ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 7, pp. 666–682.
© The Author(s), 2025 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2025.
Russian Text © The Author(s), 2025, published in Avtomatika i Telemekhanika, 2025, No. 7, pp. 90–111.

Randomized Machine Learning Methods for Generating Random Data Ensembles with Given Numerical Characteristics

Yu. S. Popkov[*,**,a], **A. Yu. Popkov**[*,b], **and Yu. A. Dubnov**[*,c]

[*] *Federal Research Center "Computer Science and Control," Russian Academy of Sciences, Moscow, Russia*
[**] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a] *popkov@isa.ru,* [b] *apopkov@isa.ru,* [c] *yury.dubnov@phystech.edu*

Received March 25, 2024
Revised February 16, 2025
Accepted May 29, 2025

Abstract—This paper considers the problem of generating random data ensembles with given numerical characteristics. A solution method is developed using randomized machine learning procedures based on a sequence of functional entropy-linear programming problems with constraints in the form of normalized moments. The generation problem is reduced to a system of nonlinear equations with integral components. The authors' asymptotic analytical method is adapted to transform these equations into a system of equations with a polynomial left-hand side. The analytical methods are applied to generate random data ensembles for asset price dynamics.

Keywords: entropy, randomization, normalized moments, random ensembles, machine learning, power series, analytic functions, multidimensional integrals, multidimensional polynomials

DOI: 10.31857/S0005117925070069

1. INTRODUCTION

Machine learning methods for many applied and scientific-practical problems are based on data with required properties. In the context of machine learning methods of the statistical approach, data requirements are implemented in the form of their probabilistic and numerical characteristics [1]. In addition to traditional classification and forecasting [2–7], we mention other areas such as software testing [8–10] and knowledge control [11–13], where data serve either to train process models or evaluate hypotheses statistically. In parallel, the so-called scenario approach is being developed: parameter arrays are compiled for a parameterized process model (most often, by experts), and corresponding data ensembles are generated [14, 15]. In any case, data properties must be properly considered for the correct use of the corresponding theoretical methods and approaches as well as in the practical application of the developed and trained models.

Nowadays, rich data are available in many areas, as they are collected in abundance and accumulated automatically during the operation of information systems and various technical devices. At the same time, a still open challenge is to generate necessary data (with required properties) for the purpose of developing, training, and testing methods and devices. Of course, the required properties can be formalized in different ways in particular areas and problems. In this paper, such a formalization is performed within a probabilistic concept; in particular, by suitable data we mean random ensembles with appropriate probability density functions (PDFs). By assumption, appropriate PDFs can be sampled, in one way or another, i.e., transformed into corresponding random sequences.

Thus, suitable data ensembles are generated by reconstructing optimal, in an accepted sense, PDFs considering given requirements. Rather high uncertainty arises when formulating such requirements, and entropy is a natural criterion for optimizing PDFs [16–20]. However, this is not enough, and some additional properties of PDFs are often necessary. Some of them can be formulated in terms of numerical characteristics, namely, moments, semi-invariants, etc. Therefore, the construction of desired PDFs is reduced to constrained maximization of an information entropy functional. In its formal description, this problem is close to some mathematical models of randomized machine learning (RML) studied in [21, 22]. Certain differences are associated with the system of constraints in the problem under consideration.

In this paper, we further develop RML methods in the following directions:

- randomized learning under additional moment-type constraints;
- adaptation of the analytical method for solving nonlinear equations with integral components to the problem under consideration;
- effectiveness analysis of the methods in asset price forecasting as an illustrative example.

Note that RML problems and the problem of constructing desired distributions involve essentially nonlinear equations with the so-called integral components. They represent multidimensional integrals with exponential subintegral functions that are defined on simple sets (parallelepipeds) and parameterized by Lagrange multipliers. Using the analytic properties of exponential functions, these integrals are approximated by parameterized integrals of multidimensional polynomials as integrand functions. The latter are calculated analytically.

With the indicated transformations of multidimensional parameterized integrals, a system of nonlinear equations with integral components is approximated by a system of equations with a polynomial left-hand side. They are solved by an analytical method based on abstract power series [23, 24].

In view of the aforesaid, the remainder of this paper is organized as follows. Section 2 presents the general problem statement addressed. The solution approach and necessary theoretical tools are described in Section 3. Next, Section 4 is devoted to asset price forecasting. Section 5 discusses the features of the results and directions for further research. Finally, the outcomes of this paper are summarized in Section 6.

2. PROBLEM STATEMENT

Consider a random sequence $u[n]$, where $n \in \mathcal{N} = \overline{1, N}$, and let $Y_{(s\times N)}$ be a given data matrix of numerical characteristics whose elements describe the values of *normalized moments*[1] of order $k = \overline{1, s}$ at observation points (time instants) $n = \overline{1, N}$:

$$Y_{(s\times N)} = \left\{ \Big(\mathcal{M}\{u^k[n]\}\Big)^{1/k} \right\} = \left\{ y^k[n] \,|\, k = \overline{1, s}, n = \overline{1, N} \right\}. \tag{1}$$

In particular, this kind of information often appears in the forecasting of financial instruments prices at stock exchanges [26–28].

The problem of generating data with given properties can be formulated as follows:

At each time instant n, it is required to generate ensembles $\mathcal{Z}_n$ of random sequences $z[n]$, $n = \overline{1, N}$, with s normalized moments

$$m^{(k)}[n] = \Big(\mathcal{M}\{z^k[n]\}\Big)^{1/k}, \quad k = \overline{1, s}, \tag{2}$$

that equal given normalized moments $y^k[n]$ (1).

[1] Here, normalized moments are selected as numerical characteristics. But it is possible to consider, e.g., semi-invariants or the mathematical expectations of continuous functions of a random sequence.

The generator of the ensemble $\mathcal{Z}_n$ is an *input-output model* or an *auto-model*.[2] In both cases, the model has random parameters $\mathbf{a} \in \mathcal{A} \subset R^r$ of the interval type:

$$\mathbf{a} \in \mathcal{A} \subset R^r, \quad \mathcal{A} = \left[\mathbf{a}^-, \mathbf{a}^+\right]. \tag{3}$$

For each time instant n, they are characterized by a continuously differentiable PDF $P_{(n)}(\mathbf{a})$.

Depending on the availability of a priori information about the data origin, either a static or dynamic input-output model is used.

A *static input-output model* generating random sequences $z[n]$ is characterized by a nonlinear differentiable function φ with parameters $\mathbf{a}$:

$$z[n|\mathbf{a}] = \varphi\left(\mathbf{x}[n] \,|\, \mathbf{a}\right), \quad n = \overline{1, N}, \tag{4}$$

where $\mathbf{x}[n] = \{x_1[n], \dots, x_m[n]\}$ and z are the input and output of this model.

From this point onwards, the symbol "|" indicates that random parameters $\mathbf{a}$ with a PDF $P_{(n)}(\mathbf{a})$ are realized for each time instant n.

A *dynamic model* is characterized by a continuous nonlinear functional $\mathcal{B}$:

$$z[n|\mathbf{a}] = \mathcal{B}\left[\mathbf{x}[\tau],\, n - p \leqslant \tau \leqslant n \,|\, \mathbf{a}\right], \quad n = \overline{1, N}, \tag{5}$$

where p means the model memory (the number of previous input values affecting the current output value). The random parameters $\mathbf{a}$ are of the interval type (3).

This problem is solved in two stages. The *first* stage is to find the optimal probabilistic characteristics of the random parameters, namely, the PDFs $P_{(n)}(\mathbf{a})$ for all $n = \overline{1, N}$ in the static model (4) or for all $n = \overline{1 - p, N}$ in the dynamic model (5). The *second* stage consists in the sampling of these PDFs, i.e., transforming them into corresponding random ensembles $\mathcal{Z}_n$.

3. MATERIALS AND METHODS

3.1. Optimization of PDFs of Model Parameters

To solve the first-stage problem, we utilize the methodology of randomized machine learning [22]: the specified sequences will be generated by a mathematical input-output model with random parameters optimized by the information entropy criterion.

It is convenient to introduce the following s-dimensional vectors for further considerations:

- the vector of given normalized moments for each n,

$$\mathbf{y}^{(n)} = \{y_1[n], \dots, y_s[n]\}; \tag{6}$$

- the model output vector for each n,

$$\mathbf{z}^{(n)}(\mathbf{a}) = \{z_1[n \,|\, \mathbf{a}], \dots, z_s[n \,|\, \mathbf{a}]\}; \tag{7}$$

- the vector of model output's normalized moments for each n,

$$\mathbf{m}^{(n)} = \left\{ \int\limits_{\mathcal{A}} P_{(n)}(\mathbf{a})\, z[n \,|\, \mathbf{a}] d\mathbf{a}, \dots, \left(\int\limits_{\mathcal{A}} P_{(n)}(\mathbf{a})\, z^s[n \,|\, \mathbf{a}] d\mathbf{a} \right)^{1/s} \right\}. \tag{8}$$

According to [22], the problem of finding the optimal PDFs $P_{(n)}(\mathbf{a})$ can be formulated for each $n = \overline{1, N}$ as follows: maximize the information entropy functional

$$\mathcal{H}_{(n)}[P_{(n)}(\mathbf{a})] = -\int\limits_{\mathcal{A}} P_{(n)}(\mathbf{a}) \ln P_{(n)}(\mathbf{a}) d\mathbf{a} \Rightarrow \max \tag{9}$$

[2] Here, we use an input-output model. In the illustrative example below, an auto-model is described by difference equations.

subject to the constraints

- $$\int\limits_{\mathcal{A}} P_{(n)}(\mathbf{a})\, d\mathbf{a} = 1 \tag{10}$$

(the normalization conditions) and

- $$\mathbf{m}^{(n)} = \mathbf{y}^{(n)} \tag{11}$$

(the balances of the model output's normalized moments with the data).

Assuming that the PDFs are continuously differentiable, the solution parameterized by the Lagrange multipliers $\boldsymbol{\lambda}^{(n)} = \left\{\lambda_1^{(n)}, \ldots, \lambda_s^{(n)}\right\}$ has the form [22]

$$P^*_{(n)}(\mathbf{a}) = \frac{\exp\left(-\langle \boldsymbol{\lambda}^{(n)}, \mathbf{z}^{(n)}(\mathbf{a})\rangle\right)}{\mathbb{P}_{(n)}(\boldsymbol{\lambda}^{(n)})}, \tag{12}$$

where

$$\mathbb{P}_{(n)}(\boldsymbol{\lambda}^{(n)}) = \int\limits_{\mathcal{A}} \exp\left(-\langle \boldsymbol{\lambda}^{(n)}, \mathbf{z}^{(n)}(\mathbf{a})\rangle\right)\, d\mathbf{a} \tag{13}$$

and $\langle \bullet, \bullet \rangle$ denotes the inner product of vectors with s components.

The Lagrange multipliers $\boldsymbol{\lambda}^{(n)}$ satisfy the following system of s equations, called the *balance* equations:

$$\int\limits_{A} \exp\left(-\langle \boldsymbol{\lambda}^{(n)}, \mathbf{z}^{(n)}(\mathbf{a})\rangle\right) \left[\mathbf{z}^{(n)}(\mathbf{a}) - \mathbf{y}^{(n)}\right] d\mathbf{a} = \mathbf{0}. \tag{14}$$

This system determines the vectors of the Lagrange multipliers $\boldsymbol{\lambda}^{(n)}$ for each time instant n from the interval $[1, N]$.

The second-stage problem (the transformation of the entropy-optimal PDF into the corresponding random sequence) can be implemented using the methods described in [25].

3.2. An Analytical Method for Calculating Multidimensional Integrals

In view of equation (14), to determine the optimal PDF functions, it is necessary to calculate multidimensional integrals and then solve the resulting nonlinear equations. The analytical method developed allows combining these two stages.

The problem under consideration has some useful features that can be utilized when constructing an approximate analytical method. In particular, they include a simple definitional domain of the multidimensional integral (namely, a parallelepiped) and the subintegral functions represented by exponentials of continuous functions.

Note that the exponential function is analytic, and the Lagrange multipliers and the model output are bounded. Therefore, we have the following polynomial approximation of degree q :

$$\exp(-v_{(n)}) = \sum_{h=0}^{q} \frac{(-1)^h}{h!} v_{(n)}^h, \tag{15}$$

where

$$v_{(n)} = \langle \boldsymbol{\lambda}^{(n)}, \mathbf{z}^{(n)}(\mathbf{a})\rangle < \pm M < \infty \tag{16}$$

and[3]

$$(v_{(n)})^h = \sum_{i_j\geqslant 0;\,\sum_{j=1}^{s} i_j = h}^{s} A^{(h)}_{i_1,\ldots,i_h}\,(\lambda_1^{(n)})^{i_1}\cdots(\lambda_s^{(n)})^{i_h}(z^1[n\,|\,\mathbf{a}])^{i_1}\cdots(z^s[n\,|\,\mathbf{a}])^{i_h}, \tag{17}$$

$$A^{(h)}_{i_1,\ldots,i_h} = \frac{h!}{i_1!\cdots i_h!}. \tag{18}$$

With this notation, system (14) takes the form

$$\sum_{h=0}^{q}\frac{(-1)^h}{h!}\sum_{(i_1,\ldots,i_h)=1}^{s}\lambda^{(n)}_{i_1}\cdots\lambda^{(n)}_{i_h}\,u^{(n)}_{i_1,\ldots,i_h;k} - v_n^k = 0, \tag{19}$$

where

$$u^{(n)}_{i_1,\ldots,i_h;k} = \int\limits_{\mathcal{A}} z^{i_1}[n\,|\,\mathbf{a}]\cdots z^{i_q}[n\,|\,\mathbf{a}]\left(z^k[n\,|\,\mathbf{a}] - y^{(k)}[n]\right)d\mathbf{a},\quad i_1,\ldots,i_h,\; k=\overline{1,s}. \tag{20}$$

The numbers $i_1,\ldots,i_h$ take values in the interval $[1,s]$. This system contains s variables (the Lagrange multipliers), and each equation is a multivariate polynomial of degree q.

By assumption, the model is characterized by a continuous function (or functional). Hence, it can be represented by a multivariate polynomial, and then the multidimensional integrals in (20) are transformed into the product of one-dimensional integrals calculated analytically. Therefore, the balance system (19) has a polynomial left-hand side.

Consider the following family of systems in the parameter $\varepsilon\in[0,1]$:

$$-\sum_{i_1=1}^{s}\lambda^{(n)}_{i_1}\,u^{(n)}_{i_1,k} + \varepsilon\sum_{h=1}^{q}\frac{(-1)^h}{h!}\sum_{\substack{(i_1,\ldots,i_h)=1\\ \sum_{j=1}^{h} i_j = h}}^{s}\lambda^{(n)}_{i_1}\cdots\lambda^{(n)}_{i_h}\,u^{(n)}_{i_2,\ldots,i_h;k} - v_n^k = 0. \tag{21}$$

For $\varepsilon=0$, we have the so-called *basic* system, linear with a square matrix $U^{(n)} = [u^n_{i_1,k},\,|\,(i_1,k) = \overline{1,s}]$:

$$U^{(n)}\boldsymbol{\lambda}^{(n)} = -\mathbf{v}_n, \tag{22}$$

where $\mathbf{v}_n = \{v_n^1,\ldots,v_n^s\}$.

If $\det U^{(n)}\neq 0$, then the Lagrange multipliers yielding the *basic* solution are

$$\boldsymbol{\lambda}^{(n)}_{(\bullet)} = -[U^{(n)}]^{-1}\mathbf{v}_n. \tag{23}$$

We write the solution of system (21) as an abstract power series in the parameter ε [23, 24]:

$$\lambda^{(n)}_{k,\star} = \lambda^{(n)}_{k,\bullet} + \varepsilon\lambda^{(n)}_{k,I} + \varepsilon^2\lambda^{(n)}_{k,II} + \cdots,\quad k=\overline{1,s}, \tag{24}$$

where $\lambda^{(n)}_{k,I}, \lambda^{(n)}_{k,II},\ldots$ are the first, second,…, corrections to the basic solution, respectively.

To find the corrections sequentially, we apply the method of undetermined coefficients [23] to obtain:

— the first correction vector

$$^{*}\boldsymbol{\lambda}^{(n)}_{I} = -[U^{(n)}]^{-1}\,\mathbf{b}^{(n)}_{(1)}(\boldsymbol{\lambda}^{(n)}_{(\bullet)}), \tag{25}$$

where

$$\mathbf{b}^{(n)}_{(I)}(\boldsymbol{\lambda}^{(n)}_{(\bullet)}) = \left\{\frac{1}{2}\sum_{(i_1,i_2)=1}^{s}\lambda^{(n)}_{i_1,\bullet}\lambda^{(n)}_{i_2,\bullet}u^{(n)}_{i_1,i_2,1},\ldots,\frac{1}{2}\sum_{(i_1,i_2)=1}^{s}\lambda^{(n)}_{i_1,\bullet}\lambda^{(n)}_{i_2,\bullet}u^{(n)}_{i_1,i_2,s}\right\}, \tag{26}$$

[3] The expression on the right-hand side of (17) is Newton's polynomials of degree h [31].

and the second correction vector

$$ {}^{*}\boldsymbol{\lambda}^{n}_{(II)} = -[U^{(n)}]^{-1}\,\mathbf{b}^{(n)}_{(II)}(\boldsymbol{\lambda}^{n}_{(\bullet)}, \boldsymbol{\lambda}^{n}_{(I)}), \tag{27}$$

where

$$\left\{\frac{1}{2}\sum_{(i_1,i_2)=1}^{s}\lambda^{(n)}_{i_1,I}\lambda^{(n)}_{i_2,\bullet}u^{(n)}_{i_1,i_2,k} + \frac{1}{3!}\sum_{(i_1,i_2,i_3)=1}^{s}\lambda^{(n)}_{i_1,\bullet}\lambda^{(n)}_{i_2,\bullet}\lambda^{(n)}_{i_3,\bullet}u^{(n)}_{i_1,i_2,i_3,k}\right\}. \tag{28}$$

Thus, the solution of the balance system (21) is written as

$$\boldsymbol{\lambda}^{(n)}_{\star} = \boldsymbol{\lambda}^{(n)}_{(\bullet)} +^{*} \boldsymbol{\lambda}^{(n)}_{I} +^{*} \boldsymbol{\lambda}^{(n)}_{II} + \cdots . \tag{29}$$

3.3. Asset Pricing Model

Consider the pricing process of an asset during trading sessions and apply the above method for generating random data with given numerical characteristics to forecast price dynamics.

By a theoretical consensus, the price of an asset at any time instant is the product of balancing real demand and real supply. However, a trader is guided by the expected demand and supply, which may significantly differ from the real ones. In these conditions, price forecasting under sufficiently high uncertainty becomes crucial.

Therefore, it seems natural to attempt to maximize the information entropy as a measure of uncertainty on training retrospective data containing the values of mean price and its second moment (variance describing mean volatility).

3.3.1. Price Dynamics Model

We adopt an autonomous model in the form of a linear difference equation of order p, but with random interval-type parameters:

$$C[t] = \sum_{i=1}^{p} a_i C[t-i], \quad t \in \mathcal{T}, \quad a_i \in \mathcal{A}_i = [d, w], \ \mathcal{A}^p = \prod_{i=1}^{p}\mathcal{A}_i. \tag{30}$$

(All parameters have the same intervals $[d, w]$.)

We will use this model at the stages of *training*, $\mathcal{T}_{trn} = [t_0, t_0 + p]$, and *forecasting*, $\mathcal{T}_{frc} = [t_0 + p + 1, t_0 + p + 1 + t_{frc}]$, where t_{frc} are one- or two-day forecasts. (Longer forecasts can be considered by analogy.)

The probabilistic properties of the parameters are characterized by a continuously differentiable PDF $P_t(\mathbf{a})$.

By assumption, based on the results of each real trading session t, two price indicators are formed: the mean price $m^{*}_{C}[t]$ and its second moment $D^{*}_{C}[t]$ as a characteristic of mean volatility:

$$D^{*}_{C}[t] = (V^{*}_{C}[t])^2 + (m^{*}_{C}[t])^2, \tag{31}$$

where $V^{*}_{C}[t]$ is the estimated standard deviation of the price, constructed from the value of its maximum and minimum deviation.

3.3.2. Data

Consider real price and volatility dynamics data available on a past interval $[t_0 - p, t_0 - 1]$:

$$m^{*}_{C}[t_0 - p], m^{*}_{C}[t_0 - p + 1], \ldots, m^{*}_{C}[t_0 - 1]$$

and

$$D^{*}_{C}[t_0 - p], D^{*}_{C}[t_0 - p + 1], \ldots, D^{*}_{C}[t_0 - 1],$$

respectively.

Using these data, *the model* (30) *generates* the prices on the training interval $\mathcal{T}_{trn}$:

$$\mathbb{C}[\mathbf{a}\,|\,t] = \sum_{i=1}^{p} a_i m_C^*[t-i], \quad t \in \mathcal{T}_{trn}. \tag{32}$$

3.3.3. The Entropy-Optimal Estimation of the PDFs of Model Parameters

To optimize the PDF $P_t(\mathbf{a})$, we will apply randomized machine learning [22], see formulas (9)–(11). This methodology is reduced to solving the following problems for each trading session $t \in \mathcal{T}_{trn}$:

$$\mathcal{H}_t[P_t(\mathbf{a})] = -\int\limits_{\mathcal{A}^p} P_t(\mathbf{a})\, lnP_t(\mathbf{a})\, d\mathbf{a} \Rightarrow \max_{P(\mathbf{a})} \tag{33}$$

subject to the constraints

$$\int\limits_{\mathcal{A}^p} P_t(\mathbf{a})\, d\mathbf{a} = 1, \tag{34}$$

$$\int\limits_{\mathcal{A}^p} P_t(\mathbf{a})\, \mathbb{C}[\mathbf{a}\,|\,t]\, d\mathbf{a} = m_C^*[t], \quad \int\limits_{\mathcal{A}} P_t(\mathbf{a}) \mathbb{C}^2[\mathbf{a}\,|\,t]\, d\mathbf{a} = D_C^*[t], \; t \in \mathcal{T}_{trn}, \tag{35}$$

where $\mathbb{C}_t[\mathbf{a}\,|\,t]$ are given by (32).

According to (12) and (13), problem (33)–(35) has the analytical solution

$$\begin{aligned} P_t^*(\mathbf{a}) &= \frac{\exp\left(-\lambda_1^{(t)}\mathbb{C}[\mathbf{a}\,|\,t] - \lambda_2^{(t)}\mathbb{C}^2[\mathbf{a}\,|\,t]\right)}{\mathcal{P}_t(\lambda_1^{(t)}, \lambda_2^{(t)})}, \\ \mathcal{P}_t^*(\lambda_1^{(t)}, \lambda_2^{(t)}) &= \int\limits_{\mathcal{A}^p} \exp\left(-\lambda_1^{(t)}\mathbb{C}[\mathbf{a}\,|\,t] - \lambda_2^{(t)}\mathbb{C}^2[\mathbf{a}\,|\,t]\right) d\mathbf{a}, \end{aligned} \qquad t \in \mathcal{T}_{trn}. \tag{36}$$

The Lagrange multipliers $\lambda_1^{(t)}, \lambda_2^{(t)}$ are found by solving the two balance equations

$$\begin{aligned} &\int\limits_{\mathcal{A}^p} \exp\left(-\lambda_1^{(t)}\mathbb{C}[\mathbf{a}\,|\,t] - \lambda_2^{(t)}\mathbb{C}^2[\mathbf{a}\,|\,t]\right)(\mathbb{C}[\mathbf{a}\,|\,t-1] - m_C^*[t])\, d\mathbf{a} = 0, \\ &\int\limits_{\mathcal{A}^p} \exp\left(-\lambda_1^{(t)}\mathbb{C}[\mathbf{a}\,|\,t] - \lambda_2^{(t)}\mathbb{C}^2[\mathbf{a}\,|\,t]\right)\left(\mathbb{C}^2[\mathbf{a}\,|\,t-1] - D_C^*[t]\right) d\mathbf{a} = 0, \end{aligned} \qquad t \in \mathcal{T}_{trn}. \tag{37}$$

According to the method developed in [29], we approximate the exponent by a polynomial of degree 2 :

$$\begin{aligned} \exp(x) &\approx \left(1 - \lambda_1^{(t)}\mathbb{C}[\mathbf{a}\,|\,t] + \frac{1}{2}\left(\lambda_1^{(t)}\mathbb{C}[\mathbf{a}\,|\,t]\right)^2\right)\mathbb{C}[\mathbf{a}\,|\,t], \\ \exp(y) &\approx \left(1 - \lambda_2^{(t)}\,\mathbb{C}^2[\mathbf{a}\,|\,t] + \frac{1}{2}\left(\lambda_2^{(t)}\,\mathbb{C}^2[\mathbf{a}\,|\,t]\right)^2\right)\mathbb{C}^2[\mathbf{a}\,|\,t]. \end{aligned} \tag{38}$$

With the approximations (38), the balance equations take the form

$$\begin{aligned} &\lambda_1 B_1(t) + \lambda_2 B_2(t) + \lambda_1^2 B_3(t) + \lambda_2^2 B_4(t) + \lambda_1\lambda_2 B_5(t) = B_0(t), \\ &\lambda_1 Z_1(t) + \lambda_2 Z_2(t) + \lambda_1^2 Z_3(t) + \lambda_2^2 Z_4(t) + \lambda_1\lambda_2 Z_5(t) = Z_0(t), \end{aligned} \qquad t \in \mathcal{T} = [t_0, t_0 + p]. \tag{39}$$

In the first equation above, the coefficients are

$$\begin{aligned} B_0(t) &= A\, m_C^*[t] - I_{p,1}^{(t)}, \quad B_1(t) = m_C^*[t]\, I_{p,1}^{(t)} - I_{p,2}^{(t)}, \\ B_2(t) &= m_C^*[t]\, I_{p,2}^{(t)} - I_{p,3}^{(t)}, \quad B_3(t) = -\frac{1}{2}\, B_2^{(2,3)}(t), \\ B_4(t) &= -\frac{1}{2}\left(m_C^*[t]\, I_{p,4}^{(t)} - I_{p,5}^{(t)}\right), \quad B_5(t) = -\left(m_C^*[t]\, I_{p,3}^{(t)} - I_{p,4}^{(t)}\right). \end{aligned} \tag{40}$$

In the second equation, the coefficients are

$$\begin{gathered}
Z_0(t) = A\,D^*_C[t] - I^{(t)}_{p,2}, \quad Z_1(t) = D^*_C[t]\,I^{(t)}_{p,1} - I^{(t)}_{p,3}, \\
Z_2(t) = D^*_C[t]\,I^{(t)}_{p,2} - I^{(t)}_{p,4}, \quad Z_3(t) = -\frac{1}{2}\,Z_2^{(2,4)}(t), \\
Z_4(t) = -\frac{1}{2}\left(D^*_C[t]\,I^{(t)}_{p,4} - I^{(t)}_{p,6}\right), \quad Z_5(t) = -\left(D^*_C[t]\,I^{(t)}_{p,3} - I^{(t)}_{p,5}\right).
\end{gathered} \tag{41}$$

In these expressions,

$$\begin{gathered}
A = \int\limits_{\mathcal{A}} d\mathbf{a} = (w-d)^p, \\
I^{(t)}_{p,n}(k_1,\ldots,k_n) = \underbrace{\int\limits_d^w \cdots \int\limits_d^w}_{p} \mathbb{C}^n[\mathbf{a}\,|\,t]\,d\mathbf{a} \\
= \sum_{k_j\geqslant 0;\ \sum_{j=1}^n k_j = n} \frac{n!}{k_1!\cdots k_n!} \left(\underbrace{\int\limits_d^w \cdots \int\limits_d^w}_{p} a_1^{k_1}\cdots a_p^{k_n}\,da_1\cdots da_p \right) \\
\times (m^*_C[t])^{k_1}\cdots(m^*_C[t-p])^{k_n}, \quad n = \overline{0,6}.
\end{gathered} \tag{42}$$

4. FORECASTING THE FUTURE PRICE OF A TRADED ASSET

We will experimentally test the method for the *one- and two-day forecasting* of the mean price and mean volatility of a traded asset, i.e., Gazprom's stocks during 2020 on the Moscow Exchange.

Consider twelve trading sessions, each at the beginning of the month. For the convenience of further calculations, let us introduce a conventional monetary unit (c.m.u. = 1000 rubles). The stock price data in c.m.u. are combined in Table 1.

Table 1. Gazprom's stock price quotations in 2020

Month	Jan.	Feb.	Mar.	Apr.	May	Jun.
Data t	1	2	3	4	5	6
Price m^*_C	0.259	0.223	0.208	0.178	0.188	0.200
Max $C^*_{\max}$	0.262	0.240	0.212	0.196	0.202	0.208
Min $C^*_{\min}$	0.227	0.201	0.158	0.177	0.182	0.190
V^*_C	0.017	0.020	0.027	0.009	0.011	0.009
D^*_C	0.084	0.069	0.070	0.041	0.046	0.049
Month	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Date t	7	8	9	10	11	12
Price m^*_C	0.195	0.182	0.181	0.171	0.154	0.183
Max $C^*_{\max}$	0.202	0.195	0.186	0.173	0.189	0.215
Min $C^*_{\min}$	0.179	0.180	0.170	0.154	0.152	0.182
V^*_C	0.011	0.007	0.009	0.010	0.019	0.017
D^*_C	0.206	0.189	0.190	0.181	0.173	0.200

4.1. Training of the Price Model

The price model (30) possesses a memory of $p = 2$, the initial training instant is $t_0 = 3 \to$ Mar., and the interval limits of the two parameters are $d = -1$, $w = 2$.

The price model will be trained on the interval $\mathcal{T} = [t_0, t_0 + 2] = [3, 5]$. The historical period is $\mathcal{I}_p = [t_0 - 2, t_0 - 1] = [1, 2]$.

Since model (32) contains the two parameters,

$$\mathbb{C}[\mathbf{a} \,|\, t] = a_1\, m_C^*[t-1] + a_2\, m_C^*[t-2]. \tag{43}$$

The PDFs in the corresponding trading sessions have the form (36) with model (43). Note that the entropy-optimal PDFs of the parameters generated by the linear model (43) differ from the Gaussian distribution.

To find the Lagrange multipliers (solve the balance equations), we apply the approximate analytical method from subsection 2.2.

For trading sessions $t = 3, 4, 5$, the entropy-optimal PDFs with the approximate Lagrange multipliers within the first correction have the form

$$P_3^*(\mathbf{a} \,|\, 1.068;\ -0.871) = 0.131\ \exp(-0.238a_1 - 0.277a_2 + 0.043a_1^2 + 0.058a_2^2 + 0.100a_1a_2), \tag{44}$$

$$P_4^*(\mathbf{a} \,|\, 0.958;\ 0.102) = 0.133\ \exp(-0.199a_1 - 0.214a_2 - 0.004a_1^2 - 0.005a_2^2 - 0.005a_1a_2), \tag{45}$$

$$P_5^*(\mathbf{a} \,|\, -1.994;\ 2.609) = 0.092\,(0.355a_1 + 0.415a_2 - 0.083a_1^2 - 0.112a_2^2 - 0.193a_1a_2). \tag{46}$$

Their plots are shown in Figs. 1–3.

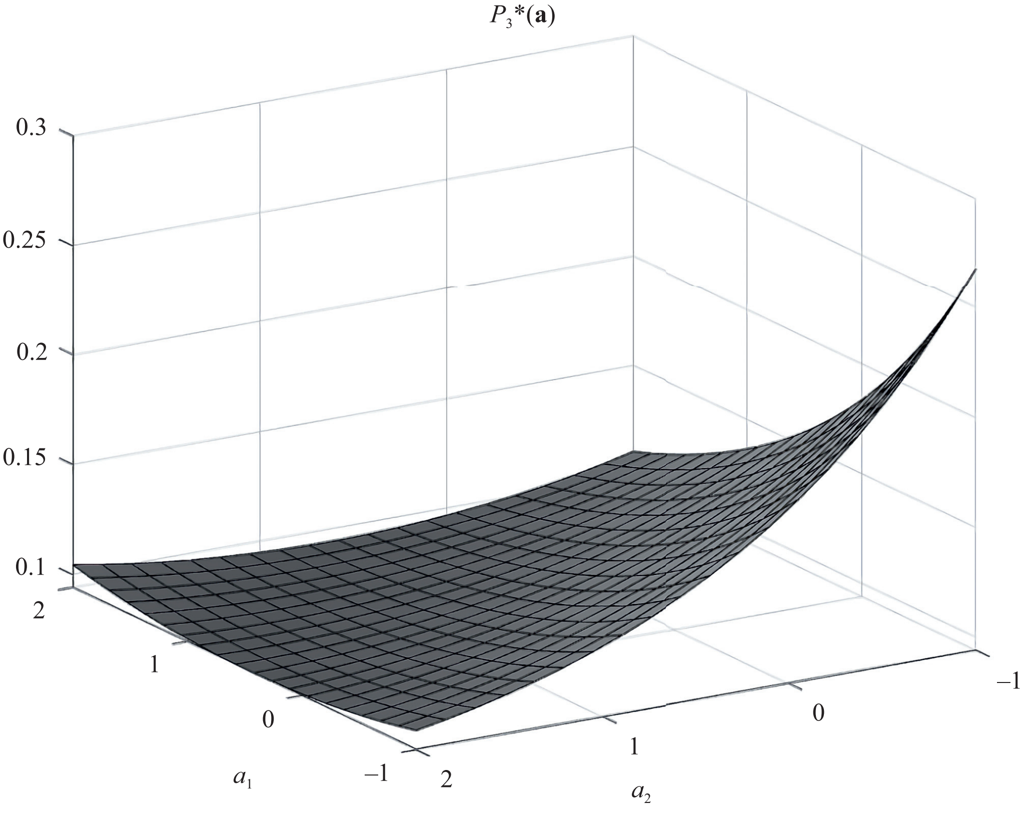

Fig. 1. The distribution $P_3^*(\mathbf{a})$.

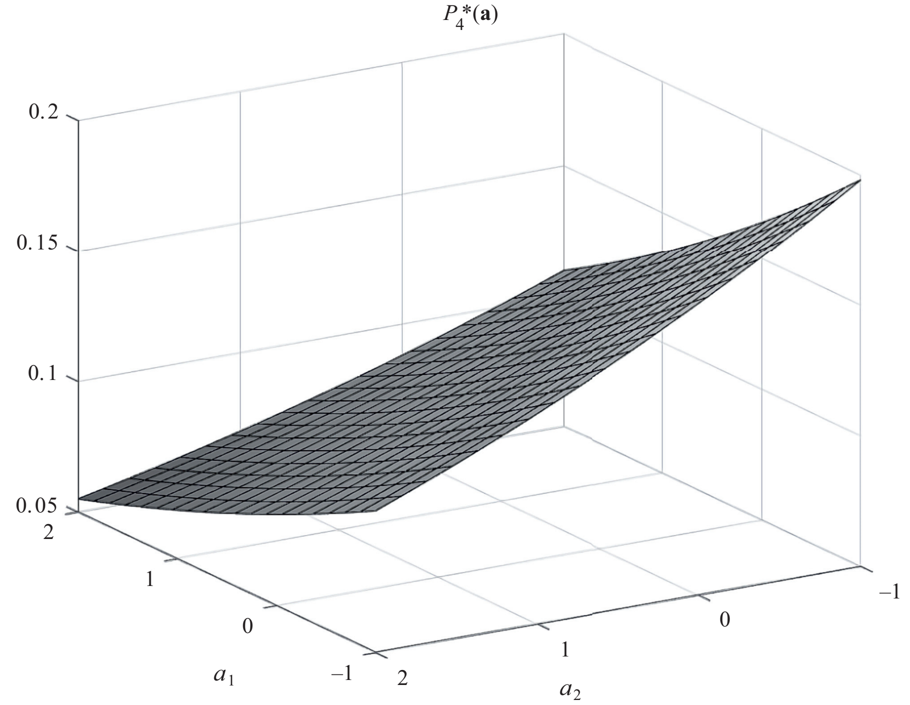

Fig. 2. The distribution $P_4^*(\mathbf{a})$.

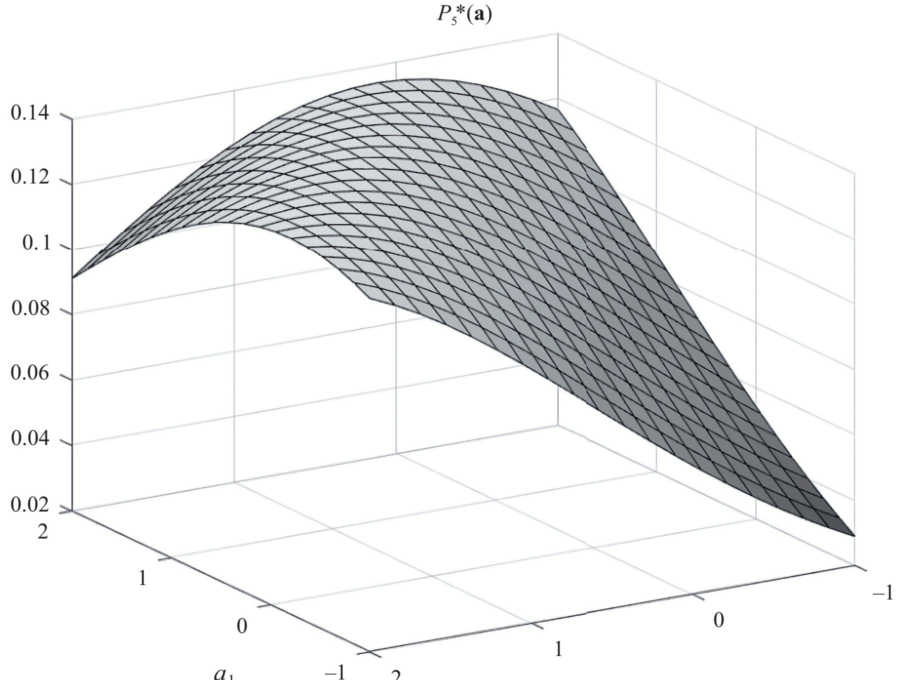

Fig. 3. The distribution $P_5^*(\mathbf{a})$.

4.2. Forecasting the Mean Price and Mean Volatility

The above entropy-optimal PDFs $P_3^*(\mathbf{a}), P_4^*(\mathbf{a})$, and $P_5^*(\mathbf{a})$ of the model parameters (32) will be used to generate data ensembles and then calculate the forecasted values of $m_C[t]$ and $D_C[t]$ in trading sessions from Apr. (4) to Nov. (11). The realized values of these variables are known (Table 1); hence, it is possible to estimate the accuracy of different forecasting strategies.

4.2.1. One-Day Forecasts $P_k^*(\mathbf{a}) \to (m_C[k+1], V_C[k+1])$

For one-day forecasts, we use the optimal PDF for trading session k to predict the results of trading session $(k+1)$. Consider the procedure for constructing the forecast 3 → 4. For this purpose, it is necessary to use the PDF $P_3^*(\mathbf{a})$ (44) and the forecasting model (32); in this example, the model takes the form

$$\mathbb{C}[\mathbf{a}\,|\,4] = a_1 m_C^*[3] + a_2 m_C^*[2]. \tag{47}$$

We transform the PDF $P_3^*(\mathbf{a})$ (44) into the random sequence $\{a_1, a_2\}$. The generated ensemble contains 1000 values $\mathbb{C}[\mathbf{a}\,|\,4]$. We calculate $\bar{m}_C[4] = \tilde{\mathcal{M}}(\mathbb{C}[\mathbf{a}\,|\,4])$ and $\bar{\sigma}_C^2[4] = \tilde{\mathcal{M}}\{(\mathbb{C}[\mathbf{a}\,|\,4] - \bar{m}_C[4])^2\}$, where $\tilde{\mathcal{M}}\{\bullet\}$ denotes the empirical mean operator.

The forecasts 4 → 5 and 5 → 6 are constructed by analogy. Table 2 presents the resulting one-day forecasts and their accuracy estimates compared to the realized values in the trading sessions.

Table 2. One-day forecasts

Forecast $\bullet$	3 → 4	4 → 5	5 → 6
$\bar{m}_C[\bullet]$	0.175	0.145	0.229
$m_C^*[\bullet]$	0.178	0.188	0.200
$\bar{\sigma}_C^2[\bullet]$	0.076	0.056	0.041
$V_C^*[\bullet]$	0.041	0.046	0.049
$\lvert\delta_m[\bullet]\rvert$	0.003	0.043	0.029
$\lvert\delta_\sigma[\bullet]\rvert$	0.035	0.010	0.008

In this table, the variables are

$$\delta_m[\bullet] = \bar{m}_C[\bullet] - m_C^*[\bullet], \quad \delta_\sigma[\bullet] = \bar{\sigma}_C^2[\bullet] - V_C^*[\bullet]. \tag{48}$$

Figures 4–6 show the empirical PDFs of the forecasted prices in trading sessions 4, 5, and 6 under one-day forecasting.

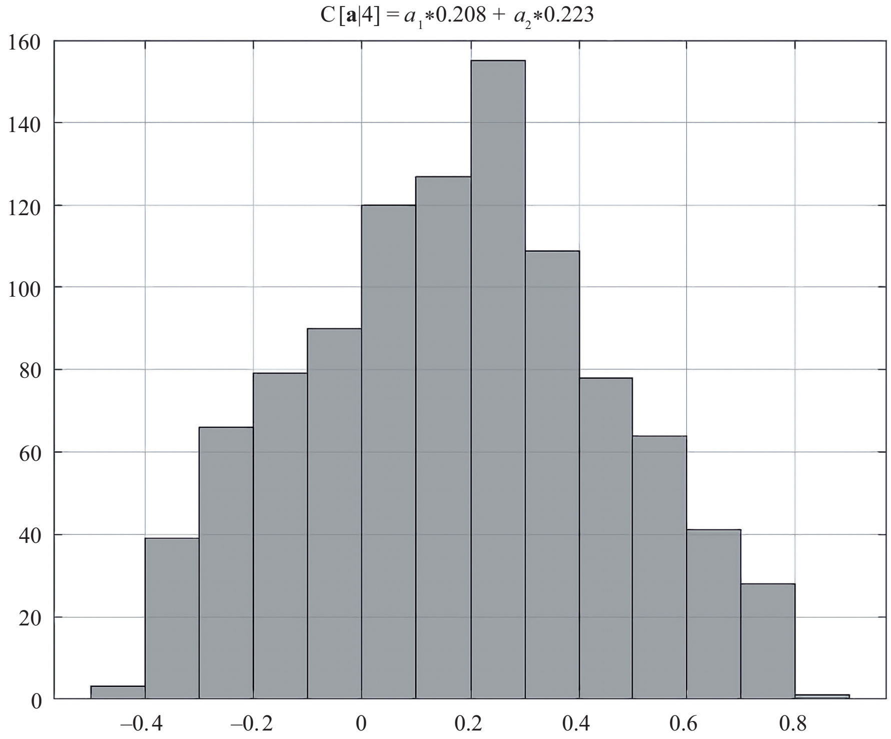

Fig. 4. The empirical PDF $\mathbb{C}[\mathbf{a}\,|\,4]$ (one-day forecast 3 → 4).

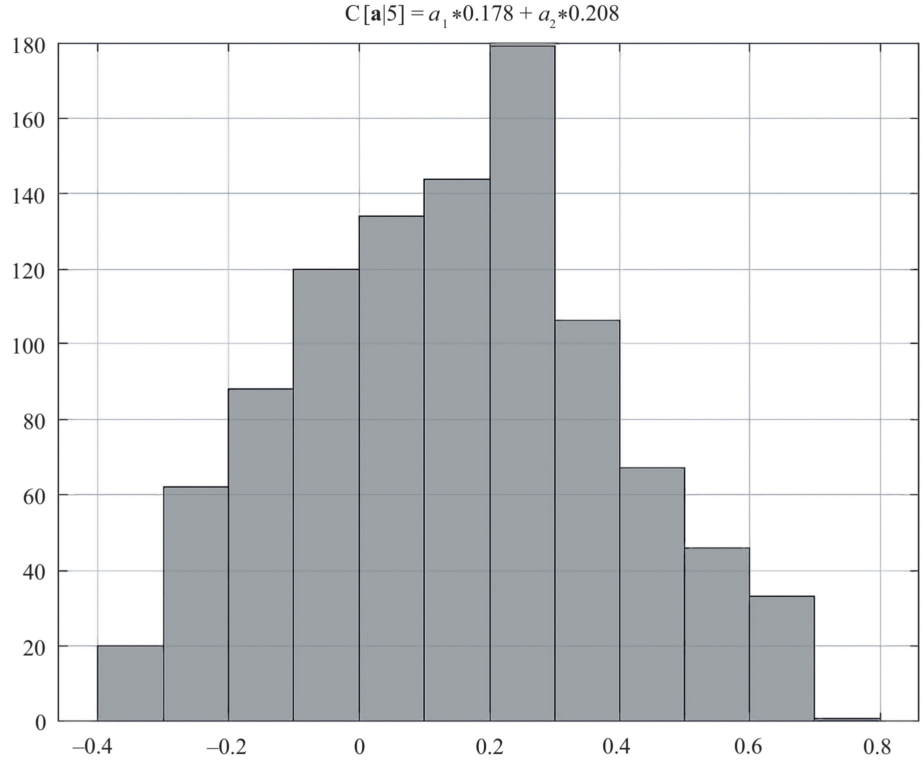

Fig. 5. The empirical PDF $\mathbb{C}[\mathbf{a}\,|\,5]$ (one-day forecast 4 → 5).

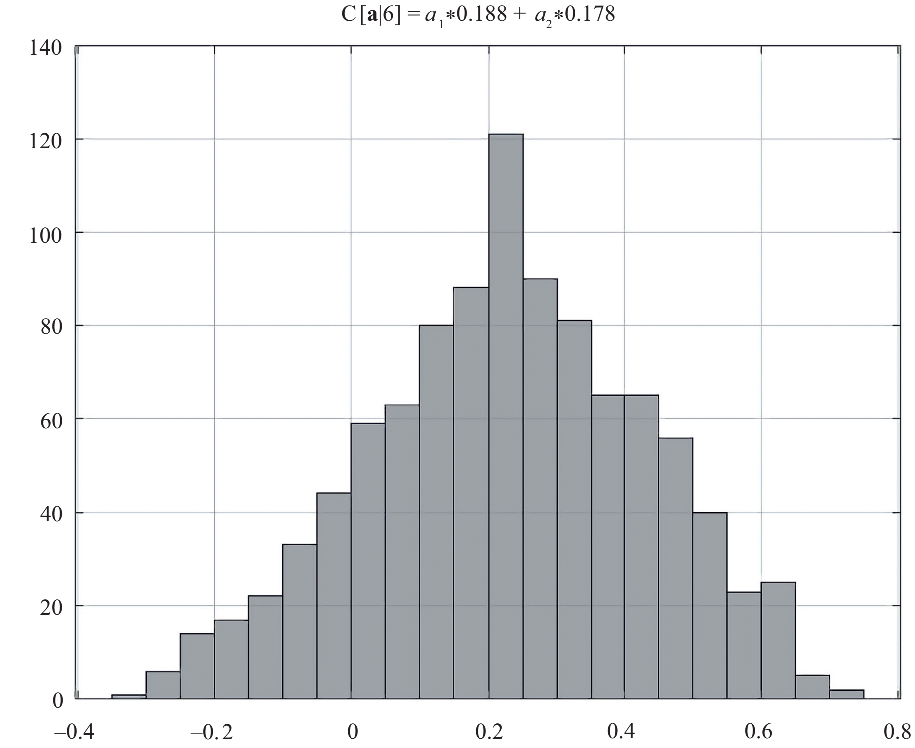

Fig. 6. The empirical PDF $\mathbb{C}[\mathbf{a}\,|\,6]$ (one-day forecast 5 → 6).

The integral relative error of the *mean price forecast* under one-day forecasting is given by

$$\Delta_m = \frac{\sqrt{\sum_{t=4}^{6} \delta_m^2[t]}}{\sqrt{\sum_{t=4}^{6} m_C^2[t]} + \sqrt{\sum_{t=4}^{6} (m_C^*[t])^2}} = 8\%. \tag{49}$$

The integral relative error of the *mean volatility forecast* under one-day forecasting is given by

$$\Delta_\sigma = \frac{\sqrt{\sum_{t=4}^{6} \delta_\sigma^2[t]}}{\sqrt{\sum_{t=4}^{6} \sigma_C^2[t]} + \sqrt{\sum_{t=4}^{6} (V_C^*[t])^2}} = 17\%. \tag{50}$$

4.2.2. Two-Day Forecasts $P_k^*(\mathbf{a}) \to (m_C[k+1], V_C[k+1]), (m_C[k+2], V_C[k+2])$

In two-day forecasts: the optimal PDF for trading session k is used to predict the results of trading session $(k+2)$.

The forecast $3 \to 4, 5$ with the PDF $P_3^*(\mathbf{a})$ can be implemented as

$$\begin{aligned} \mathbb{C}[\mathbf{a}\,|\,4] &= a_1\, m^*[3] + a_2\, m_C^*[2], \\ \mathbb{C}[\mathbf{a}\,|\,5] &= a_1\, m^*[4] + a_2\, m_C^*[3] \end{aligned} \tag{51}$$

(sequential one-day forecasts) or as

$$\begin{aligned} \mathbb{C}[\mathbf{a}\,|\,4] &= a_1\, m^*[3] + a_2\, m_C^*[2] = a_1\, 0.208 + a_2\, 0.223, \\ \mathbb{C}[\mathbf{a}\,|\,5] &= a_1\, \bar{m}^*[4] + a_2\, m_C^*[3] = a_1\, \bar{m}^*[4] + a_2\, 0.208, \\ \bar{m}^*[4] &= \tilde{\mathcal{M}}\{\mathbb{C}[\mathbf{a}\,|\,4]\} \end{aligned} \tag{52}$$

(using information from the first one-day forecast).

The two-day forecasts $4 \to 5, 6$ and $5 \to 6, 7$ are constructed by analogy. Table 3 presents the resulting two-day forecasts and their accuracy estimates compared to the realized values in the trading sessions.

Table 3. Two-day forecasts

•	3 → 4	3 → 5	4 → 5	4 → 6	5 → 6	5 → 7
$\bar{m}_C[\bullet]$	0.175	0.185	0.145	0.206	0.229	0.200
$m_C^*[\bullet]$	0.178	0.188	0.188	0.200	0.200	0.195
$\bar{\sigma}_C^2[\bullet]$	0.056	0.060	0.056	0.012	0.041	0.024
$V_C^*[\bullet]$	0.041	0.051	0.046	0.009	0.049	0.011
$\lvert\delta_m[\bullet]\rvert$	0.003	0.033	0.043	0.078	0.029	0.065
$\lvert\delta_\sigma[\bullet]\rvert$	0.035	0.049	0.010	0.030	0.009	0.043

The integral relative error of the *mean price forecast* under two-day forecasting is given by

$$\Delta_m = \frac{\sqrt{\sum_{t=5}^{7} \delta_m^2[t]}}{\sqrt{\sum_{t=5}^{7} m_C^2[t]} + \sqrt{\sum_{t=5}^{7} (m_C^*[t])^2}} = 7.2\%. \tag{53}$$

The integral relative error of the *mean volatility forecast* under two-day forecasting is given by

$$\Delta_\sigma = \frac{\sqrt{\sum_{t=5}^{7} \delta_\sigma^2[t]}}{\sqrt{\sum_{t=5}^{7} \sigma_C^2[t]} + \sqrt{\sum_{t=5}^{7} (V_C^*[t])^2}} = 25\%. \tag{54}$$

Figures 7–9 show the empirical PDFs of the forecasted prices in trading sessions 5–7 under two-day forecasting.

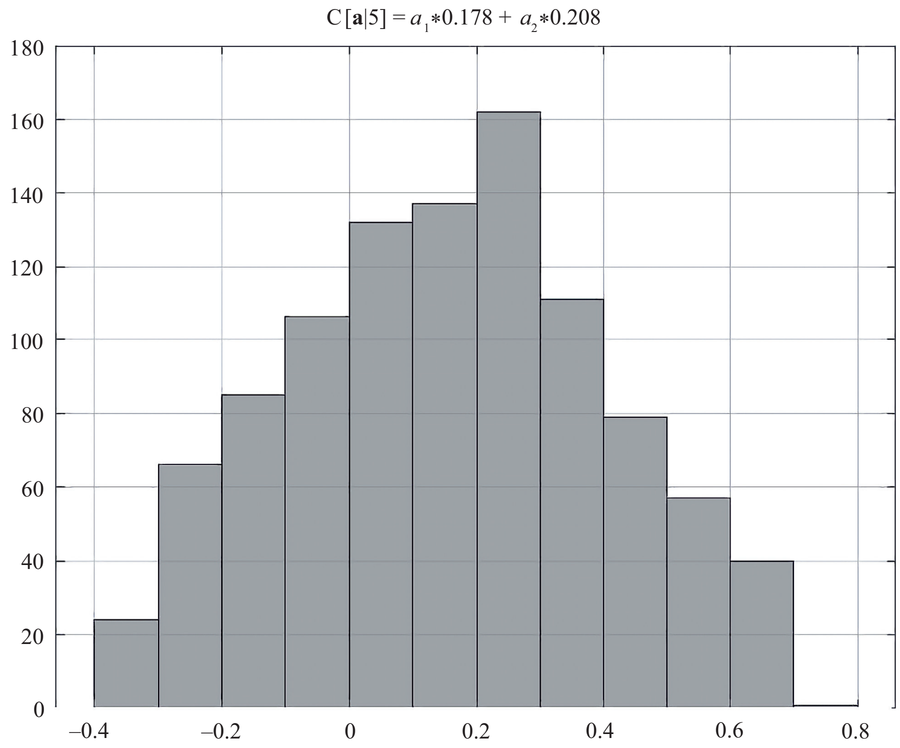

Fig. 7. The empirical PDF $\mathbb{C}[\mathbf{a}\,|\,5]$ (two-day forecast 3 → 4, 5).

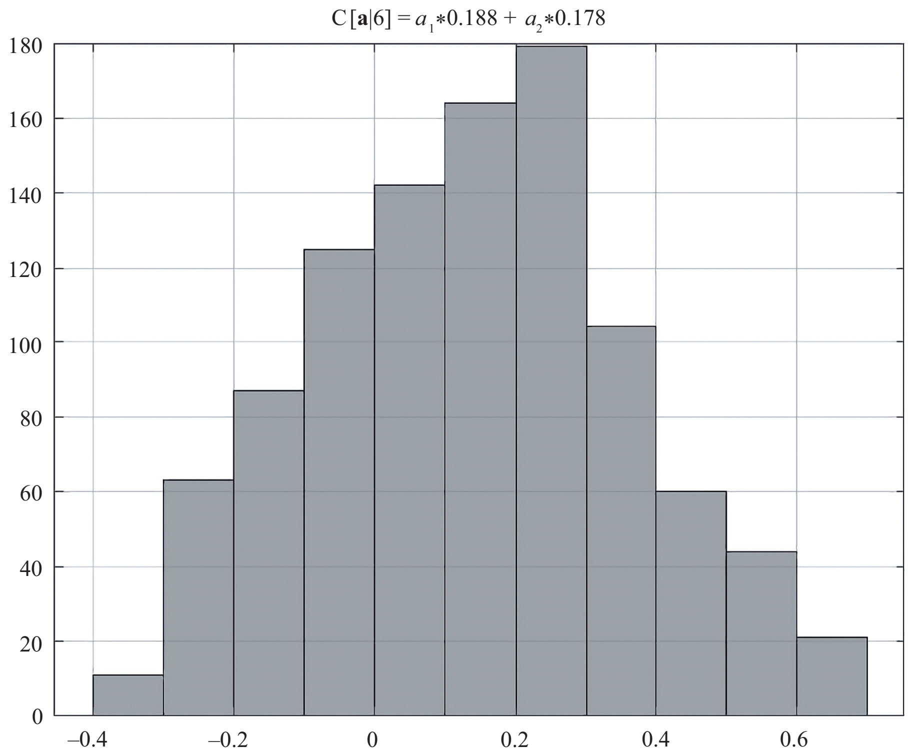

Fig. 8. The empirical PDF $\mathbb{C}[\mathbf{a}\,|\,6]$ (two-day forecast 4 → 5, 6).

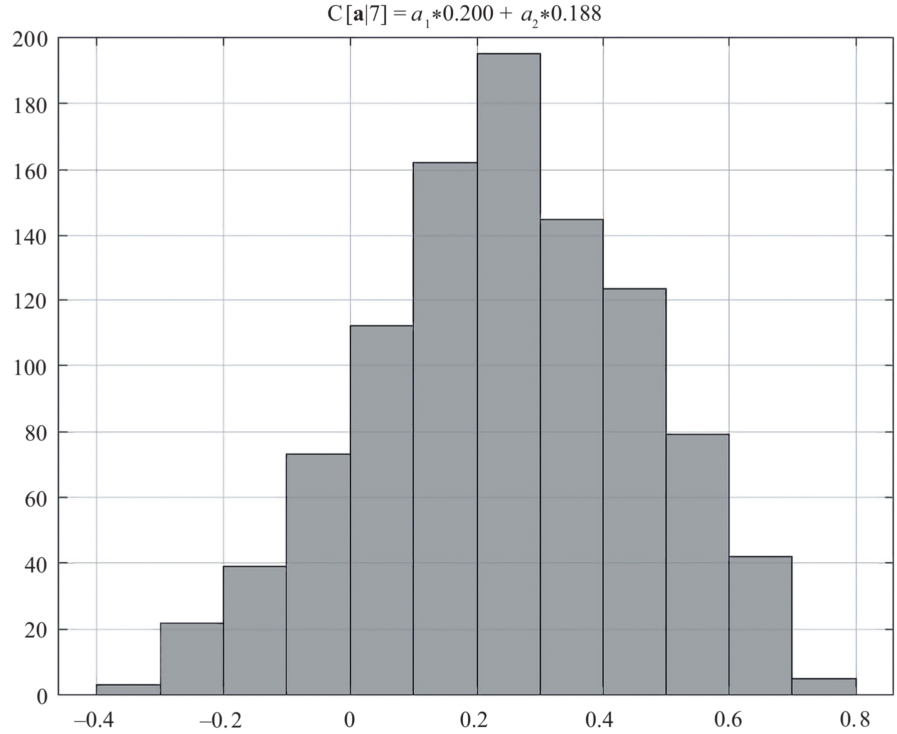

Fig. 9. The empirical PDF $\mathbb{C}[\mathbf{a}\,|\,7]$ (two-day forecast $5 \to 6, 7$)

5. DISCUSSION

The problem of generating suitable data for testing and forecasting is quite popular in modern computer science. In this paper, we have adapted and further developed the technology of randomized machine learning for generating data ensembles with given numerical characteristics.

Unlike the existing technology, an extension has been proposed to consider moment characteristics from the 1st to sth order. According to the results, this approach leads to non-Gaussian PDFs even in the case of a linear data model. Similar to the existing technology, the extension is reduced to solving the corresponding balance equations with integral components. The paper has presented an approximate analytical solution method for these equations based on power series and the method of undetermined coefficients.

It has been applied to forecast the price of an asset, and the results have been compared with the realized data for one- and two-day forecasts. Within this study, quite acceptable accuracy of the approximate solution has been discovered through two corrections. However, in-depth research into the approximate method is necessary, both in its theoretical aspects and numerical simulation.

6. CONCLUSIONS

This paper has presented a theory and algorithm for generating test data ensembles with specified properties (numerical characteristics) based on a structural modification of the randomized machine learning procedure [22]. As is known, the core of this procedure is the balance equations for the Lagrange multipliers, which contain the so-called integral components (the multidimensional integrals of arbitrary subintegral functions).

The *analytical* solution method developed in [29] has been adapted for solving these equations. With this method, the multidimensional integration problem is reduced to calculating the sum of products of the one-dimensional integrals of power functions.

Finally, a randomized forecasting method has been developed and applied to construct one- and two-day forecasts of the mean price and mean volatility of a traded asset.

FUNDING

This work was supported by the Ministry of Science and Higher Education of the Russian Federation, project no. 075-15-2024-544.

REFERENCES

1. Rubinstein, R.Y. and Kroese, D.P., *Simulation and the Monte Carlo Method*, John Wiley & Sons, 2016.
2. Vapnik, V.N., *Statistical Learning Theory*, Wiley, 1998.
3. Bishop, C.M., *Pattern Recognition and Machine Learning*, Springer, 2006.
4. Hastie, T., Tibshirant, R., and Friedman, J., *The Elements of Statistical Learning*, Springer, 2009.
5. Vovk, V. and Shafer, G., Good Randomized Sequential Probability Forecasting Is Always Possible, *Journal of Royal Statistical Society B*, 2005, vol. 67, no. 5, pp. 747–763.
6. Hong, T., Prinson, P., Fan, S., Zareijpour, H., Triccoli, A., and Hyndman, R.J., Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 and Beyond, *Inter. Journal of Forecasting*, 2016, vol. 32, no. 3, pp. 896–913.
7. Zhang, L., Aggarwal, C.C., and Qi, G.-J., Stock Price Prediction via Discovering Multy-Frequency Trading Patterns, *Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2017, pp. 2141–2149.
8. Myers, G.J., *The Art of Software Testing*, John Wiley & Sons, 1979.
9. Gorodetsky, V.I., Grushitsky, M.S., and Khabalov, A.V., Multiagent Systems, *Artificial Intelligence News*, 1998, no. 2, pp. 64–116.
10. Patton, R., *Software Testing*, SAWS Publishers, 2005.
11. Lysikov, A.I., But, G.S., and Didenko, D.A., A Computer Test Development System, 2002. http://www.bytic.ru/cue99m/cf7pvke.html
12. Micel, A.A. and Poguda, A.A., Neural Networks Model Approach to Knowledge Testing, *Applied Informatics*, 2011, no. 5(35), pp. 60–67.
13. Zaozerskaya, L.A. and Platonova, V.A., Mathematical Models to Form an Optimal Set of Test Structures for Knowledge Control, *Omsk Scientific Bulletin*, 2012, no. 3, pp. 33–36.
14. Campi, M.C., Garatti, S., and Prandini, M., The Scenario Approach for Systems and Control Design, *Ann. Rev. Control*, 2009, vol. 33, no. 2, pp. 149–157.
15. Chi, Z., Liu, Y., Turrini, A., Zhang, L., and Jansen, D.N., A Scenario Approach for Parametric Markov Decision Processes, in *Principles of Verification: Cycling the Probabilistic Landscape: Essay Dedicated to Joost-Pieter Katoen on the Occasion of His 60th Birthday, Part II*, Cham: Springer, 2024, pp. 234–266.
16. Boltzmann, L., *Vorlesungen uber Gastheory*, Leipzig, 1896, vol. 1, J.A. Barth; 1898, vol. 2, J.A. Barth.
17. Jaynes, E.T., Information Theory and Statistical Mechanics, *Physical Review*, 1957, vol. 106, no. 4, pp. 620–630.
18. Jaynes, E.T., Gibbs vs. Boltzmann Entropy, *American Journal of Physics*, 1965, vol. 33, pp. 391–398.
19. Rosenkrantz, R.D., *E.T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, Kluwer Academic Publishers, 1989.
20. Jaynes, E.T., *Probability Theory: the Logic of Science*, Cambridge: Cambridge Univ. Press, 2003.

21. Popkov, Y.S., Asymptotic Efficiency of Maximum Entropy Estimates, *Dokl. Math.*, 2020, vol. 102, pp. 350–352.

22. Popkov, Yu.S., Popkov, A.Yu., and Dubnov, Yu.A., *Entropy Randomization in Machine Learning*, CRC Press, 2023.

23. Krasnosel'skii, M.A., Vainikko, G.M., Zabreiko, P.P., Rutitski, Ja.B., and Stecenko, V.Ja., *Approximated Solutions of Operator Equations*, Groningen: Walters-Noordhoff, 1972.

24. Malkin, I.G., *Some Problems in the Theory of Nonlinear Oscillation*, U.S. Atomic Energy Commission, Technical Information Service, 1959.

25. Darkhovsky, B.S., Popkov, Y.S., Popkov, A.Y., and Aliev, A.S., A Method of Generating Random Vectors with a Given Probability Density Function, *Autom. Remote Control*, 2018, vol. 79, no. 9, pp. 1569–1581. https://doi.org/10.1134/S0005117918090035

26. Avellaneda, M., Minimum-Relative-Entropy Calibration of Asset-Pricing Models, *International Journal of Theoretical and Applied Finance*, 1998, vol. 1, no. 04, pp. 447–472.

27. Jackwerth, J.C., Recovering Risk Aversion from Option Prices and Realized Returns, *The Review of Financial Studies*, 2000, vol. 13, no. 2, pp. 433–451.

28. Ant-Sahalia, Y. and Lo, A.W., Nonparametric Risk Management and Implied Risk Aversion, *Journal of Economics*, 2000, vol. 94, no. 1–2, pp. 9–51.

29. Popkov, Yu.S., Analytic Method for Solving One Class of Nonlinear Equations, *Dokl. Math.*, 2024, vol. 110, pp. 404–407. https://doi.org/10.1134/S1064562424601392

30. Fikhtengol'ts, G.M., *Kurs differentsial'nogo i integral'nogo ischisleniya* (Course on Differential and Integral Calculus), Moscow: Fizmatgiz, 1962.

31. Feller, W., *An Introduction to Probability Theory and Its Applications*, Wiley, 1968.

32. Sobol, I.M., *The Monte Carlo Method*, Moscow: Mir Publishers, 1975.

33. Bakhvalov, N.S., Zhidkov, N.P., and Kobel'kov, G.M., *Chislennye metody* (Numerical Methods), Moscow: Binom, 2003.

This paper was recommended for publication by O.N. Granichin, a member of the Editorial Board