\mathrm{tr}\left[C_t\left(h_0\left(Y_t,\hat{X}_t,\Theta,t\right)+\right.\right.$$

$$\left.\left.+\sum_{\nu=1}^{N}\frac{c_\nu h_\nu\left(Y_t,\hat{X}_t,t\right)}{\mu_t}\right)\times\right.$$

$$\times\sigma_2\left(Y_t,\Theta,t\right)\left(h_0\left(Y_t,\hat{X}_t,\Theta,t\right)^{\mathrm{T}}+\right.$$

$$\left.\left.\left.+\sum_{\nu=1}^{N}\frac{c_\nu h_\nu\left(Y_t,\hat{X}_t,\Theta,t\right)^{\mathrm{T}}}{\mu_t}\right)\right]-\right.$$

$$
\begin{aligned}
&- 2\hat{X}_t^{\mathrm{T}} C_t \left(h_0\left(Y_t, \hat{X}_t, \Theta, t\right) + \right. \\
&\left. + \sum_{\nu=1}^{N} \frac{c_\nu h_\nu\left(Y_t, \hat{X}_t, \Theta, t\right)}{\mu_t} \right) \sigma_2\left(Y_t, \Theta, t\right) \times \\
&\times \left(b_0\left(Y_t, \hat{X}_t, \Theta, t\right)^{\mathrm{T}} + \sum_{\nu=1}^{N} \frac{c_\nu b_\nu\left(Y_t, \hat{X}_t, \Theta, t\right)^{\mathrm{T}}}{\mu_t} \right) + \\
&+ \hat{X}_t^{\mathrm{T}} C_t \hat{X}_t \left(b_0\left(Y_t, \hat{X}_t, \Theta, t\right) + \right. \\
&\left. + \sum_{\nu=1}^{N} c_\nu \frac{b_\nu\left(Y_t, \hat{X}_t, \Theta, t\right)}{\mu_t} \right) \sigma_2\left(Y_t, \Theta, t\right) \times \\
&\times \left(b_0\left(Y_t, \hat{X}_t, \Theta, t\right) + \sum_{\nu=1}^{N} c_\nu \frac{b_\nu\left(Y_t, \hat{X}_t, \Theta, t\right)^{\mathrm{T}}}{\mu_t} \right) \Bigg\} dt + \\
&+ \left\{ \mu_t \eta_{\chi 0}\left(Y_t, \hat{X}_t, \Theta, t\right) + \sum_{\nu=1}^{N} c_\nu \eta_{\chi\nu}\left(Y_t, \hat{X}_t, \Theta, t\right) \right\} dY_t \\
&\qquad (\chi = 1, \ldots, N). \quad (17)
\end{aligned}
$$

Уравнения (11), (12), (17) и соотношение $\hat{X}_t = m_t/\mu_t$ при начальных условиях

$$
\left.
\begin{gathered}
m\left(t_0\right) = \mathrm{M}\left[X_0 \mid Y_0\right]; \;\; \mu\left(t_0\right) = 1; \\
c_\chi\left(t_0\right) = c_{\chi 0} \;\; (\chi = 1, \ldots, N),
\end{gathered}
\right\} \quad (18)
$$

где $c_{\chi 0}$ $(\chi = 1, \ldots, N)$ — коэффициенты разложения (9) условной плотности вероятности $\tilde{p}_{t_0}(x) = p_0(x \mid Y_0)$ вектора X_0 относительно Y_0, определяют МЭСОФ.

После решения уравнений (12) и (17) с.к. оптимальная оценка вектора состояния и ковариационная матрица ошибки фильтрации в МЭСОФ определяются по следующим приближенным формулам:

$$
\hat{X}_t = \frac{m_t}{\mu_t}; \quad (19)
$$

$$
\begin{aligned}
R_t &= \mathsf{M}^{w}_{\Delta^x}\left[\left(X - \frac{m_t}{\mu_t}\right)\left(X^{\mathrm{T}} - \frac{m_t^{\mathrm{T}}}{\mu_t}\right)\right] + \\
&+ \sum_{\nu=1}^{N} \frac{c_\nu}{\mu_t} \mathsf{M}^{wp_\nu}_{\Delta^x}\left[\left(X - \frac{m_t}{\mu_t}\right)\left(X^{\mathrm{T}} - \frac{m_t^{\mathrm{T}}}{\mu_t}\right)\right]. \quad (20)
\end{aligned}
$$

Порядок МЭСОФ особенно при большой размерности n вектора состояния системы значительно ниже порядка других УОФ. Так, при учете в разложении (12) моментов до десятого порядка, уже при $n > 3$, $N = 5$ имеем $n + N + 1 \le n(n+3)/2$. При $n > 3$, $N = 5$ МЭУОФ имеет более низкий порядок, чем фильтры МНА, обобщенный фильтр Калмана–Бьюси, фильтры второго порядка и гауссовский фильтр.

Таким образом, в основе алгоритма модифицированной эллипсоидальной условно-оптимальной нелинейной фильтрации лежат следующие утверждения.

Теорема 3. *В условиях теоремы* 1*, если МЭУОФ существует, то он определяется уравнениями* (11), (12) и (17) *при условиях* (18)–(20).

Теорема 4. *В условиях теоремы* 2*, если МЭУОФ существует, то он определяется уравнениями теоремы* 3 *при условиях* (8).

Следуя [6] для приближенного анализа фильтрационных уравнений и учитывая случайность параметров Θ, придем к следующим уравнениям для функций чувствительности первого порядка [3]:

$$
\left.
\begin{aligned}
d\nabla^\Theta \hat{X}_s &= \nabla^\Theta A^{\hat{X}_s}\, dt + \nabla^\Theta B^{\hat{X}_s}\, dY_t\,, \\
&\qquad \nabla^\Theta B^{\hat{X}_s}(t_0) = 0\,; \\
d\nabla^\Theta R_{sq} &= \nabla^\Theta A^{R_{sq}}\, dt + \nabla^\Theta B^{R_{sq}}\, dY_t\,, \\
&\qquad \nabla^\Theta R_{sq}(t_0) = 0\,; \\
d\nabla^\Theta c_\kappa &= \nabla^\Theta A^{c_\kappa}\, dt + \nabla^\Theta B^{c_\kappa}\, dY_t\,, \\
&\qquad \nabla^\Theta c_\kappa(t_0) = 0\,.
\end{aligned}
\right\} \quad (21)
$$

Здесь процедура взятия производных осуществляется по всем входящим переменным, а коэффициенты чувствительности вычисляются при $\Theta = m^\Theta$. При этом предполагается малость дисперсий по сравнению с их математическими ожиданиями. Очевидно, что при дифференцировании по Θ $(\nabla^\Theta = \partial/\partial\Theta)$ порядок уравнений возрастает пропорционально числу производных. Аналогично составляются уравнения для элементов матриц вторых функций чувствительности.

Для оценки качества МЭСОФ, следуя [6], при гауссовских Θ с математическим ожиданием m^Θ и ковариационной матрицей K^Θ введем условную функцию потерь, допускающую квадратичную аппроксимацию:

$$
\begin{aligned}
\rho^{\hat{X}_s} = \rho^{\hat{X}_s}(\Theta) &= \rho\left(m^\Theta\right) + \sum_{ii=1}^{n^\Theta} \rho'_i\left(m^\Theta\right)\Theta^0_s + \\
&+ \sum_{i,j=1}^{n^\Theta}\sum_{i,j=1}^{n^\Theta} \rho''_{ij}\left(m^\Theta\right)\Theta^0_i\Theta^0_j\,, \quad (22)
\end{aligned}
$$

а также показатель ε

$$
\varepsilon = \varepsilon_2^{1/4}\,, \quad (23)
$$

где введено обозначение:

$$\varepsilon_2 = \mathsf{M}^{\text{ЭА}}\left[\rho(\Theta)^2\right] - \rho\left(m^{\Theta}\right)^2. \qquad (24)$$

Здесь

$$\mathsf{M}^{\text{ЭА}}\left[\rho(\Theta)^2\right] = \rho\left(m^{\Theta}\right)^2 + \rho'\left(m^{\Theta}\right)^{\mathrm{T}} K^{\Theta}\rho'\left(m^{\Theta}\right) + \\ + 2\rho\left(m^{\Theta}\right)\mathrm{tr}\left[\rho''\left(m^{\Theta}\right)K^{\Theta}\right] + \left\{\mathrm{tr}\left[\rho''\left(m^{\Theta}\right)K^{\Theta}\right]\right\}^2 + \\ + 2\mathrm{tr}\left[\rho''\left(m^{\Theta}\right)K^{\Theta}\right]^2,$$

при этом функции ρ' и ρ'' по известным формулам определяются на основе первых и вторых функций чувствительности. Таким образом, в основе оценки качества МЭСОФ, в условиях теорем 3 и 4, лежат уравнения (21)–(24) при существовании соответствующих производных в правых частях уравнений (21) (**теорема 5**).

Изложенные выше методы синтеза МЭСОФ дают принципиальную возможность получить фильтр, близкий к оптимальному по оценке с любой степенью точности. Чем выше максимальный порядок учитываемых моментов, коэффициент ЭА, тем выше будет точность приближения к оптимальной оценке. Однако число уравнений, определяющих параметры апостериорного одномерного эллипсоидального распределения, быстро растет с увеличением числа учитываемых параметров.

4 Основные классы модифицированных эллипсоидальных условно-оптимальных фильтров

4.1. Рассмотрим задачу условно-оптимальной фильтрации, когда требуется найти оптимальную оценку $\hat{X}_t$ процесса X_t в момент $t > t_0$ по результатам наблюдения этого процесса до момента t (т. е. на интервале $[t_0, t)$) в классе допустимых оценок, определяемых формулой $\hat{X}_t = AU_t$ и стохастическим дифференциальным уравнением вида

$$dU_t = \\ + \left[\alpha_t\xi(Y_t, \hat{X}_t, t) + \gamma_t\right]dt + \beta_t\eta\left(Y_t, \hat{X}_t, t\right)dY_t \qquad (25)$$

при заданных векторной и матричной структурных функциях ξ и η и при всех возможных функциях времени α_t, β_t и γ_t (α_t и β_t — матрицы; γ_t — вектор). В качестве критерия оптимальности примем критерий минимума с.к. ошибки оценки $\hat{X}_t$. Как известно [5], самым сложным вопросом в практических применениях теории условно-оптимальной фильтрации является вопрос о выборе класса допустимых фильтров, определяемого заданием структурных функций ξ и η в уравнении (25). В принципе, их можно задать произвольно. При желании можно выбрать ξ и η так, чтобы класс допустимых фильтров содержал произвольно заданный УОФ. В этом случае такой фильтр будет практически всегда точнее заданного условно-оптимального. В то же время, выбрав в качестве компонент векторной функции ξ и элементов матричной функции η конечный отрезок некоторого базиса в соответствующем гильбертовом пространстве L_2, можно получить приближение с любой степенью точности к неизвестным оптимальным функциям ξ и η. По-видимому, это наиболее рациональный способ выбора функций ξ, η, основанный на уравнениях теории условно-оптимальной фильтрации (см. разд. 3). При этом новые возможности открывают уравнения УОФ, полученные из уравнения для ненормированной апостериорной характеристической функции.

4.2. Для применения уравнений МЭСОФ, полученных из ненормированных уравнений для апостериорного распределения, необходимо изменить постановку задач условно-оптимальной фильтрации [5] так, чтобы использовать уравнение для множителя μ_t.

С этой целью воспользуемся для определения класса допустимых МЭУОФ (25) уравнениями:

$$d\mu_t = \rho_t\chi\left(Y_t, \hat{X}_t, t\right)dY_t; \qquad (26)$$
$$\hat{X}_t = \frac{A\hat{X}_t}{\mu_t},$$

где $\chi(Y_t, \hat{X}_t, t)$ — некоторая заданная матричная функция; ρ_t — матрица-строка коэффициентов, зависящих от t и подлежащих оптимизации наряду с коэффициентами α_t, β_t и γ_t в уравнении фильтра (25).

Основываясь на результатах разд. 3, можно выделить два класса допустимых МЭУОФ.

4.3. Первый класс. Этот важный класс допустимых МЭУОФ можно получить, положив $U_t = m_t$, $A = I_n$ и определив функции ξ, η и χ в (25) и (26), руководствуясь уравнениями (11) и (12). Это дает следующие выражения для структурных функций:

$$\xi = \xi\left(Y_t, \hat{X}_t, t\right) = \\ = \left[\mu_t f_0\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}} f_1\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}} \cdots \right. \\ \left. \cdots f_N\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}}\right]^{\mathrm{T}};$$

$$\eta = \eta\left(Y_t, \hat{X}_t, t\right) = \left[\mu_t h_0\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}} h_1\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}} \cdots \cdots h_N\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}}\right]^{\mathrm{T}};$$

$$\chi = \chi\left(Y_t, \hat{X}_t, t\right) = \left[\mu_t b_0\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}} b_1\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}} \cdots \cdots b_N\left(Y_t, \frac{\hat{X}_t}{\mu_t}, t\right)^{\mathrm{T}}\right]^{\mathrm{T}},$$

при этом порядок МЭУОФ, определяемого уравнениями (25) и (26), будет равен $n+1$.

4.4. Второй класс. Для получения более широкого класса допустимых МЭУОФ перепишем уравнение (17) в виде:

$$dc_\chi = \left\{F_{\chi 0}\left(Y_t, \hat{X}_t, t\right) + \sum_{\nu=1}^{N} c_\nu F_{\chi\nu}\left(Y_t, \hat{X}_t, t\right) + \sum_{\lambda,\nu=1}^{N} c_\lambda c_\nu F_{\chi\lambda\nu}\left(Y_t, \hat{X}_t, t\right) + c_{\chi-1}\sum_{\lambda,\nu=1}^{N} c_\lambda c_\nu F'_{\chi\lambda\nu}\left(Y_t, \hat{X}_t, t\right)\right\} dt + \left\{\mu_t \eta_{\chi 0}\left(Y_t, \hat{X}_t, t\right) + \sum_{\nu=1}^{N} c_\nu \eta_{\chi\nu}\left(Y_t, \hat{X}_t, t\right)\right\} dY_t. \quad (27)$$

Здесь введены следующие обозначения:

$$F_{\chi 0}\left(Y_t, \hat{X}_t, t\right) = \mu_t \gamma_{\chi 0}\left(Y_t, \hat{X}_t, t\right) + \mu_t \mathrm{tr}\left[\left(h_0\left(Y_t, \hat{X}_t, t\right) + \hat{X}_t b_0\left(Y_t, \hat{X}_t, t\right)\right)\varepsilon_{\chi 0}\left(Y_t, \hat{X}_t, t\right)\right];$$

$$F_{\chi\nu}\left(Y_t, \hat{X}_t, t\right) = \gamma_{\chi 0}\left(Y_t, \hat{X}_t, t\right) + \mathrm{tr}\left[\left(h_\nu\left(Y_t, \hat{X}_t, t\right) + \hat{X}_t b_\nu\left(Y_t, \hat{X}_t, t\right)\right)\varepsilon_{\chi 0}\left(Y_t, \hat{X}_t, t\right) + \left(h_0\left(Y_t, \hat{X}_t, t\right) + \hat{X}_t b_0\left(Y_t, \hat{X}_t, t\right)\right)\varepsilon_{\chi\nu}\left(Y_t, \hat{X}_t, t\right)\right] +$$

$$+ \frac{1}{2n}\delta_{\chi-1,\nu}\left\{\mathrm{tr}\left[b_\nu K_t + C_t h_0\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) h_0\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}}\right] - 2\hat{X}_t^{\mathrm{T}} C_t h_0\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) b_0\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}} + \hat{X}_t^{\mathrm{T}} C_t \hat{X}_t b_0\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) b_0\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}}\right\} + \frac{1}{n}\chi\delta_{\chi\nu}\mathrm{tr}\left[\dot{C}_t K_t\right];$$

$$F_{\chi\lambda\nu} = \frac{1}{\mu_t}\mathrm{tr}\left[C_t\left(h_\lambda\left(Y_t, \hat{X}_t, t\right) + \hat{X}_t b_\lambda\left(Y_t, \hat{X}_t, t\right)\right)\varepsilon_{\chi\nu}\left(Y_t, \hat{X}_t, t\right)\right] + \frac{1}{n}\delta_{\chi-1,\lambda}\frac{1}{\mu_t} \times \left\{\mathrm{tr}\left[C_t h_0\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) h_\nu\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}}\right] - \hat{X}_t^{\mathrm{T}} C_t h_0\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) b_\nu\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}} + h_\nu\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) b_0\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}} + \hat{X}_t^{\mathrm{T}} C_t \hat{X}_t b_0\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) b_\nu\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}}\right\};$$

$$F'_{\chi\lambda\nu} = \frac{1}{2n}\frac{1}{\mu_t^2}\left\{\mathrm{tr}\left[C_t h_\lambda\left(Y_t, \hat{X}_t, t\right) + \sigma_2\left(Y_t, t\right) h_\nu\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}}\right] - 2\hat{X}_t^{\mathrm{T}} C_t h_\lambda\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) b_\nu\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}} + \hat{X}_t^{\mathrm{T}} C_t \hat{X}_t b_\lambda\left(Y_t, \hat{X}_t, t\right)\sigma_2\left(Y_t, t\right) b_\nu\left(Y_t, \hat{X}_t, t\right)^{\mathrm{T}}\right\}.$$

Взяв за основу для построения класса допустимых МЭУОФ уравнения (11), (12) и (27), следует принять за компоненты вектора $\hat{X}_t$ все компоненты вектора m_t и коэффициенты $c_1, \ldots, c_N$, так что $\hat{X}_t = \left[m_t^{\mathrm{T}} c_1 \cdots c_N\right]^{\mathrm{T}}$. Порядок всех допустимых фильтров при этом равен $n+N+1$.

4.5. Для нахождения коэффициентов α_t, β_t и γ_t уравнения МЭУОФ (25) необходимо знать совместное одномерное распределение случайных процессов X_t и $\hat{X}_t$. Это распределение находится путем решения задачи анализа системы, описываемой стохастическими дифференциальными уравнениями (25) и (26). Как всегда в теории условно-оптимальной фильтрации, все сложные вычисления, необходимые для нахождения оптимальных коэффициентов уравнения МЭУОФ (25) или (26), основаны только на априорных данных и поэтому могут быть выполнены заранее в процессе проектирования МЭУОФ. При этом может быть определена и точность фильтрации для каждого допустимо-

го МЭУОФ. Сам же процесс фильтрации сводится к решению дифференциального уравнения, что дает возможность производить фильтрацию в реальном времени.

Таким образом, приходим к следующему результату.

Теорема 6. *В условиях теоремы* 1 *уравнения МЭУОФ вида* (25), (26) *будут совпадать с уравнениями МЭСОФ* (*см. разд.* 3)*, если структурные функции УОФ выбрать из описанных выше классов. При этом качество МЭУОФ оценивается согласно теореме* 5.

5 Квазилинейные модифицированные эллипсоидальные условно-оптимальные фильтры

Построим квазилинейный МЭУОФ на основе МЭЛ для МСтС (1), (2) при $\psi' = \psi'(\Theta, t)$, $\psi'' = \psi''(\Theta, t, v)$, $\psi_1' = \psi_1'(\Theta, t)$ и $\psi_1'' = \psi_1''(\Theta, t, v)$ (т. е. с аддитивными винеровскими и пуассоновскими шумами). Уравнения ЭСОФ проще получаются, если нелинейные функции φ и φ_1 на основе эллипсоидального распределения с известным c_ν заменить на статистически линеаризованные:

$$\varphi = \varphi\left(X_t, Y_t, \Theta, t\right) \approx \approx \varphi_0^{\text{э}} + k_x^{\text{э}\varphi}\left(X_t - m_t^x\right) + k_y^{\text{э}\varphi}\left(Y_t - m_t^y\right); \quad (28)$$

$$\varphi_1 = \varphi_1\left(X_t, Y_t, \Theta, t\right) \approx \approx \varphi_{10}^{\text{э}} + k_x^{\text{э}\varphi_1}\left(X_t - m_t^x\right) + k_y^{\text{э}\varphi_1}\left(Y_t - m_t^y\right), \quad (29)$$

а затем использовать уравнения линейной фильтрации [5]. Входящие в (28), (29) коэффициенты с МЭЛ зависят от математических ожиданий, дисперсий и ковариаций:

$$Z_t = \begin{bmatrix} X_t \\ Y_t \end{bmatrix}; \; m_t^z = \begin{bmatrix} m_t^x \\ m_t^y \end{bmatrix}; \; K_t^z = \begin{bmatrix} K_t^x & K_t^{xy} \\ K_t^{xy} & K_t^y \end{bmatrix}.$$

Они определяются из уравнений:

$$\dot{Z}_t = A^z Z_t + A_0^z + B_0^z V, \;\; V = \dot{W};$$
$$\dot{m}_t^z = A^z m_t^z + A_0^z, \;\; m_{t_0}^Z = m_0^z;$$
$$\dot{K}_t^z = B^z K_t^z + K_t^z (B^z)^{\mathrm{T}} + B_0^z \nu^m (B_0^z)^{\mathrm{T}}, \; K_{t_0}^z = K_0^z.$$

Здесь введены следующие обозначения:

$$A_0^z = \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}; \; A^z = \begin{bmatrix} a_1 & a \\ b_1 & b \end{bmatrix}; \; B_0^z = \begin{bmatrix} \bar{\psi} \\ \bar{\psi}_1 \end{bmatrix};$$

$$a = k_y^{\text{э}\varphi}; \;\; a_1 = k_x^{\text{э}\varphi}; \;\; a_0 = \varphi_0^{\text{э}} - k_x^{\text{э}\varphi} m_t^x - k_y^{\text{э}\varphi} m_t^y;$$
$$b = k_y^{\text{э}\varphi_1}; \;\; b_1 = k_x^{\text{э}\varphi_1}; \;\; b_0 = \varphi_0^{\text{э}} - k_x^{\text{э}\varphi_1} m_t^x - k_y^{\text{э}\varphi_1} m_t^y;$$

$$\left.\begin{aligned} \psi\, dW_0 + \int\limits_{R_0^q} \psi'' P^0(dt, dv) &= \bar{\psi}\, dW; \\ \psi_1'\, dW_0 + \int\limits_{R_0^q} \psi_1'' P^0(dt, dv) &= \bar{\psi}_1\, dW, \end{aligned}\right\} \quad (30)$$

где ν^W — интенсивность СтП с независимыми приращениями, состоящего из винеровской и пуассоновской частей (30). Тогда уравнения квазилинейного ЭУОФ будут иметь вид:

$$\hat{X}_t = aY_t + a_1\hat{X}_t + a_0 + + \beta_t\left[Z_t - \left(bY_t + b_1\hat{X}_t + b_0\right)\right]. \quad (31)$$

Здесь

$$\beta_t = \left(R_t b_1^{\mathrm{T}} + \bar{\psi}\nu^W\bar{\psi}_1^{\mathrm{T}}\right)\left(\bar{\psi}_1\nu^W\bar{\psi}_1^{\mathrm{T}}\right)^{-1}, \quad (32)$$

где

$$\dot{R}_t = a_1 R_t + R_t a_1^{\mathrm{T}} + \bar{\psi}\nu^W\bar{\psi}^{\mathrm{T}} - - \left(R_t b_1^{\mathrm{T}} + \bar{\psi}\nu^W\bar{\psi}_1^{\mathrm{T}}\right)\left(\bar{\psi}_1\nu^W\bar{\psi}_1^{\mathrm{T}}\right)^{-1} \times \times \left(b_1 R_t + \bar{\psi}_1\nu^W\bar{\psi}^{\mathrm{T}}\right). \quad (33)$$

Теорема 7. *Пусть МСтС* (1), (2) *содержит только аддитивные винеровские и пуассоновские шумы и допускает замену статистически линеаризованной, а матрица* $\sigma_1 = \bar{\psi}_1\nu^W\bar{\psi}_1^{\mathrm{T}}$ *не вырождена. Тогда в основе алгоритма квазилинейного МЭУОФ лежат уравнения* (31)–(33) *при соответствующих начальных условиях.*

6 Заключение

Разработана теория аналитического синтеза МЭУОФ для нелинейных дифференциальных гауссовских и негауссовских МСтС (3), (4) и (6), (7) на основе МЭА и МЭЛ. Алгоритмы ЭУОФ положены в основу разрабатываемого модуля экспериментального программного обеспечения StS-Filter (version 2017).

Результаты допускают обобщение на случай дискретных и непрерывно-дискретных МСтС, а также автокоррелированных МСтС.

Теоретический и практический интерес представляет теория МЭСОФ и МЭУОФ для МСтС (1) и (2).

Литература

1. *Синицын И. Н.* Ортогональные субоптимальные фильтры для нелинейных стохастических систем на многообразиях // Информатика и её применения, 2016. Т. 10. Вып. 1. С. 34–44.

2. *Синицын И. Н.* Нормальные и ортогональные субоптимальные фильтры для нелинейных стохастических систем на многообразиях // Системы и средства информатики, 2016. Т. 26. № 1. С. 199–226.
3. *Синицын И. Н., Синицын В. И., Корепанов Э. Р.* Эллипсоидальные субоптимальные фильтры для нелинейных стохастических систем на многообразиях // Информатика и её применения, 2016. Т. 10. Вып. 2. С. 24–35.
4. *Синицын И. Н., Синицын В. И.* Лекции по теории нормальной и эллипсоидальной аппроксимации распределений в стохастических системах. — М.: ТОРУС ПРЕСС, 2013. 488 с.
5. *Синицын И. Н.* Фильтры Калмана и Пугачева. — 2-е изд. — М.: Логос, 2007. 776 с.
6. *Синицын И. Н., Синицын В. И., Корепанов Э. Р.* Модифицированные эллипсоидальные субоптимальные фильтры для нелинейных стохастических систем на многообразиях // Системы и средства информатики, 2016. Т. 26. № 2. С. 79–97.
7. *Пугачёв В. С., Синицын И. Н.* Теория стохастических систем. — М.: Логос, 2000; 2004. 1000 с.
8. *Wonham W. M.* Some applications of stochastic differential equations to optimal nonlinear filtering // J. Soc. Ind. Appl. Math. A, 1964. Vol. 2. No. 3. P. 347–369.
9. Справочник по теории вероятностей и математической статистике / Под ред. В. С. Королюка, Н. И. Портенко, А. В. Скорохода, А. Ф. Турбина. — М.: Наука, 1985. 640 с.
10. *Zakai M.* On the optimal filtering of diffusion processes // Ztschr. Wahrschein lichkeitstheor. Verm. Geb., 1969. Bd. 11. S. 230–243.

Поступила в редакцию 08.02.17

MODIFICATED ELLIPSOIDAL CONDITIONALLY OPTIMAL FILTERS FOR NONLINEAR STOCHASTIC SYSTEMS ON MANIFOLDS

I. N. Sinitsyn, V. I. Sinitsyn, and E. R. Korepanov

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The analytical synthesis theory for modificated ellipsoidal conditionally optimal filters (MECOF) for nonlinear stochastic systems on manifolds (MStS) based on the nonnormed a posteriori characteristic function is developed. Gaussian and non-Gaussian MStS are considered. The MECOF algorithms are more simple than the ECOF algorithms. The MECOF algorithms are the basis of the software tool "StS-Filter" (version 2017).

Keywords: accuracy and sensitivity equations; ellipsoidal approximation and linearization methods (EAM & ELM); ellipsoidal conditionally optimal filter (ECOF); modificated ellipsoidal conditionally optimal filter (MECOF); nonnormed characteristic function; Poisson noise; conditionally optimal filter (COF); Wiener noise

DOI: 10.14357/19922264170211

Acknowledgments

The research was supported by the Russian Foundation for Basic Research (project 15-07-002244).

References

1. Sinitsyn, I. N. 2016. Ortogonal'nye suboptimal'nye fil'try dlya nelineynykh stokhasticheskikh sistem na mnogoobraziyakh [Orthogonal suboptimal filters for nonlinear stochastic systems on manifolds]. *Informatika i ee Primeneniya — Inform. Appl.* 10(1):34–44.
2. Sinitsyn, I. N. 2016. Normal'nye i ortogonal'nye suboptimal'nye fil'try dlya nelineynykh stokhasticheskikh sistem na mnogoobraziyakh [Normal and orthogonal conditionally optimal filters for nonlinear stochastic systems on manifolds]. *Informatika i ee Primeneniya — Inform. Appl.* 10(1):199–226.
3. Sinitsyn, I. N., V. I. Sinitsyn, and E. R. Korepanov. 2016. Ellipsoidal'nye suboptimal'nye fil'try dlya nelineynykh stokhasticheskikh sistem na mnogoobraziyakh [Ellipsoidal conditionally optimal filters for nonlinear stochastic systems on manifolds]. *Informatika i ee Primeneniya — Inform. Appl.* 10(1):24–35.
4. Sinitsyn, I. N., and V. I. Sinitsyn. 2013. Lektsii po teorii normal'noy i ellipsoidal'noy approkskimatsii raspredeleniy v stokhasticheskikh sistemakh [Lectures on normal and ellipsoidal approximation of distributions in stochastic systems]. Moscow: TORUS PRESS. 488 p.
5. Sinitsyn, I. N. 2007. *Fil'try Kalmana i Pugacheva* [Kalman and Pugachev filters]. 2nd ed. Moscow: Logos. 776 p.

6. Sinitsyn, I. N., V. I. Sinitsyn, and E. R. Korepanov. 2016. Modifitsirovannye ellipsoidal'nye suboptimal'nye fil'try dlya nelineynykh stokhasticheskikh sistem na mnogoobraziyakh [Modificated ellipsoidal conditionally optimal filters for nonlinear stochastic systems on manifolds]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 26(2):79–97.
7. Pugachev, V. S., and I. N. Sinitsyn. 2000, 2004. Teoriya stokhasticheskikh sistem [Stochastic systems. Theory and applications]. Moscow: Logos. 1000 p.
8. Wonham, M. 1965. Some applications of stochastic differential equations to optimal nonlinear filtering. *J. Soc. Ind. Appl. Math. A* 2(3):347–369.
9. Korolyuk, V. S., N. I. Portenko, A. V. Skorokhod, and A. F. Turbin, eds. 1985. *Spravochnik po teorii veroyatnosti i matematicheskoy statistike* [Handbook: Probability theory and mathematical statistics]. Moscow: Nauka. 640 p.
10. Zakai, M. 1969. On the optimal filtering of diffusion processes. *Ztschr. Wahrschein lichkeitstheor. Verm. Geb.* 11:230–243.

Received February 8, 2017

Contributors

Sinitsyn Igor N. (b. 1940) — Doctor of Science in technology, professor, Honored scientist of RF, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sinitsin@dol.ru

Sinitsyn Vladimir I. (b. 1968) — Doctor of Science in physics and mathematics, associate professor, Head of Department, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; VSinitsyn@ipiran.ru

Korepanov Eduard R. (b. 1966) — Candidate of Science (PhD) in technology, Head of Department, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; Ekorepanov@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 112–116

ОДНОКАНАЛЬНАЯ СИСТЕМА ОБСЛУЖИВАНИЯ С ЗАВИСИМЫМИ ИНТЕРВАЛАМИ ВРЕМЕНИ МЕЖДУ ПОСТУПЛЕНИЯМИ ТРЕБОВАНИЙ*

В. Г. Ушаков[1], Н. Г. Ушаков[2]

Аннотация: Изучена одноканальная система массового обслуживания с бесконечным числом мест для ожидания и произвольным распределением времени обслуживания. Входящий поток требований является пуассоновским потоком со случайной интенсивностью. Текущее значение интенсивности выбирается из конечного множества с заданными вероятностями в момент начала отсчета времени до следующего поступления требования. Последовательные интенсивности образуют цепь Маркова специального вида. Частными случаями таких потоков являются гиперэкспоненциальные потоки и потоки, возникающие при исследовании байесовских моделей систем обслуживания с дискретным априорным распределением. Рассматриваемые потоки хорошо описывают работу систем массового обслуживания, функционирующих в случайной среде с конечным множеством различных состояний и марковской зависимостью между ними. Кроме того, такими потоками можно достаточно точно аппроксимировать реальные потоки в сетях передачи данных. Исследовано поведение длины очереди в нестационарном режиме.

Ключевые слова: пуассоновский поток; случайная интенсивность; гиперэкспоненциальный поток; цепь Маркова; одноканальная система; длина очереди

DOI: 10.14357/19922264170212

1 Введение

Статистический анализ трафика в различных телекоммуникационных сетях показывает ярко выраженную зависимость интервалов времени между поступлениями пакетов данных. В последние годы появилось много работ, в которых при построении математических моделей учитывается это явление. При моделировании процесса передачи информации важно концентрироваться на тех характеристиках трафика, которые могут быть эффективно оценены из реальных данных и поддаются физической интерпретации. С этой точки зрения перспективными являются модели, в которых те или иные параметры входящих потоков предполагаются случайными величинами, связанными регрессионной зависимостью небольшого порядка.

В работах [1–6] содержатся различные постановки задач в этом направлении и приведена обширная библиография. Потоки, рассматриваемые в настоящей статье, также относятся к потокам с регрессионной зависимостью параметров, в качестве которых выступают интенсивности. Кроме того, рассматриваемый класс потоков хорошо описывает работу систем обслуживания, функционирующих в случайной среде. Множество состояний среды конечно, и последовательные состояния связаны марковской зависимостью.

2 Входящий поток

В изучаемой системе обслуживания входящий поток имеет следующую структуру. Интервал времени до поступления первого требования z_1 и интервалы между поступлениями $(n-1)$-го и n-го требований z_n имеют показательное распределение со случайным параметром $a^{(n)}$, $n = 1, 2, \ldots$ Значение $a^{(n)}$ выбирается непосредственно перед началом промежутка z_n, причем $\mathbf{P}(a^{(1)} = a_j) = c_j$, $a_i \neq a_j$, $i \neq j$, $c_j > 0$, $j = 1, \ldots, N$, $\sum_{j=1}^{N} c_j = 1$ и $a^{(n)} = \alpha a^{(n-1)} + (1-\alpha) b^{(n)}$, где $b^{(n)}$, $n = 1, 2, \ldots,$ — последовательность независимых и независящих от последовательности $a^{(n)}$, $n = 1, 2, \ldots,$ одинаково

*Работа выполнена при финансовой поддержке РФФИ (проект 15-07-02354).

[1]Факультет вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, vgushakov@mail.ru

[2]Институт проблем технологии микроэлектроники и особочистых материалов Российской академии наук; Норвежский научно-технологический университет, Тронхейм, ushakov@math.ntnu.no

распределенных случайных величин, распределение которых такое же, как у $a^{(1)}$, а α не зависит от $a^{(n)}$ и $b^{(n)}$, $n = 1, 2, \ldots$, $\alpha = \begin{cases} 1, & p, \\ 0, & 1-p. \end{cases}$

Легко видеть, что

$$\mathbf{P}(z_n < t) = \sum_{j=1}^{N} c_j \left(1 - e^{-a_j t}\right) ;$$

$$\mathbf{P}(z_n < t_1, z_{n+1} < t_2) =$$
$$= (1-p) \sum_{j=1}^{N} c_j \left(1 - e^{-a_j t_1}\right) \sum_{k=1}^{N} c_k \left(1 - e^{-a_k t_2}\right) +$$
$$+ p \sum_{k=1}^{N} c_k \left(1 - e^{-a_k t_1}\right) \left(1 - e^{-a_k t_2}\right), \; n = 1, 2, \ldots$$

В частности, при $p = 0$ входящий поток будет гиперэкспоненциальным. При $p = 1$ получается система, в которой в начальный момент времени случайно выбирается значение интенсивности из множества $\{a_1, \ldots, a_N\}$ с вероятностями $c_1, \ldots, c_N$ и в дальнейшем система функционирует как система с пуассоновским входящим потоком с выбранной интенсивностью. В статье этот случай рассматриваться не будет.

Известно, что для любых $\mu > 0$ и $\sigma > \mu$ существует гиперэкспоненциальный поток второго порядка ($N = 2$), у которого математическое ожидание и дисперсия интервалов между поступлениями требований равны μ и σ^2. Коэффициент корреляции двух соседних интервалов для рассматриваемых в статье потоков при $N = 2$ составляет $(p/2)\left(1 - (\mu/\sigma)^2\right)$. Таким образом, появляется возможность не только подогнать первые два момента интервалов между поступления реального потока, но и учесть их зависимость.

3 Обозначения и определения

Пусть $B(x)$ и $b(x)$ — соответственно функция распределения и плотность распределения времен обслуживания требований. Обозначим

$$\beta(s) = \int_0^\infty e^{-sx} b(x)\, dx; \quad \eta(x) = \frac{b(x)}{1 - B(x)} .$$

Введем следующие случайные процессы:

$L(t)$ — число требований в системе в момент времени t;

$j(t)$ — процесс с состояниями $1, \ldots, N$ такой, что $j(t) = j$, если в момент времени t интенсивность входящего потока есть a_j;

$x(t)$ — время, прошедшее с начала обслуживания требования, находящегося на приборе, до момента t, если $L(t) > 0$, и $x(t) = 0$, если $L(t) = 0$.

Трехмерный случайный процесс $(L(t), x(t), j(t))$ является однородным марковским процессом. Положим

$$P_j(n, x, t) =$$
$$= \frac{\partial}{\partial x} \mathbf{P}(L(t) = n\,,\; j(t) = j\,,\; \nu(t) = 1,\; x(t) < x)\,,$$
$$n > 0\,,\;\; x \geqslant 0\,,$$
$$p_j(z, x, s) = \sum_{n=1}^{\infty} z^n \int_0^\infty e^{-st} P_j(n, x, t)\, dt\,;$$
$$P_{0j}(t) = \mathbf{P}(L(t) = 0\,,\; j(t) = j)\,;$$
$$p_{0j}(s) = \int_0^\infty e^{-st} P_{0j}(t)\, dt\,,\;\; j = 1, \ldots, N\,.$$

Справедливы следующие леммы.

Лемма1. *Уравнение*

$$(1-p) z \sum_{j=1}^{N} \frac{c_j a_j}{\mu + a_j (1 - pz)} = 1 \qquad (1)$$

имеет N непрерывных в области $|z| \leqslant 1$ решений $\mu_1(z), \ldots, \mu_N(z)$, причем:

(а) *только одна из этих функций обращается в нуль в точке $z = 1$;*

(б) *для всех $j = 1, \ldots, N$ при $|z| < 1$ справедливы неравенства $\mathrm{Re}(\mu_j(z)) < 0$;*

(в) *функции $\mu_i(z) \neq \mu_j(z)$ при $i \neq j$.*

Не ограничивая общности, будем считать, что $\mu_1(1) = 0$. Обозначим

$$\alpha_k(z) = \prod_{j \neq k} (\mu_k(z) - \mu_j(z))\,;\;\; \delta_{ij} = \begin{cases} 1, & i = j\,, \\ 0, & i \neq j\,. \end{cases}$$

Лемма 2. *При каждом k, $k = 1, \ldots, N$, уравнение*

$$z = \beta\left(s - \mu_k(z)\right)$$

имеет в области $\mathrm{Re}\, s > 0$ единственное решение $z = z_k(s)$ такое, что $|z_k(s)| < 1$.

Положим

$$\mu_k^{(*)}(s) = \mu_k\left(z_k(s)\right) .$$

4 Распределение длины очереди

Прямые уравнения Колмогорова для распределения процесса $(L(t),\ x(t),\ j(t))$ при $x > 0$ имеют вид:

$$\frac{\partial P_j(n,x,t)}{\partial t} + \frac{\partial P_j(n,x,t)}{\partial x} = \\ = -(a_j + \eta(x))P_j(n,x,t) + (1-\delta_{n,1})\times \\ \times \left(p\, a_j P_j(n-1,x,t) + \right. \\ \left. + (1-p)c_j \sum_{k=1}^{N} a_k P_k(n-1,x,t) \right), \quad (2)$$

а краевые условия при $x = 0$ и уравнения для $P_{0j}(t)$:

$$P_j(n,0,t) = \int_0^\infty P_j(n+1,x,t)\eta(x)\,dx + \delta_{n,1} \times \\ \times \left(p\, a_j P_{0j}(t) + (1-p)c_j \sum_{k=1}^{N} a_k P_{0k}(t) \right); \quad (3)$$

$$\frac{\partial P_{0j}(t)}{\partial t} = -a_j P_{0j}(t) + \int_0^\infty P_j(1,x,t)\eta(x)\,dx. \quad (4)$$

Начальные условия при $t = 0$ имеют вид:

$$P_j(n,x,0) = 0,\ \ n > 0,\ \ P_{0j}(0) = c_j,\ \ j = 1,\dots,N.$$

Переходя в (2)–(4) к производящим функциям и преобразованиям Лапласа, получаем:

$$\frac{\partial p_j(z,x,s)}{\partial x} = \\ = -\left(s + a_j - p\, a_j z + \eta(x)\right) p_j(z,x,s) + \\ + (1-p)c_j z \sum_{k=1}^{N} a_k p_k(z,x,s); \quad (5)$$

$$p_j(z,0,s) = z^{-1}\int_0^\infty p_j(z,x,s)\eta(x)\,dx + c_j - \\ - (s+a_j)p_{0j}(s) + \\ + z\left(p\, a_j p_{0j}(s) + (1-p)c_j \sum_{k=1}^{N} a_k p_{0k}(s) \right), \\ j = 1,\dots,N. \quad (6)$$

Общее решение системы дифференциальных уравнений (5) имеет вид:

$$p_j(z,x,s) = (1-B(x))c_j \times \\ \times \sum_{k=1}^{N} \frac{\gamma^{(k)}(z,s)}{\mu_k(z) + a_j(1-pz)}\, e^{-(s-\mu_k(z))x}, \quad (7)$$

где функции $\gamma^{(k)}(z,s)$ определяются из краевых условий. Подставляя (7) в (6), получаем:

$$\sum_{k=1}^{N} \frac{1 - z^{-1}\beta(s-\mu_k(z))}{\mu_k(z) + a_j(1-pz)}\,\gamma^{(k)}(z,s) = f_j(z,s), \quad (8)$$

где

$$f_j(z,s) = 1 - (s + a_j - a_j pz))\, c_j^{-1} p_{0j}(s) + \\ + (1-p)z\sum_{k=1}^{N} a_k p_{0k}(s).$$

Решая систему уравнений (8) относительно $\gamma^{(k)}(z,s)$, находим:

$$\left(1 - z^{-1}\beta(s-\mu_k(z))\right)\gamma^{(k)}(z,s) = \alpha_k^{-1}(z)\times \\ \times \sum_{l=1}^{N} f_l(z,s) \left(\prod_{j=1}^{N} \left(\mu_k(z) + a_j(1-pz)\right)\left(\mu_j(z) + \right.\right. \\ \left.\left. + a_l(1-pz)\right) \right) \Bigg/ \left(\prod_{\nu\neq l} (1-pz)\left(a_l - a_\nu\right)\left(\mu_k(z) + \right.\right. \\ \left.\left. + a_l(1-pz)\right) \right), \ \ k = 1,\dots,N.$$

Так как $\mu_1(z),\dots,\mu_N(z)$ являются решениями уравнения (1),

$$\prod_{j=1}^{N}(\mu + a_j(1-pz)) - \\ -(1-p)z\sum_{l=1}^{N} c_l a_l \prod_{j\neq l}(\mu + a_j(1-pz)) = \prod_{\nu=1}^{N}(\mu - \mu_\nu(z)).$$

Подставляя сюда $\mu = -a_l(1-pz)$, получаем:

$$\frac{\prod_{j=1}^{N}(\mu_j(z) + a_l(1-pz))}{\prod_{\nu\neq l}((1-pz)(a_l - a_\nu))} = (1-p)z c_l a_l.$$

Следовательно,

$$\left(1 - z^{-1}\beta\left(s-\mu_k(z)\right)\right)\gamma^{(k)}(z,s) = \alpha_k^{-1}(z)(1-)z \times \\ \times \prod_{j=1}^{N}(\mu_k(z) + a_j(1-pz)) \sum_{l=1}^{N} \frac{c_l a_l f_l(z,s)}{\mu_k(z) + a_l(1-pz)}, \\ k = 1,\dots,N. \quad (9)$$

В силу леммы 2 левая часть (9) обращается в 0 при $z = z^{(k)}(s)$. Значит,

$$\sum_{l=1}^{N} \frac{c_l\, a_l\, f_l(z^{(k)}(s), s)}{\mu_k^*(s) + a_l(1 - pz^{(k)}(s))} = 0, \quad k = 1, \ldots, N\,.$$

Отсюда, учитывая (1) и определение $f_l(z, s)$, получаем систему линейных уравнений для нахождения функций $p_{0j}(s)$, $j = 1, \ldots, N$:

$$\sum_{l=1}^{N} \frac{c_l a_l}{\mu_k^*(s) + a_l(1 - pz^{(k)}(s))} \Big(1 - \\ - (s + a_l(1 - pz^{(k)}(s)))c_l^{-1} p_{0l}(s)\Big) + \\ + \sum_{\nu=1}^{N} a_\nu p_{0\nu}(s) = 0, \quad k = 1, \ldots, N\,.$$

Таким образом, все функции, определяющие $p_j(z, x, s)$ и $p_{0j}(s)$, $j = 1, \ldots, N$, найдены.

Литература

1. *Hwang G. U., Choi B. D., Kim J.-K.* The waiting time analysis of a discrete-time queue with arrivals as an autoregressive process of order 1 // J. Appl. Probab., 2002. Vol. 39. No. 3. P. 619–629.
2. *Hwang G. U., Sohraby K.* On the exact analysis of a discrete-time queueing system with autoregressive inputs // Queueing Syst., 2003. Vol. 43. P. 29–41.
3. *Kamoun F.* The discrete-time queue with autoregressive inputs revisited // Queueing Syst., 2006. Vol. 54. P. 185–192.
4. *Леонтьев Н. Д., Ушаков В. Г.* Анализ системы обслуживания с входящим потоком авторегрессионного типа // Информатика и её применения, 2014. Т. 8. Вып. 3. С. 39–44.
5. *Леонтьев Н. Д., Ушаков В. Г.* Исследование систем обслуживания с дискретным временем, входящим потоком авторегрессионного типа и обратной связью // Системы и средства информатики, 2015. Т. 25. № 2. С. 61–71.
6. *Леонтьев Н. Д., Ушаков В. Г.* Анализ системы обслуживания с входящим потоком авторегрессионного типа и относительным приоритетом // Информатика и её применения, 2016. Т. 10. Вып. 3. С. 15–22.

Поступила в редакцию 02.03.17

SINGLE SERVER QUEUEING SYSTEM WITH DEPENDENT INTERARRIVAL TIMES

V. G. Ushakov[1,2] and N. G. Ushakov[3,4]

[1]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, Moscow 119991, GSP-1, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[3]Institute of Microelectronics Technology and High-Purity Materials of the Russian Academy of Sciences, 6 Academician Osipyan Str., Chernogolovka, Moscow Region 142432, Russian Federation
[4]Norwegian University of Science and Technology, 15A S. P. Andersensvei, Trondheim 7491, Norway

Abstract: The paper studies a single server queueing system with an infinite number of positions in the queue and random distribution of the service time. The incoming flow of claims is a Poisson flow with a random intensity. The current intensity value is selected from a finite set with given probabilities at the start of the countdown to the next receipt of the claim. Sequential intensities form a Markov chain of a special kind. Particular cases of such flows are hyperexponential flows and flows arising in the study of Bayesian models of queueing systems with a discrete prior distribution. Considered flows describe well the work of queueing systems operating in a random environment with a finite set of different states and Markov relationship between them. Furthermore, such flows can accurately approximate real flows in data networks. The nonstationary behavior of the queue length is studied.

Keywords: Poisson flow; random intensity; hyperexponential flow; Markov chain; single server; queue length

DOI: 10.14357/19922264170212

Acknowledgments

This work was supported by the Russian Foundation for Basic Research (project No. 15-07-02354).

References

1. Hwang, G. U., B. D. Choi, and J.-K. Kim. 2002. The waiting time analysis of a discrete-time queue with arrivals as an autoregressive process of order 1. *J. Appl. Probab.* 39(3):619–629.
2. Hwang, G. U., and K. Sohraby. 2003. On the exact analysis of a discrete-time queueing system with autoregressive inputs. *Queueing Syst.* 43:29–41.
3. Kamoun, F. 2006. The discrete-time queue with autoregressive inputs revisited. *Queueing Syst.* 54:185–192.
4. Leont'ev, N. D., and V. G. Ushakov. 2014. Analiz sistemy obsluzhivaniya s vkhodyashchim potokom avtoregressionnogo tipa [Analysis of a queueing system with autoregressive arrivals]. *Informatika i ee Primeneniya — Inform. Appl.* 8(3):39–44.
5. Leont'ev, N. D., and V. G. Ushakov. 2015. Issledovanie sistem obsluzhivaniya s diskretnym vremenem, vkhodyashchim potokom avtoregressionnogo tipa i obratnoy svyaz'yu [A study of queueing systems with discrete time, autoregressive arrivals, and feedback]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 25(2):61–71.
6. Leont'ev, N. D., and V. G. Ushakov. 2016. Analiz sistemy obsluzhivaniya s vkhodyashchim potokom avtoregressionnogo tipa i otnositel'nym prioritetom [Analysis of a queueing system with autoregressive arrivals and nonpreemptive priority]. *Informatika i ee Primeneniya — Inform. Appl.* 10(3):15–22.

Received March 2, 2017

Contributors

Ushakov Vladimir G. (b. 1952) — Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, Moscow 119991, GSP-1, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; vgushakov@mail.ru

Ushakov Nikolai G. (b. 1952) — Doctor of Science in physics and mathematics, leading scientist, Institute of Microelectronics Technology and High-Purity Materials of the Russian Academy of Sciences, 6 Academician Osipyan Str., Chernogolovka, Moscow Region 142432, Russian Federation; professor, Norwegian University of Science and Technology, 15A S. P. Andersensvei, Trondheim 7491, Norway; ushakov@math.ntnu.no

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 117–121

СИЛЬНАЯ СОСТОЯТЕЛЬНОСТЬ ОЦЕНКИ СРЕДНЕКВАДРАТИЧНОЙ ПОГРЕШНОСТИ ПРИ РЕШЕНИИ ОБРАТНЫХ СТАТИСТИЧЕСКИХ ЗАДАЧ*

О. В. Шестаков[1]

Аннотация: Нелинейные методы обработки сигналов и изображений с помощью процедур пороговой обработки коэффициентов вейвлет-разложений стали популярным аппаратом для задач подавления шума и компрессии. Объясняется это тем, что вейвлет-анализ позволяет гораздо более эффективно исследовать нестационарные сигналы, чем традиционный Фурье-анализ, благодаря возможности лучшей адаптации к функциям, имеющим на разных участках различную степень регулярности. Анализ погрешностей этих методов представляет собой важную практическую задачу, поскольку позволяет оценить качество как самих методов, так и используемого оборудования. В некоторых приложениях данные наблюдаются не напрямую, а после применения некоторого линейного преобразования. Задача обращения такого преобразования, как правило, некорректно поставлена, что приводит к росту дисперсии шума. В работе исследуются асимптотические свойства оценки среднеквадратичной погрешности при обращении линейных однородных операторов методами вейвлет-вейглет-разложения и пороговой обработки. При довольно слабых ограничениях доказывается сильная состоятельность этой оценки.

Ключевые слова: вейвлеты; пороговая обработка; несмещенная оценка риска; коррелированный шум; асимптотическая нормальность

DOI: 10.14357/19922264170213

1 Введение

Методы вейвлет-анализа широко применяются при анализе и обработке зашумленных данных. Во многих статистических задачах эти данные измеряются не напрямую, а после некоторого преобразования. В таких случаях построение вейвлет-оценок осуществляется с помощью вейвлет-вейглет-разложения и процедуры мягкой пороговой обработки. Порог обычно зависит от уровня разложения, и его можно выбирать различными способами, исходя из постановки задачи и целей обработки. Наличие шума неизбежно приводит к погрешностям. Свойства оценки таких погрешностей (риска) в модели с независимым шумом подробно исследовались в [1–6]. Модель с коррелированным шумом исследовалась в [7–9]. Показано, что при определенных условиях оценка риска является состоятельной и асимптотически нормальной. В данной работе доказывается сильная состоятельность оценки риска.

2 Модель данных и вейвлет-вейглет-разложение

В данной работе рассматривается следующая модель данных:

$$Y_i = (Kf)_i + \epsilon_i\,,\ \ i = 1, \ldots, 2^J\,, \tag{1}$$

где K — некоторый линейный оператор; f — функция, которую необходимо оценить (предполагается, что f принадлежит области определения K); $\{\epsilon_i,\ i \in Z\}$ — стационарный гауссовский процесс с ковариационной последовательностью $r_k = \mathrm{cov}(\epsilon_i, \epsilon_{i+k})$ (предполагается, что ϵ_i имеют нулевое среднее и единичную дисперсию).

В данной работе рассматривается модель долгосрочной зависимости $r_k \sim Ak^{-\alpha}$, $0 < \alpha < 1$, $A > 0$. Как показано в работе [10], случай краткосрочной зависимости, т. е. когда $\sum_{-\infty}^{+\infty} |r_k| < \infty$, эквивалентен модели с независимым шумом.

Во многих случаях нельзя оценить функцию f, просто применив к данным обратный оператор K^{-1}, поскольку такой оператор либо не существует, либо не является ограниченным. Статистические задачи такого рода называются некорректно поставленными.

Для решения такого рода задач в работе [11] предложен метод так называемого вейвлет-вейглет-разложения, который хорошо зарекомендовал себя при обращении линейных однородных операторов, т. е. таких операторов K, для которых выполнено

*Работа выполнена при частичной финансовой поддержке РФФИ (проект 15-07-02652).

[1]Московский государственный университет им. М. В. Ломоносова, кафедра математической статистики факультета вычислительной математики и кибернетики; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, oshestakov@cs.msu.su

$$K\left[f\left(a\left(x-x_0\right)\right)\right]=a^{-\beta}(Kf)\left[a\left(x-x_0\right)\right]$$

для любого x_0 и любого $a>0$. Параметр β называется показателем однородности. Примерами линейных однородных операторов служат оператор интегрирования, преобразование Абеля и операторы Рисса.

Вейвлет-разложение функции $f\in L^2(\mathbb{R})$ представляет собой ряд

$$f=\sum_{j,k\in Z}\langle f,\psi_{jk}\rangle\psi_{jk}, \tag{2}$$

где $\psi_{jk}(t)=2^{j/2}\psi(2^jt-k)$, а $\psi(t)$ — некоторая материнская вейвлет-функция (семейство $\{\psi_{jk}\}_{jk\in Z}$ образует ортонормированный базис в $L^2(\mathbb{R})$). Индекс j в (2) называется масштабом, а индекс k — сдвигом.

В дискретной постановке задачи функция задана в отсчетах на конечном отрезке. Дискретное вейвлет-преобразование представляет собой умножение вектора значений функции f на ортогональную матрицу W, определяемую вейвлет-функцией ψ. При этом в силу ортогональности матрицы дискретные вейвлет-коэффициенты связаны с непрерывными следующим образом: $\mu_{jk}\approx 2^{J/2}\langle f,\psi_{jk}\rangle$, где 2^J — число отсчетов функции f [12]. Всюду далее предполагается, что используются вейвлеты Мейера [12], преобразование Фурье которых обладает необходимым количеством непрерывных производных.

Поскольку наблюдается не функция f, а Kf, коэффициенты разложения в (2) вычислить напрямую нельзя. Идея метода вейвлет-вейглет-разложения заключается в том, чтобы выразить коэффициенты разложения в (2) через Kf. Если оператор K однороден, то существует последовательность функций ξ_{jk} такая, что $\langle Kf,\xi_{jk}\rangle=\langle f,\psi_{jk}\rangle$. Нормированные функции $u_{jk}=2^{-\beta j}\xi_{jk}$ получили название «вейглеты». Своими свойствами они очень похожи на вейвлеты (если соответствующие вейвлеты удовлетворяют определенным условиям гладкости [11]), за исключением свойства ортогональности. Таким образом, ряд (2) можно переписать в виде

$$f=\sum_{j,k\in Z}2^{-\beta j}\langle Kf,u_{jk}\rangle\psi_{jk}\,,$$

которое и представляет собой вейвлет-вейглет-разложение.

При применении дискретного аналога этого разложения к модели (1) по аналогии с дискретным вейвлет-преобразованием получается следующая модель дискретных вейглет-коэффициентов [7]:

$$X_{jk}=\mu_{j,k}+2^{J(1-\alpha)/2}\varepsilon_{jk}\,,\qquad j=0,\ldots,J-1;\;\; k=0,\ldots,2^j-1\,,$$

где $\mu_{j,k}=2^{J/2}\langle Kf,\xi_{j,k}\rangle$, $\varepsilon_{jk}=\int\xi_{j,k}\,d\mathbf{B}_H$, а $\mathbf{B}_H(t)$ — процесс дробного броуновского движения с $H=1-\alpha/2$. Дисперсии коэффициентов X_{jk} не зависят от k и равны [10]

$$\sigma_j^2=C2^{(J-j)(1-\alpha)}2^{2\beta j}\,,$$

где C — положительная константа, зависящая от параметров A, α и β.

3 Оценка среднеквадратичной погрешности

Для подавления шума в методе вейвлет-вейглет-разложения используется процедура пороговой обработки коэффициентов, смысл которой заключается в удалении достаточно маленьких коэффициентов, которые считаются шумом. В данной работе рассматривается так называемая мягкая пороговая обработка. К каждому коэффициенту применяется функция $\rho_T(x)=\operatorname{sgn}(x)\left(|x|-T\right)_+$, т. е. коэффициенты, которые по модулю меньше порога T, обнуляются, а абсолютные величины остальных коэффициентов уменьшаются на величину порога.

Среднеквадратичная погрешность (или риск) мягкой пороговой обработки определяется следующим образом:

$$R_J(f)=\sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\mathbb{E}\left(\mu_{jk}-\rho_T(X_{jk})\right)^2\,.$$

В [13] было предложено использовать порог $T_j=\sigma_j\sqrt{2\ln 2^j}$ и показано, что при таком пороге среднеквадратичная ошибка близка к минимальной [12]. Этот порог получил название «универсальный». В дальнейшем будет использоваться именно такой вид порога.

Вычислить $R_J(f)$ в явном виде нельзя, так как в выражении присутствуют неизвестные «чистые» коэффициенты μ_{jk}. Однако его можно оценить с помощью следующей величины:

$$\widehat{R}_J(f)=\sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}F\left[X_{jk}^2,T\right]\,, \tag{3}$$

где $F[x,T]=(x-\sigma^2)\mathbb{1}(|x|\leqslant T^2)+(\sigma^2+T^2)\mathbb{1}(|x|>T^2)$. В работе [1] было показано, что $\widehat{R}_J(f)$ является несмещенной оценкой $R_J(f)$.

4 Сильная состоятельность оценки среднеквадратичной погрешности

В работе [7] показано, что (3) при определенных условиях гладкости на функцию f является асимптотически нормальной и состоятельной оценкой $R_J(f)$. Оказывается, что эта оценка является также сильно состоятельной даже при более слабых ограничениях.

Для доказательства этого утверждения потребуется следующая лемма Боска [14], в которой оценивается вероятность отклонения суммы ограниченных слабозависимых случайных величин от ее математического ожидания.

Лемма. *Пусть $\{X_i,\ i \in Z\}$ — последовательность случайных величин таких, что $\mathbb{E}X_i = 0$ и $|X_i| \leqslant b$ п.в. для всех $i \in Z$, где $b > 0$ — некоторая константа. Тогда для любого $q \in [1, n/2]$ и любого $\varepsilon > 0$*

$$\mathsf{P}\left(\left|\sum_{i=1}^{n} X_i\right| > n\varepsilon\right) \leqslant 4\exp\left\{-\frac{\varepsilon^2}{8b^2}q\right\} + 22\left(1+\frac{4b}{\varepsilon}\right)^{1/2} q\alpha\left(\left[\frac{n}{2q}\right]\right), \quad (4)$$

где $\alpha(k)$ — коэффициент α-перемешивания последовательности $\{X_i,\ i \in Z\}$.

Докажем теперь сильную состоятельность оценки (3).

Теорема. *Пусть $f \in L^2(\mathbb{R})$ задана на конечном отрезке, а K — линейный однородный оператор с показателем $\beta > 0$. Тогда имеет место сходимость*

$$\frac{\widehat{R}_J(f) - R_J(f)}{2^{\lambda J}} \to 0 \ \text{ п.в. при } J \to \infty \qquad (5)$$

при любом $\lambda > 1/2 + 2\beta$ в случае $\alpha + 2\beta \geqslant 1/2$ и любом $\lambda > 1 - \alpha$ в случае $\alpha + 2\beta < 1/2$.

Доказательство. Пусть $0 < p < 1$ некоторое число, которое будет выбрано позднее. Представим числитель (5) в виде $\widehat{R}_J(f) - R_J(f) = \widehat{R}_1 + \widehat{R}_2$, где

$$\widehat{R}_1 = \sum_{j=0}^{[pJ]-1}\sum_{k=0}^{2^j-1}\left(F\left[X_{jk}^2, T\right] - \mathbb{E}F\left[X_{jk}^2, T\right]\right);$$

$$\widehat{R}_2 = \sum_{j=[pJ]}^{J-1}\sum_{k=0}^{2^j-1}\left(F\left[X_{jk}^2, T\right] - \mathbb{E}F\left[X_{jk}^2, T\right]\right).$$

Рассмотрим $\widehat{R}_2$. Для произвольного $\delta > 0$ имеем:

$$p_J = \mathsf{P}\left(\left|\widehat{R}_2\right| > \delta 2^{\lambda J}\right) \leqslant \leqslant \sum_{j=[pJ]}^{J-1}\mathsf{P}\left(\left|\sum_{k=0}^{2^j-1}\left(F\left[X_{jk}^2, T\right] - \mathbb{E}F\left[X_{jk}^2, T\right]\right)\right| > > \delta J^{-1}2^{\lambda J}\right). \quad (6)$$

Из вида функции $F[x,T]$ следует, что $-\sigma_j^2 \leqslant \leqslant F[X_{jk}^2, T_j] \leqslant \sigma_j^2 + T_j^2$. В работе [10] показано, что в силу свойств вейвлетов Мейера при каждом j слагаемые в сумме под вероятностью в (6) удовлетворяют свойству ρ-перемешивания с коэффициентом $\rho(k) \leqslant Ck^{-M}$, где $C > 0$ — некоторая константа, а M можно сделать достаточно большим, выбрав соответствующий вейвлет Мейера.

Известно [15], что для коэффициентов α-перемешивания и ρ-перемешивания справедливо неравенство $4\alpha(k) \leqslant \rho(k)$. Применяя неравенство (4) с $q = 2^{\theta j}$ ($\theta < 1$) для каждого j в сумме (6) и выбирая M достаточно большим, получаем:

$$p_J \leqslant c_1 J \times \times \max_{[pJ] \leqslant j \leqslant J-1}\left\{\exp\left[-c_2 J^{-3}2^{2(\lambda-1+\alpha)J+(\theta-2\alpha-4\beta)j}\right]\right\} + + o_J. \quad (7)$$

Здесь c_1 и c_2 — некоторые положительные константы, а o_J убывает по J быстрее, чем $2^{-M_0 pJ}$, где M_0 — некоторое положительное число, зависящее от M.

Если $\alpha + 2\beta \geqslant 1/2$, то $\theta - 2\alpha - 4\beta < 0$, и при $j = J$ правая часть (7) не превосходит $c_1 J \exp\left[-c_2 J^{-3}2^{(2\lambda-2+\theta-4\beta)J}\right]$. Поскольку $\theta < 1$ можно выбрать произвольно, для того чтобы выполнялось неравенство $2\lambda - 2 + \theta - 4\beta > 0$, достаточно потребовать $\lambda > 1/2 + 2\beta$. Если же $\alpha + 2\beta < 1/2$, то можно выбрать $\theta < 1$ так, что $\theta - 2\alpha - 4\beta > 0$, и правая часть (7) не превосходит $c_1 J \exp\left[-c_2 J^{-3}2^{2(\lambda-1+\alpha)J}\right]$. Следовательно, чтобы выполнялось неравенство $\lambda - 1 + \alpha > 0$, достаточно потребовать $\lambda > 1 - \alpha$. При таком выборе λ

$$\sum_{J=1}^{\infty} p_J < \infty,$$

и в силу леммы Бореля–Кантелли для любого $\delta > > 0$ событие $\left\{\left|\widehat{R}_2\right| > \delta 2^{\lambda J}\right\}$ осуществляется лишь конечное число раз. Следовательно, $\widehat{R}_2 2^{-\lambda J} \to 0$ п.в.

В $\widehat{R}_1$ при каждом фиксированном j ($0 \leqslant \leqslant j \leqslant [pJ] - 1$) число слагаемых равно 2^j,

а каждое слагаемое не превосходит по модулю $BJ2^{J(1-\alpha)}2^{j(2\beta+\alpha-1)}$, где $B > 0$ — некоторая константа. Следовательно, $\left|\widehat{R}_1\right| \leqslant B_1 J 2^{J(1-\alpha+p(2\beta+\alpha))}$, где B_1 — некоторая положительная константа. Если $2\beta+\alpha \geqslant 1/2$, то при $\lambda > 1/2+2\beta$ всегда можно выбрать такое $0 < p < 1$, что $\lambda-(1-\alpha+p(2\beta+\alpha)) > 0$, и, следовательно, $\widehat{R}_1 2^{-\lambda J} \to 0$ п.в. Если же $\alpha+2\beta < 1/2$, то при $\lambda > 1-\alpha$ всегда можно выбрать такое $0 < p < 1$, что $\lambda-(1-\alpha+p(2\beta+\alpha)) > 0$, и, следовательно, $\widehat{R}_1 2^{-\lambda J} \to 0$ п.в. Теорема доказана.

Литература

1. *Donoho D., Johnstone I. M.* Adapting to unknown smoothness via wavelet shrinkage // J. Amer. Stat. Assoc., 1995. Vol. 90. P. 1200–1224.
2. *Маркин А. В.* Предельное распределение оценки риска при пороговой обработке вейвлет-коэффициентов // Информатика и её применения, 2009. Т. 3. Вып. 4. С. 57–63.
3. *Маркин А. В., Шестаков О. В.* О состоятельности оценки риска при пороговой обработке вейвлет-коэффициентов // Вестн. Моск. ун-та. Сер. 15: Вычисл. матем. и киберн., 2010. № 1. С. 26–34.
4. *Кудрявцев А. А., Шестаков О. В.* Асимптотика оценки риска при вейглет-вейвлет-разложении наблюдаемого сигнала // T-Comm: Телекоммуникации и транспорт, 2011. № 2. С. 54–57.
5. *Шестаков О. В.* Асимптотическая нормальность оценки риска пороговой обработки вейвлет-коэффициентов при выборе адаптивного порога // Докл. РАН, 2012. Т. 445. № 5. С. 513–515.
6. *Шестаков О. В.* О свойствах оценки среднеквадратичного риска при регуляризации обращения линейного однородного оператора с помощью адаптивной пороговой обработки коэффициентов вейглет-вейвлет разложения // Вестн. ТвГУ. Серия: Прикладная математика, 2012. № 8. С. 117–130.
7. *Ерошенко А. А., Шестаков О. В.* Асимптотическая нормальность оценки риска при вейвлет-вейглет-разложении функции сигнала в модели с коррелированным шумом // Вестн. Моск. ун-та. Сер. 15: Вычисл. матем. и киберн., 2014. № 3. С. 110–117.
8. *Ерошенко А. А.* Состоятельность оценок риска при вейвлет-вейглет и вейглет-вейвлет-разложениях функции сигнала в модели с коррелированным шумом // Вестн. ТвГУ. Серия: Прикладная математика, 2015. № 1. С. 103–114.
9. *Ерошенко А. А., Кудрявцев А. А., Шестаков О. В.* Предельное распределение оценки риска метода вейглет-вейвлет-разложения сигнала в модели с коррелированным шумом // Вестн. Моск. ун-та. Сер. 15: Вычисл. матем. и киберн., 2015. № 1. С. 12–18.
10. *Johnstone I. M.* Wavelet shrinkage for correlated data and inverse problems: Adaptivity results // Stat. Sinica, 1999. Vol. 9. No. 1. P. 51–83.
11. *Donoho D.* Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition // Appl. Comput. Harmon. Anal., 1995. Vol. 2. P. 101–126.
12. *Mallat S.* A wavelet tour of signal processing. — New York, NY, USA: Academic Press, 1999. 857 p.
13. *Kolaczyk E. D.* Wavelet methods for the inversion of certain homogeneous linear operators in the presence of noisy data. — Stanford, CA, USA: Stanford University, 1994. PhD Thesis.
14. *Bosq D.* Nonparametric statistics for stochastic processes: Estimation and prediction. — New York, NY, USA: Springer-Verlag, 1996. 169 p.
15. *Bradley R. C.* Basic properties of strong mixing conditions. A survey and some open questions // Probab. Surveys, 2005. Vol. 2. P. 107–144.

Поступила в редакцию 11.11.16

STRONG CONSISTENCY OF THE MEAN SQUARE RISK ESTIMATE IN THE INVERSE STATISTICAL PROBLEMS

O. V. Shestakov[1,2]

[1]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Nonlinear methods of digital signal processing based on thresholding of wavelet coefficients became a popular tool for solving the problems of signal de-noising and compression. This is explained by the fact that the wavelet methods allow much more effective analysis of nonstationary signals than traditional Fourier analysis, thanks to the better adaptation to the functions with varying degrees of regularity. Wavelet thresholding risk analysis is an important practical task, because it allows determining the quality of techniques themselves and the equipment

which is being used. In some applications, the data are observed not directly but after applying a linear transformation. The problem of inverting this transformation is usually set incorrectly, leading to an increase in the noise variance. In this paper, the asymptotic properties of the mean square error (MSE) estimate are studied when inverting linear homogeneous operators by means of wavelet vaguelette decomposition and thresholding. The strong consistency of this estimate has been proved under mild conditions.

Keywords: wavelets; thresholding; MSE risk estimate; correlated noise; asymptotic normality

DOI: 10.14357/19922264170213

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research projects Nos. 15-37-20611 and 16-07-00677).

References

1. Donoho, D., and I. M. Johnstone. 1995. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Stat. Assoc*. 90:1200–1224.
2. Markin, A. V. 2009. Predel'noe raspredelenie otsenki riska pri porogovoy obrabotke veyvlet-koeffitsientov [Limit distribution of risk estimate of wavelet coefficient thresholding]. *Informatika i ee Primeneniya — Inform. Appl.* 3(4):57–63.
3. Markin, A. V., and O. V. Shestakov. 2010. Consistency of risk estimation with thresholding of wavelet coefficients. *Moscow Univ. Comput. Math. Cybern.* 34(1):22–30.
4. Kudryavtsev, A. A., and O. V. Shestakov. 2011. Asimptotika otsenki riska pri veyglet-veyvlet-razlozhenii nablyudaemogo signala [The asymptotic behavior of the risk estimate under wavelet-vaguelette decomposition of the observed signal]. *T-Comm: Telekommunikatsii i Transport* [T-Comm: Telecommunications and Transport] 2:54–57.
5. Shestakov, O. V. 2012. Asymptotic normality of adaptive wavelet thresholding risk estimation. *Dokl. Math.* 86(1):556–558.
6. Shestakov, O. V. 2012. O svoystvakh otsenki srednekvadratichnogo riska pri regulyarizatsii obrashcheniya lineynogo odnorodnogo operatora s pomoshch'yu adaptivnoy porogovoy obrabotki koeffitsientov veyglet-veyvlet razlozheniya [The properties of mean square error estimate when regularizing the inversion of the homogeneous linear operator using adaptive thresholding of wavelet-vaguelette decomposition coefficients]. *Vestn. TvGU. Seriya: Prikladnaya matematika* [Herald of Tver State University. Series: Applied Mathematics] 8:117–130.
7. Eroshenko, A. A., and O. V. Shestakov. 2014. Asymptotic normality of estimating risk upon the wavelet-vaguelette decomposition of a signal function in a model with correlated noise. *Moscow Univ. Comput. Math. Cybern.* 38(3):110–117.
8. Eroshenko, A. A. 2015. Sostoyatel'nost' otsenok riska pri veyvlet-veyglet i veyglet-veyvlet-razlozheniyakh funktsii signala v modeli s korrelirovannym shumom [Consistency of risk estimates for wavelet-vaguelette and vaguelette-wavelet decompositions of signal function in the model of data with correlated noise]. *Vestn. TvGU. Seriya: Prikladnaya matematika* [Herald of Tver State University. Series: Applied Mathematics] 1:103–114.
9. Eroshenko, A. A., A. A. Kudryavtsev, and O. V. Shestakov. 2015. Limit distribution of a risk estimate using the vaguelette-wavelet decomposition of signals in a model with correlated noise. *Moscow Univ. Comput. Math. Cybern.* 39(1):6-13.
10. Johnstone, I. M. 1999. Wavelet shrinkage for correlated data and inverse problems: Adaptivity results. *Stat. Sinica* 9(1):51–83.
11. Donoho, D. 1995. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.* 2:101–126.
12. Mallat, S. 1999. *A wavelet tour of signal processing.* New York, NY: Academic Press. 857 p.
13. Kolaczyk, E. D. 1994. Wavelet methods for the inversion of certain homogeneous linear operators in the presence of noisy data. Stanford, CA: Stanford University. PhD Thesis. 163 p.
14. Bosq, D. 1996. *Nonparametric statistics for stochastic processes: Estimation and prediction.* New York, NY: Springer-Verlag. 169 p.
15. Bradley, R. C. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surveys* 2:107–144.

Received November 11, 2016

Contributor

Shestakov Oleg V. (b. 1976) — Doctor of Science in physics and mathematics, associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; oshestakov@cs.msu.su

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 2. С. 122–125

УНИВЕРСАЛЬНАЯ ПОРОГОВАЯ ОБРАБОТКА В МОДЕЛЯХ С НЕГАУССОВЫМ ШУМОМ

О. В. Шестаков[1]

Аннотация: В задачах непараметрического оценивания сигнала обычно предполагается, что функция сигнала принадлежит некоторому специальному классу. Например, она может быть кусочно-непрерывной или кусочно-дифференцируемой и иметь компактный носитель. Эти предположения, как правило, позволяют экономно представить функцию сигнала в некотором специально подобранном базисе таким образом, что полезный сигнал оказывается сосредоточенным в относительно небольшом количестве больших по абсолютному значению коэффициентов разложения. Затем осуществляется пороговая обработка с целью удаления шумовых коэффициентов. Обычно распределение шума предполагается гауссовым. Эта модель хорошо изучена в литературе, и для разных классов функций сигналов вычислены оптимальные параметры пороговой обработки. В данной работе рассматривается задача построения оценки функции сигнала по наблюдениям, содержащим аддитивный шум, распределение которого принадлежит достаточно широкому классу. Вычисляются значения универсальных параметров пороговой обработки, при которых среднеквадратичный риск близок к минимальному.

Ключевые слова: пороговая обработка; негауссовый шум; среднеквадратичный риск

DOI: 10.14357/19922264170214

1 Введение

Современные методы построения оценок функции сигнала по зашумленным наблюдениям часто основаны на разложении этой функции по базису, обеспечивающему «экономное» представление данных, т. е. коэффициенты такого разложения убывают достаточно быстро (примерами подобных базисов могут служить различные классы вейвлетов). Затем происходит обнуление части коэффициентов, которые по предположению содержат в основном шум. В предположении о гауссовском распределении шума эти методы хорошо разработаны [1–3]. В данной работе рассматривается модель с аддитивным шумом, который необязательно имеет гауссово распределение, и вычисляются универсальные параметры диагональных методов подавления шума, при которых среднеквадратичный риск близок к минимальному.

2 Модель данных и методы подавления шума

Предположим, что данные имеют вид:

$$X_i = f_i + z_i\,,\ \ i = 1, \ldots, N\,,$$

где f_i — «чистый» сигнал, а z_i — «шумовые» коэффициенты, относительно которых предполагается, что они независимы и имеют распределение с симметричной дифференцируемой плотностью $h(x)$. Также предположим, что $\sup\limits_{x\in\mathbf{R}} |h'(x)| < A$ с некоторой константой $A > 0$ и что

$$h(x) \asymp x^{\alpha} e^{-\theta x^{\beta}} \text{ при } x \to \infty\,,\ \ \alpha \in \mathbf{R}\,,\ \theta > 0\,,\ \beta > 0\,.$$

Дисперсию z_i обозначим через σ^2. Класс распределений такого вида достаточно широк. Распределения из этого класса могут иметь как более легкие, так и более тяжелые хвосты, чем гауссово распределение.

При построении оценки сигнала будем рассматривать только диагональные методы, т. е. когда для получения оценки $\hat{f}_i$ коэффициента f_i используется только величина X_i. Определим среднеквадратичный риск оценки:

$$R(\hat{f}) = \sum_{i=1}^{N} \mathsf{E}\left(\hat{f}_i - f_i\right)^2\,. \tag{1}$$

Рассмотрим метод построения оценки, который заключается в том, что каждый коэффициент либо обнуляется, либо остается неизменным:

$$\hat{f}_i = \rho_\delta\left(X_i\right) = \delta_i X_i\,,\ \ \delta_i \in \{0, 1\}\,,\ i = 1, \ldots, N\,.$$

[1]Московский государственный университет им. М. В. Ломоносова, кафедра математической статистики факультета вычислительной математики и кибернетики; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, oshestakov@cs.msu.su

Предположим, что известны «идеальные» параметры δ_i, которые минимизируют риск (1). Поскольку слагаемые в (1) равны f_i^2, если $\delta_i = 0$, и σ^2, если $\delta_i = 1$, минимальный среднеквадратичный риск равен

$$R_{\mathrm{Min}}(\hat{f}) = \sum_{i=1}^{N} \min\left(f_i^2, \sigma^2\right), \qquad (2)$$

а «идеальные» параметры равны $\delta_i = \mathbf{1}(|f_i| > \sigma)$. На практике вычислить эти параметры нельзя и невозможно построить оценку, риск которой равен (2). Однако в работе [1] показано, что при использовании процедуры пороговой обработки в модели с гауссовым можно обеспечить порядок среднеквадратичного риска, который близок к (2) с точностью до логарифмического множителя.

Пороговая обработка является одним из самых популярных методов подавления шума. Ее смысл заключается в обнулении коэффициентов, чьи абсолютные значения не превышают заданного порога. Оценка $\hat{f}_i$ вычисляется с помощью пороговой функции $\rho_T(x)$ с порогом T. Наиболее популярными являются функция жесткой пороговой обработки $\rho_T^{(h)}(x) = x \cdot \mathbf{1}(|x| > T)$ и мягкой пороговой обработки $\rho_T^{(s)}(x) = \mathrm{sign}(x)(|x| - T)_+$. Среднеквадратичный риск пороговой обработки обозначим через $R_T(\hat{f})$.

3 Универсальный порог

Одной из основных проблем при пороговой обработке является стратегия выбора порога. В работе [1] предложен так называемый универсальный порог для модели с гауссовским шумом. Этот порог является в некотором смысле максимальным среди «разумных» порогов, и среднеквадратичный риск при таком пороге близок к (2). Более точные значения порога для различных функций потерь при дополнительных условиях на гладкость функции сигнала получены в работах [3–6]. В данной работе предлагается аналог универсального порога для определенной выше более общей модели шума и показывается, что он обладает практически такими же свойствами, как в модели с гауссовским шумом.

Пусть $T_U = \left(\theta^{-1} \ln N\right)^{1/\beta}$. По аналогии с гауссовской моделью шума назовем этот порог универсальным.

Теорема 1. *Существует такая константа $C > 0$, зависящая только от $h(x)$, что начиная с некоторого N при жесткой пороговой обработке*

$$R_{T_U}\left(\hat{f}\right) \le C(\ln N)^{\delta(\alpha,\beta)} \left(\sigma^2 + R_{\mathrm{Min}}\left(\hat{f}\right)\right), \qquad (3)$$

где $\delta(\alpha, \beta) = \max(2/\beta, (3 + \alpha - \beta)/\beta)$.

Доказательство. Рассмотрим $\mathsf{E}(\rho_{T_U}^{(h)}(X_i) - f_i)^2$. Пусть $|f_i| > T_U$. Тогда

$$\mathsf{E}\left(\rho_{T_U}^{(h)}(X_i) - f_i\right)^2 = \sigma^2 - \\ - \mathsf{E}\left(X_i - f_i\right)^2 \mathbf{1}\left(|X_i| \le T_U\right) + f_i^2 \mathsf{E}\mathbf{1}\left(|X_i| \le T_U\right) \le \\ \le \sigma^2 + f_i^2 \mathsf{E}\mathbf{1}\left(|X_i| \le T_U\right).$$

В силу конечности второго момента плотности $h(x)$ существует такая положительная константа $C^{(h)}$, зависящая от $h(x)$, что $f_i^2 \mathsf{E}\mathbf{1}(|X_i| \le T) \le C^{(h)} T^2$ при любом $T > 0$. Следовательно, существует такая константа $C_1 > 0$, что

$$\mathsf{E}\left(\rho_{T_U}^{(h)}(X_i) - f_i\right)^2 \le \\ \le \sigma^2 + C^{(h)} T_U^2 \le C_1 \left(\frac{\sigma^2}{N} + \sigma^2\right) (\ln N)^{2/\beta}.$$

Пусть теперь $|f_i| \le T_U$. Тогда

$$\mathsf{E}\left(\rho_{T_U}^{(h)}(X_i) - f_i\right)^2 \le \mathsf{E}\left(X_i - f_i\right)^2 \mathbf{1}\left(|X_i| > T_U\right) + f_i^2.$$

Обозначим

$$g(f_i) = \mathsf{E}\left(X_i - f_i\right)^2 \mathbf{1}\left(|X_i| > T_U\right).$$

Поскольку $g(f_i)$ симметрична относительно 0 и $|h'(x)|$ ограничена,

$$g(f_i) \le g(0) + \frac{1}{2}\left(\sup_{f \in \mathbf{R}} |g''(f)|\right) f_i^2.$$

В силу определения $h(x)$ существует такая константа $C_0 > 0$, что

$$g(0) \le C_0 T_U^{3+\alpha-\beta} e^{-\theta T_U^\beta}.$$

Следовательно, существует такая константа $C_2 > 0$, что

$$\mathsf{E}\left(\rho_{T_U}^{(h)}(X_i) - f_i\right)^2 \le C_2 \left(\frac{\sigma^2 (\ln N)^{(3+\alpha-\beta)/\beta}}{N} + f_i^2\right).$$

Таким образом, начиная с некоторого N

$$\mathsf{E}\left(\rho_{T_U}^{(h)}(X_i) - f_i\right)^2 \le \\ \le C(\ln N)^{\delta(\alpha,\beta)} \left(\frac{\sigma^2}{N} + \min(f_i^2, \sigma^2)\right),$$

где C зависит только от $h(x)$. Суммируя по i, получаем (3). Теорема доказана.

Теорема 2. *Существует такая константа $C > 0$, зависящая только от $h(x)$, что начиная с некоторого N при мягкой пороговой обработке*

$$R_{T_U}\left(\hat{f}\right) \le C(\ln N)^{\delta(\alpha,\beta)}\left(\sigma^2 + R_{\mathrm{Min}}\left(\hat{f}\right)\right), \quad (4)$$

где $\delta(\alpha,\beta) = \max(2/\beta, (3+\alpha-3\beta)/\beta)$.

Доказательство этой теоремы аналогично доказательству теоремы 1. Основное отличие заключается в том, что при мягкой пороговой обработке в силу определения $h(x)$ выполнено

$$g(0) = \mathsf{E}\left(X_i - T_U\right)^2 \mathbf{1}\left(X_i > T_U\right) + \\ + \mathsf{E}\left(X_i + T_U\right)^2 \mathbf{1}\left(X_i < -T_U\right) \le C_0 T_U^{3+\alpha-3\beta} e^{-\theta T_U^{\beta}}$$

с некоторой константой $C_0 > 0$.

Таким образом, оценки (3) и (4) демонстрируют, что в рассматриваемой модели универсальный порог обеспечивает порядок среднеквадратичного риска, который отличается от минимального лишь наличием логарифмического множителя в степени, зависящей от характеристик распределения шума.

Установим теперь еще одно свойство порога T_U, которое показывает, что, как и в случае гауссовского шума, универсальный порог является в некотором смысле максимальным среди разумных порогов.

Теорема 3. *Пусть случайные величины z_i, $i = 1, \ldots, N$, независимы и имеют плотность распределения $h(x)$, удовлетворяющую перечисленным выше условиям. Тогда*

$$\mathsf{P}\left(T_U^- \le \max_{1\le i\le N} |z_i| \le T_U^+\right) \to 1 \text{ при } N \to \infty. \quad (5)$$

Здесь

$$T_U^- = \left(\frac{\ln N + \gamma' \ln\ln N}{\theta}\right)^{1/\beta},$$

где γ' — произвольное число, удовлетворяющее $\gamma' < \beta^{-1}(1+\alpha-\beta)$;

$$T_U^+ = \begin{cases} \left(\dfrac{\ln N + \gamma'' \ln\ln N}{\theta}\right)^{1/\beta} & \text{при } 1+\alpha-\beta \ge 0; \\ T_U & \text{при } 1+\alpha-\beta < 0, \end{cases}$$

где γ'' — произвольное число, удовлетворяющее $\gamma'' > \beta^{-1}(1+\alpha-\beta)$.

Доказательство. Утверждение (5) является простым следствием более общих утверждений о распределениях экстремумов случайных последовательностей. Легко видеть, что $NH(T_U^-) \to \infty$ и $NH(T_U^+) \to 0$ при $N \to \infty$, где $H(x)$ — функция распределения, соответствующая плотности $h(x)$. Следовательно, в силу теоремы 1.5.1 из [7] при $N \to \infty$

$$\mathsf{P}\left(\max_{1\le i\le N} |z_i| \le T_U^-\right) \to 0;$$
$$\mathsf{P}\left(\max_{1\le i\le N} |z_i| \le T_U^+\right) \to 1,$$

т. е. выполнено (5). Теорема доказана.

Утверждение теоремы 3 означает, что максимальная амплитуда шума с вероятностью, стремящейся к единице, находится в некоторой окрестности T_U. Следовательно, при достаточно большом N нет смысла выбирать порог, превосходящий T_U.

Литература

1. *Donoho D., Johnstone I. M.* Ideal spatial adaptation via wavelet shrinkage // Biometrika, 1994. Vol. 81. No. 3. P. 425–455.
2. *Donoho D., Johnstone I. M.* Minimax estimation via wavelet shrinkage // Ann. Stat., 1998. Vol. 26. No. 3. P. 879–921.
3. *Jansen M.* Noise reduction by wavelet thresholding. — Lecture notes in statistics ser. — New York, NY, USA: Springer Verlag, 2001. Vol. 161. 217 p.
4. *Шестаков О. В.* Асимптотическая нормальность оценки риска пороговой обработки вейвлет-коэффициентов при выборе адаптивного порога // Докл. РАН, 2012. Т. 445. № 5. С. 513–515.
5. *Кудрявцев А. А., Шестаков О. В.* Асимптотическое поведение порога, минимизирующего усредненную вероятность ошибки вычисления вейвлет-коэффициентов // Докл. РАН, 2016. Т. 468. № 5. С. 487–491.
6. *Кудрявцев А. А., Шестаков О. В.* Асимптотически оптимальная пороговая обработка вейвлет-коэффициентов в моделях с негауссовым распределением шума // Докл. РАН, 2016. Т. 471. № 1. С. 11–15.
7. *Лидбеттер М., Линдгрен Г., Ротсен Х.* Экстремумы случайных последовательностей и процессов / Пер с англ. — М.: Мир, 1989. 392 с. (*Leadbetter M., Lindgren G., Rootzen H.* Extremes and related properties of random sequences and processes. — New York, NY, USA: Springer-Verlag, 1983. 336 p.)

Поступила в редакцию 01.03.17

UNIVERSAL THRESHOLDING IN THE MODELS WITH NON-GAUSSIAN NOISE

O. V. Shestakov[1,2]

[1]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: A common assumption in nonparametric signal estimation is that the signal function belongs to a certain class. For example, it may be piecewise continuous or piecewise differentiable and have a compact support. These assumptions, as a rule, make it possible to economically represent a signal function in a specially selected basis in such a way that the useful signal is concentrated in a relatively small number of large expansion coefficients. Then, threshold processing removes noisy coefficients. Typically, the noise distribution is assumed to be Gaussian. This model has been well studied in the literature and optimal thresholding parameters have been calculated for different classes of signal functions. The paper considers the problem of constructing an estimate for the signal function from the observations containing additive noise, whose distribution belongs to quite a wide class. The authors calculate the values of universal thresholding parameters for which the mean-square risk is close to the minimum.

Keywords: thresholding; non-Gaussian noise; mean-square risk

DOI: 10.14357/19922264170214

References

1. Donoho, D., and I. M. Johnstone. 1994. Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81(3):425–455.
2. Donoho, D., and I. M. Johnstone. 1998. Minimax estimation via wavelet shrinkage *Ann. Stat.* 26(3):879–921.
3. Jansen, M. 2001. *Noise reduction by wavelet thresholding.* Lecture notes in statistics ser. New York, NY: Springer Verlag. Vol. 161. 217 p.
4. Shestakov, O. V. 2012. Asymptotic normality of adaptive wavelet thresholding risk estimation. *Dokl. Math.* 86(1):556–558.
5. Kudryavtsev, A. A., and O. V. Shestakov. 2016. Asymptotic behavior of the threshold minimizing the average probability of error in calculation of wavelet coefficients. *Dokl. Math.* 93(3):295–299.
6. Kudryavtsev, A. A., and O. V. Shestakov. 2016. Asymptotically optimal wavelet thresholding in the models with non-Gaussian noise distributions. *Dokl. Math.* 94(3):615–619.
7. Leadbetter, M., G. Lindgren, and H. Rootzen. 1983. *Extremes and related properties of random sequences and processes.* New York, NY: Springer-Verlag. 336 p.

Received March 1, 2017

Contributor

Shestakov Oleg V. (b. 1976) — Doctor of Science in physics and mathematics, associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; oshestakov@cs.msu.su

ОБ АВТОРАХ

Агаларов Явер Мирзабекович (р. 1952) — кандидат технических наук, доцент, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Атаева Ольга Муратовна (р. 1978) — младший научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Борисов Андрей Владимирович (р. 1965) — доктор физико-математических наук, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Васильев Николай Семенович (р. 1952) — доктор физико-математических наук, профессор Московского государственного технического университета им. Н. Э. Баумана

Гайдамака Юлия Васильевна (р. 1971) — кандидат физико-математических наук, доцент Российского университета дружбы народов, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Кабанов Юрий Михайлович (р. 1948) — профессор Лаборатории математики Университета Франш-Конте, г. Безансон, Франция; ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; преподаватель Национального исследовательского университета «МЭИ»

Корепанов Эдуард Рудольфович (р. 1966) — кандидат технических наук, заведующий отделом Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Кривенко Михаил Петрович (р. 1946) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Лукашенко Олег Викторович (р. 1986) — кандидат физико-математических наук, научный сотрудник Института прикладных математических исследований Карельского научного центра Российской академии наук; преподаватель Петрозаводского государственного университета

Мокбель Рита (р. 1981) — аспирант Лаборатории математики Университета Франш-Конте, г. Безансон, Франция

Молчанов Дмитрий Александрович (р. 1978) — кандидат технических наук, доцент Российского университета дружбы народов

Морозов Евсей Викторович (р. 1947) — доктор физико-математических наук, профессор, ведущий научный сотрудник Института прикладных математических исследований Карельского научного центра Российской академии наук; профессор Петрозаводского государственного университета

Орлов Юрий Николаевич (р. 1963) — доктор физико-математических наук, профессор, руководитель сектора Института прикладной математики им. М. В. Келдыша Российской академии наук

Пагано Микеле (р. 1968) — доктор наук (PhD) по электронике, доцент Университета г. Пиза, Италия

Пархоменко Валерий Павлович (р. 1951) — кандидат физико-математических наук, заведующий сектором Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Рудой Георгий Игоревич (р. 1991) — аспирант Московского физико-технического института

Самуйлов Андрей Константинович (р. 1988) — кандидат физико-математических наук, доцент Российского университета дружбы народов

Серебряков Владимир Алексеевич (р. 1946) — доктор физико-математических наук, профессор, заведующий отделом Вычислительного центра им.

А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Синицын Владимир Игоревич (р. 1968) — доктор физико-математических наук, доцент, заведующий отделом Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Синицын Игорь Николаевич (р. 1940) — доктор технических наук, профессор, заслуженный деятель науки РФ, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ушаков Владимир Георгиевич (р. 1952) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ушаков Николай Георгиевич (р. 1954) — доктор физико-математических наук, ведущий научный сотрудник Института проблем технологии микроэлектроники и особо чистых материалов Российской академии наук (Черноголовка); профессор Норвежского научно-технологического университета (г. Тронхейм)

Шестаков Олег Владимирович (р. 1976) — доктор физико-математических, доцент кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Эль-Битар Халил (р. 1981) — аспирант Лаборатории математики Университета Франш-Конте, г. Безансон, Франция

Правила подготовки рукописей для публикации в журнале «Информатика и её применения»

Журнал «Информатика и её применения» публикует теоретические, обзорные и дискуссионные статьи, посвященные научным исследованиям и разработкам в области информатики и ее приложений.

Журнал издается на русском языке. По специальному решению редколлегии отдельные статьи могут печататься на английском языке.

Тематика журнала охватывает следующие направления:

- теоретические основы информатики;
- математические методы исследования сложных систем и процессов;
- информационные системы и сети;
- информационные технологии;
- архитектура и программное обеспечение вычислительных комплексов и сетей.

1. В журнале печатаются статьи, содержащие результаты, ранее не опубликованные и не предназначенные к одновременной публикации в других изданиях.

   Публикация предоставленной автором(ами) рукописи не должна нарушать положений глав 69, 70 раздела VII части IV Гражданского кодекса, которые определяют права на результаты интеллектуальной деятельности и средства индивидуализации, в том числе авторские права, в РФ.

   Ответственность за нарушение авторских прав, в случае предъявления претензий к редакции журнала, несут авторы статей.

   Направляя рукопись в редакцию, авторы сохраняют свои права на данную рукописи и при этом передают учредителям и редколлегии журнала неисключительные права на издание статьи на русском языке (или на языке статьи, если он отличен от русского) и на перевод ее на английский язык, а также на ее распространение в России и за рубежом. Каждый автор должен представить в редакцию подписанный с его стороны «Лицензионный договор о передаче неисключительных прав на использование произведения», текст которого размещен по адресу http://www.ipiran.ru/publications/licence.doc. Этот договор может быть представлен в бумажном (в 2-х экз.) или в электронном виде (отсканированная копия заполненного и подписанного документа).

   Редколлегия вправе запросить у авторов экспертное заключение о возможности публикации представленной статьи в открытой печати.

2. К статье прилагаются данные автора (авторов) (см. п. 8). При наличии нескольких авторов указывается фамилия автора, ответственного за переписку с редакцией.

3. Редакция журнала осуществляет экспертизу присланных статей в соответствии с принятой в журнале процедурой рецензирования.

   Возвращение рукописи на доработку не означает ее принятия к печати.

   Доработанный вариант с ответом на замечания рецензента необходимо прислать в редакцию.

4. Решение редколлегии о публикации статьи или ее отклонении сообщается авторам.

   Редколлегия может также направить авторам текст рецензии на их статью. Дискуссия по поводу отклоненных статей не ведется.

5. Редактура статей высылается авторам для просмотра. Замечания к редактуре должны быть присланы авторами в кратчайшие сроки.

6. Рукопись предоставляется в электронном виде в форматах MS WORD (.doc или .docx) или LaTeX (.tex), дополнительно — в формате .pdf, на дискете, лазерном диске или электронной почтой. Предоставление бумажной рукописи необязательно.

7. При подготовке рукописи в MS Word рекомендуется использовать следующие настройки.

   Параметры страницы: формат — A4; ориентация — книжная; поля (см): внутри — 2,5, снаружи — 1,5, сверху — 2, снизу — 2, от края до нижнего колонтитула — 1,3.

   Основной текст: стиль — «Обычный», шрифт — Times New Roman, размер — 14 пунктов, абзацный отступ — 0,5 см, 1,5 интервала, выравнивание — по ширине.

Рекомендуемый объем рукописи — не свыше 15 страниц указанного формата. При превышении указанного объема редколлегия вправе потребовать от автора сокращения объема рукописи.

Сокращения слов, помимо стандартных, не допускаются. Допускается минимальное количество аббревиатур.

Все страницы рукописи нумеруются.

Шаблоны примеров оформления представлены в Интернете: http://www.ipiran.ru/journal/template.doc

8. Статья должна содержать следующую информацию на ***русском и английском языках***:
   - название статьи;
   - Ф.И.О. авторов, на английском можно только имя и фамилию;
   - место работы, с указанием почтового адреса организации и электронного адреса каждого автора;
   - сведения об авторах, в соответствии с форматом, образцы которого представлены на страницах: http://www.ipiran.ru/journal/issues/2013_07_01_rus/authors.asp и http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;
   - аннотация (не менее 100 слов на каждом из языков). Аннотация — это краткое резюме работы, которое может публиковаться отдельно. Она является основным источником информации в информационных системах и базах данных. Английская аннотация должна быть оригинальной, может не быть дословным переводом русского текста и должна быть написана хорошим английским языком. В аннотации не должно быть ссылок на литературу и, по возможности, формул;
   - ключевые слова — желательно из принятых в мировой научно-технической литературе тематических тезаурусов. Предложения не могут быть ключевыми словами;
   - источники финансирования работы (ссылки на гранты, проекты, поддерживающие организации и т. п.).

9. Требования к спискам литературы.

   Ссылки на литературу в тексте статьи нумеруются (в квадратных скобках) и располагаются в каждом из списков литературы в порядке первых упоминаний.

   Списки литературы представляются в двух вариантах:

   (1) **Список литературы к русскоязычной части**. Русские и английские работы — на языке и в алфавите оригинала;

   (2) **References**. Русские работы и работы на других языках — в латинской транслитерации с переводом на английский язык; английские работы и работы на других языках — на языке оригинала.

   Необходимо для составления списка “References” пользоваться размещенной на сайте http://www.translit.net/ru/bgn/ бесплатной программой транслитерации русского текста в латиницу.

   Список литературы “References” приводится полностью отдельным блоком, повторяя все позиции из списка литературы к русскоязычной части, независимо от того, имеются или нет в нем иностранные источники. Если в списке литературы к русскоязычной части есть ссылки на иностранные публикации, набранные латиницей, они полностью повторяются в списке “References”.

   Ниже приведены примеры ссылок на различные виды публикаций в списке “References”.

   Описание статьи из журнала:

   Zagurenko, A. G., V. A. Korotovskikh, A. A. Kolesnikov, A. V. Timonov, and D. V. Kardymon. 2008. Tekhniko-ekonomicheskaya optimizatsiya dizayna gidrorazryva plasta [Technical and economic optimization of the design of hydraulic fracturing]. *Neftyanoe hozyaystvo* [*Oil Industry*] 11:54–57.

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Rus. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Описание статьи из электронного журнала:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Описание статьи из продолжающегося издания (сборника трудов):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental’noe issledovanie prochnosti soedineniy “stal’–kompozit” [Experimental study of the strength of joints “steel–composite”]. *Trudy MGTU “Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem”* [*Bauman MSTU “Mathematical Modeling of Complex Technical Systems” Proceedings*]. 593:125–130.

Описание материалов конференций:

Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

Описание книги (монографии, сборники):

Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

Latyshev, V. N. 2009. *Tribologiya rezaniya. Kn. 1: Friktsionnye protsessy pri rezanii metallov* [*Tribology of cutting. Vol. 1: Frictional processes in metal cutting*]. Ivanovo: Ivanovskii State Univ. 108 p.

Описание переводной книги (в списке литературы к русскоязычной части необходимо указать: / Пер. с англ. — после названия книги, а в конце ссылки указать оригинал книги в круглых скобках):

1. В русскоязычной части:
   Тимошенко С. П., Янг Д. Х., Уивер У. Колебания в инженерном деле / Пер. с англ. — М.: Машиностроение, 1985. 472 с. (*Timoshenko S. P., Young D. H., Weaver W.* Vibration problems in engineering. — 4th ed. — N.Y.: Wiley, 1974. 521 p.)
2. В англоязычной части:
   Timoshenko, S. P., D. H. Young, and W. Weaver. 1974. *Vibration problems in engineering*. 4th ed. N.Y.: Wiley. 521 p.

Описание неопубликованного документа:

Latypov, A. R., M. M. Khasanov, and V. A. Baikov. 2004. Geology and production (NGT GiD). Certificate on official registration of the computer program No. 2004611198. (In Russian, unpubl.)

Описание интернет-ресурса:

Pravila tsitirovaniya istochnikov [Rules for the citing of sources]. Available at: http://www.scribd.com/doc/1034528/ (accessed February 7, 2011).

Описание диссертации или автореферата диссертации:

Semenov, V. I. 2003. Matematicheskoe modelirovanie plazmy v sisteme kompaktnyy tor [Mathematical modeling of the plasma in the compact torus]. D.Sc. Diss. Moscow. 272 p.

Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. PhD Thesis. Moscow: IPI RAN. 23 p.

Описание ГОСТа:

GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. Moscow: Standardinform Publs. 10 p.

Описание патента:

Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

10. Присланные в редакцию материалы авторам не возвращаются.
11. При отправке файлов по электронной почте просим придерживаться следующих правил:
    - указывать в поле subject (тема) название журнала и фамилию автора;
    - использовать attach (присоединение);
    - в состав электронной версии статьи должны входить: файл, содержащий текст статьи, и файл(ы), содержащий(е) иллюстрации.
12. Журнал «Информатика и её применения» является некоммерческим изданием. Плата за публикацию не взимается, гонорар авторам не выплачивается.

Адрес редакции журнала «Информатика и её применения»:
Москва 119333, ул. Вавилова, д. 44, корп. 2, ФИЦ ИУ РАН
Тел.: +7 (499) 135-86-92 Факс: +7 (495) 930-45-05
e-mail: rust@ipiran.ru (Сейфуль-Мулюков Рустем Бадриевич)
http://www.ipiran.ru/journal/issues/

Requirements for manuscripts submitted to Journal "Informatics and Applications"

Journal "Informatics and Applications" (Inform. Appl.) publishes theoretical, review, and discussion articles on the research and development in the field of informatics and its applications.

The journal is published in Russian. By a special decision of the editorial board, some articles can be published in English.

The topics covered include the following areas:

- theoretical fundamentals of informatics;
- mathematical methods for studying complex systems and processes;
- information systems and networks;
- information technologies; and
- architecture and software of computational complexes and networks.

1. The Journal publishes original articles which have not been published before and are not intended for simultaneous publication in other editions. An article submitted to the Journal must not violate the Copyright law. Sending the manuscript to the Editorial Board, the authors retain all rights of the owners of the manuscript and transfer the nonexclusive rights to publish the article in Russian (or the language of the article, if not Russian) and its distribution in Russia and abroad to the Founders and the Editorial Board. Authors should submit a letter to the Editorial Board in the following form:

   Agreement on the transfer of rights to publish:

   "*We, the undersigned authors of the manuscript ". . . ", pass to the Founder and the Editorial Board of the Journal "Informatics and Applications" the nonexclusive right to publish the manuscript of the article in Russian (or in English) in both print and electronic versions of the Journal. We affirm that this publication does not violate the Copyright of other persons or organizations.*

   Author(s) signature(s): (name(s), address(es), date).

   This agreement should be submitted in paper form or in the form of a scanned copy (signed by the authors).

2. A submitted article should be attached with **the data on the author(s)** (see item 8). If there are several authors, the contact person should be indicated who is responsible for correspondence with the Editorial Board and other authors about revisions and final approval of the proofs.
3. The Editorial Board of the Journal examines the article according to the established reviewing procedure. If the authors receive their article for correction after reviewing, it does not mean that the article is approved for publication. The corrected article should be sent to the Editorial Board for the subsequent review and approval.
4. The decision on the article publication or its rejection is communicated to the authors. The Editorial Board may also send the reviews on the submitted articles to the authors. Any discussion upon the rejected articles is not possible.
5. The edited articles will be sent to the authors for proofread. The comments of the authors to the edited text of the article should be sent to the Editorial Board as soon as possible.
6. The manuscript of the article should be presented electronically in the MS WORD (.doc or .docx) or LaTeX (.tex) formats, and additionally in the .pdf format. All documents may be sent by e-mail or provided on a CD or diskette. A hard copy submission is not necessary.
7. The recommended typesetting instructions for manuscript.

   Pages parameters: format A4, portrait orientation, document margins (cm): left — 2.5, right — 1.5, above — 2.0, below — 2.0, footer 1.3.

   Text: font —Times New Roman, font size — 14, paragraph indent — 0.5, line spacing — 1.5, justified alignment.

   The recommended manuscript size: not more than 15 pages of the specified format. If the specified size exceeded, the editorial board is entitled to require the author to reduce the manuscript.

   Use only standard abbreviations. Avoid abbreviations in the title and abstract. The full term for which an abbreviation stands should precede its first use in the text unless it is a standard unit of measurement.

   All pages of the manuscript should be numbered.

   The templates for the manuscript typesetting are presented on site: http://www.ipiran.ru/journal/template.doc.
8. The articles should enclose data both in **Russian and English**:

   - title;
   - author's name and surname;
   - affiliation — organization, its address with ZIP code, city, country, and official e-mail address;
   - data on authors according to the format: (see site)
     http://www.ipiran.ru/journal/issues/2013_07_01/authors.asp and
     http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;

- abstract (not less than 100 words) both in Russian and in English. Abstract is a short summary of the article that can be published separately. The abstract is the main source of information on the article and it could be included in leading information systems and data bases. The abstract in English has to be an original text and should not be an exact translation of the Russian one. Good English is required. In abstracts, avoid references and formulae;
- indexing is performed on the basis of keywords. The use of keywords from the internationally accepted thematic Thesauri is recommended.

  Important! Keywords must not be sentences;
- Acknowledgments.

9. References. Russian references have to be presented both in English translation and Latin transliteration (refer http://www.translit.net/ru/bgn/).

   Please take into account the following examples of Russian references appearance:

   Article in journal:

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Rus. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Journal article in electronic format:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Article from the continuing publication (collection of works, proceedings):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental'noe issledovanie prochnosti soedineniy "stal'–kompozit" [Experimental study of the strength of joints "steel–composite"]. *Trudy MGTU "Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem"* [*Bauman MSTU "Mathematical Modeling of Complex Technical Systems" Proceedings*]. 593:125–130.

   Conference proceedings:

   Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

   Books and other monographs:

   Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

   Dissertation and Thesis:

   Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. PhD Thesis. Moscow: IPI RAN. 23 p.

   State standards and patents:

   GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. M.: Standardinform Publs. 10 p.

   Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

   References in Latin transcription are presented in the original language.

   References in the text are numbered according to the order of their first appearance; the number is placed in square brackets. All items from the reference list should be cited.

10. Manuscripts and additional materials are not returned to Authors by the Editorial Board.
11. Submissions of files by e-mail must include:
    - the journal title and author's name in the "Subject" field;
    - an article and additional materials have to be attached using the "attach" function;
    - an electronic version of the article should contain the file with the text and a separate file with figures.
12. "Informatics and Applications" journal is not a profit publication. There are no charges for the authors as well as there are no royalties.

Editorial Board address:
FRC CSC RAS, 44, block 2, Vavilov Str., Moscow 119333, Russia
Ph.: +7 (499) 135 86 92, Fax: +7 (495) 930 45 05
e-mail: rust@ipiran.ru (to Prof. Rustem Seyful-Mulyukov)
http://www.ipiran.ru/english/journal.asp