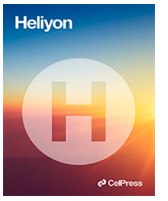
Heliyon 10 (2024) e24708

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Entropy-metric estimation of the small data models with stochastic parameters

Viacheslav Kovtun [a,*], Torki Altameem [b], Mohammed Al-Maitah [b], Wojciech Kempa [c]

[a] Department of Computer Control Systems, Vinnytsia National Technical University, Khmelnitske Shose Str., 95, Vinnytsia, 21000, Ukraine
[b] Computer Science Department, Community College, King Saud University, 11451, Riyadh, Saudi Arabia
[c] Department of Mathematics Applications and Methods for Artificial Intelligence, Silesian University of Technology, Ul. Akademicka 2A, 44-100, Gliwice, Poland

ARTICLE INFO

Keywords:
Information entropy
Machine learning
Small data model
Probability density functions estimation
Static stochastic model
Dynamic stochastic model
Parametric optimization

ABSTRACT

The formalization of dependencies between datasets, taking into account specific hypotheses about data properties, is a constantly relevant task, which is especially acute when it comes to small data. The aim of the study is to formalize the procedure for calculating optimal estimates of probability density functions of parameters of linear and nonlinear dynamic and static small data models, created taking into account specific hypotheses regarding the properties of the studied object. The research methodology includes probability theory and mathematical statistics, information theory, evaluation theory, and stochastic mathematical programming methods. The mathematical apparatus presented in the article is based on the principle of maximization of information entropy on sets determined as a result of a small number of censored measurements of "input" and "output" entities in the presence of noise. These data structures became the basis for the formalization of linear and nonlinear dynamic and static models of small data with stochastic parameters, which include both controlled and noise-oriented input and output measurement entities. For all variants of the above-mentioned small data models, the tasks of determining the optimal estimates of the probability density functions of the parameters were carried out. Formulated optimization problems are reduced to the forms canonical for the stochastic linear programming problem with probabilistic constraints.

1. Introduction

As Pierre-Simon Laplace wrote at the beginning of the 19th century: "Probability theory is common sense expressed in calculations." This idiom well defines the essence of machine learning [1–5]. It is possible to create machine learning algorithms without relying on the probability theory and Bayesian inference, but only knowledge of these scientific disciplines will ensure the clarity, validity and effectiveness of the created constricts. However, in machine learning, any theories are built on rather shaky ground, since machine learning is not abstract mathematics, but the science of applying learning algorithms to real data. Any mathematical theory is based on axioms and conditions (for example, in statistics, one of the cornerstone conditions is "independence and the same distribution of training examples in the sample"). In the context of real data, these conditions may be fulfilled to a limited extent or not at all.

* Corresponding author.
E-mail addresses: kovtun_v_v@vntu.edu.ua (V. Kovtun), altameem@ksu.edu.sa (T. Altameem), malmaitah@ksu.edu.sa (M. Al-Maitah), Wojciech.Kempa@polsl.pl (W. Kempa).

https://doi.org/10.1016/j.heliyon.2024.e24708
Received 26 June 2023; Received in revised form 27 December 2023; Accepted 12 January 2024
Available online 17 January 2024

Despite this, machine learning and statistical inference [6–9] are closely related and focused on solving practically equivalent problems. Thus, machine learning is focused on creating algorithms that train from data and automatically derive general patterns from individual examples, and statistical inference is focused on estimating the parameters distribution and testing hypotheses based on observations (that is, also obtaining general conclusions from individual examples). What then is the difference between these two branches of science? Currently, there is no unanimously recognized answer. Some say [10,11] that in statistics the goal is usually inference about whether the hypothesis is true or how the variables are related, and in machine learning the goal is prediction or generation [12,13]. It should also be noted that in machine learning, as a rule, more complex models are used, while in traditional statistics, models are usually simpler, but due to this, they are also more interpretable.

1.1. Related work and motivation

In classification and regression problems (and many other typical machine learning problems), you need to define a model to describe the relationship between input data X and output data Y as a function $Y = F(X)$. An ordinary model has parameters that are selected during training, so y can be interpreted as a function of the input x and the parameters α: $y = f(x,\alpha)$. Since there are usually many parameters, α is a vector, not a number. The general idea of stochastic modelling is that instead of a point, the model should predict a conditional probability distribution $p(y|x)$ on a set Y for a defined value $x \in X$. Since the stochastic model also has parameters, it is presented as $p(y|x,\alpha)$. With this interpretation, we allow the model to “doubt" when forecasting. We do not lose anything: the probability distribution $p(y|x,\alpha)$ contains more information than the point estimate $f(x,\alpha)$. The function $p(y|x,\alpha)$ is defined $\forall y \in Y$, so we can quantify the error of the model: the lower the probability that the model characterizes the “true" y, the greater the error. Thus, the loss function can be determined naturally.

Suppose that each x stochastic variable y is normally distributed, and the mathematical expectation of the distribution is a function x to be determined (if we achieve this, we can calculate the probability of any value of y for x and α). At the same time, for each educational example: the smaller the probability of the true value of y for defined x and α, the higher the prediction error for this example. Accordingly, we describe the loss function simply: we need to maximize $p(y|x,\alpha)$. The procedure for finding values of parameters α, at which the data probability (likelihood) $p(y|x,\alpha)$ is maximal, is known as the Maximum Likelihood Estimation (MLE) method [14–18]. However, we postulated that we use a normal distribution. Is this choice justified? The Central Limit theorem [19] states that if a stochastic quantity y is the sum of a set of independent stochastic factors, and each factor makes a vanishingly small contribution to this sum, then the quantity y is approximately normally distributed, which indicates the combinatorial nature of the normal distribution. However, even from this definition, we can conclude that everything is far from clear for small data.

We have already mentioned the hypothesis that a data sample summarizes Independent and Identically Distributed (IID) values [20–23], characteristic of both machine learning and mathematical statistics. For a large test sample, such an estimate will be quite accurate. However, there are exceptions, even with a large size sample (initial and test) and formal compliance with IID, the quality assessment will have a high error. This can happen when the impact of individual instances on the quality metric is comparable to the combined impact of all other instances (this is typical for unbalanced classes or strong outliers). In cases where it is obvious that the IID hypothesis is limited, other approaches are used [24–29]: Markov chains, autoregressive models, stochastic decision trees, etc. However, there are borderline cases where it is not clear whether IID training can be applied or not. The above information allows us to state that the formalization of dependencies between datasets, taking into account specific hypotheses regarding data properties (for example, a small data phenomenon), is a constantly relevant task of machine learning. This study is devoted to the development of the author’s concept of response to these challenges.

1.2. Scientific research attributes

The ***object*** of the study is the process of restoring dependencies between data sets determined as a result of a small number of censored measurements of “input" and “output" entities in the presence of interference.

The ***subject*** of study includes probability theory and mathematical statistics, information theory, evaluation theory, and stochastic mathematical programming methods.

The ***purpose*** of the study is to formalize the procedure for calculating optimal estimates of probability density functions of parameters of linear and nonlinear dynamic and static small data models, created taking into account specific hypotheses regarding the properties of the studied object.

The ***objectives*** of the study are.

- based on the target data structures, formalize linear and nonlinear dynamic and static small data models with stochastic parameters, among which there are both controlled and interference-oriented measurements of “input" and “output" entities;
- formalize the procedure for determining optimal probabilistic characteristics for certain types of small data models with stochastic parameters;
- to study the structural features of probability density functions of controlled parameters and interferences as part of formalized options for implementing stochastic small data models;
- demonstrate the functionality of the proposed mathematical apparatus with examples and justify its adequacy in the process of discussing the obtained results.

The ***main contribution***. The mathematical apparatus presented in the article is based on the principle of maximization of

information entropy on sets determined as a result of a small number of censored measurements of “input" and “output" entities in the presence of noise. These data structures became the basis for the formalization of linear and nonlinear dynamic and static models of small data with stochastic parameters, which include both controlled and noise-oriented input and output measurement entities. For all variants of the above-mentioned small data models, the tasks of determining the optimal estimates of the probability density functions of the parameters (taking into account the variants of both the normalized and the interval representation of the corresponding probabilities) were carried out. Formulated optimization problems are reduced to the forms canonical for the stochastic linear programming problem with probabilistic constraints. This made it possible to present solutions in an analytical form. We note that the formalization of the mentioned optimization problems with an orientation towards maximizing the information entropy guarantees obtaining the best solutions in the conditions of the maximum uncertainty of both the stochastic controlled parameters and measurement noises of the models created.

The ***highlights*** of the study are.

- linear and non-linear dynamic and static small data models,
- a lexicographic technique for bringing a nonlinear dynamic stochastic small data model to a linear form,
- analytically formalized probabilistic characteristics of small data models with stochastic parameters,
- analytical description of the structural features of optimal probability densities functions of stochastic small data models parameters.

2. Materials and methods

2.1. Statement of the research

Suppose that as a result of a controlled experiment, the results of measurements of the investigated process were obtained in the form of a matrix of stochastic input parameters X with dimension $(n \times m)$ and a vector of stochastic output parameters y with length n, where n is the number of measurement sessions, m is the number of input parameters.

The connection of the “input" and “output" entities is characterized in the first approximation by a static small data model of the form

$$w = F(X + \mu, \alpha) + \xi, \tag{1}$$

where F is a deterministic n-dimensional vector function; α is a vector of controlled parameters of length m, formed by independent stochastic components of $\alpha_j \in [\alpha_j^-, \alpha_j^+] = \mathrm{A}_j$, $j = \overline{1,m}$; $\mathrm{M} = (\mu_{ij})$ is a matrix with dimension $(n \times m)$, which models the measurement errors of the “input" entity, where $\mu_{ij} \in [\mu_{ij}^-, \mu_{ij}^+] = \mathrm{M}_{ij}$ are independent stochastic elements, $i = \overline{1,n}$, $j = \overline{1,m}$; $\xi = (\xi_i)$ is a vector of length n that models the measurement errors of the “output" entity, where $\xi_i \in [\xi_i^-, \xi_i^+] = \Xi_i$ are independent stochastic components, $i = \overline{1,n}$.

The linear version of the static model [1] will be represented as

$$w = (X + \mu)\alpha + \xi. \tag{2}$$

The discrete dynamic model of the investigated process with finite memory s summarizes the results of accurate measurements of the “input" entity (parameter $x[k]$) and the results of measurements of the “output" entity (parameter $y[k]$), which are determined with additive interference $\xi[k]$, where $k \in T$, $T = [s, s+n]$, and n is the number of measurement sessions.

We present the discrete dynamic small data model for the above interpretation of the studied process based on the Volterra polynomial [30,31] as

$$w[k] = \sum_{l=1}^{R} \sum_{(m_1,\ldots,m_l)=0}^{s} \left(v^{(l)}[m_1,\ldots,m_l] \prod_{r=1}^{l} x[k - m_r] \right) + \xi[k]. \tag{3}$$

Expression [3] is a discrete functional version of the Volterra polynomial of the R th degree with impulse characteristics m_r, $v^{(l)}$, and $m_r \in [0,s]$, $r \in [1,l]$; $v^{(l)}[m_1,\ldots,m_l] = 0 \ \forall m_r > s$.

Stochastic functions $v^{(l)}[m_1,\ldots,m_l]$ with independent ordinates that take values in the range of

$$V^{(l)} = \left[b_-^{(l)} \exp\left(-a_-^{(l)} \sum_{i=1}^{l} m_i \right), b_+^{(l)} \exp\left(-a_+^{(l)} \sum_{i=1}^{l} m_i \right) \right], \tag{4}$$

where $a_\pm^{(l)}$, $b_\pm^{(l)}$ are constants.

Let’s perform the lexicographic ordering [32] of the $\{m_1,\ldots,m_l\}$ variables to move to the linear form of the discrete dynamic model of the studied process (at the current stage, it is non-linear, due to the non-linearity of relation [3]). The resulting sets are reindexed from 0 to $t_l = (s+1)^l - 1$ (according to the lexicographic rule). Let’s enter the local index $j^{(l)} \in [0, t_l]$ and the sequence

$$\{m_1,\ldots,m_l\} \rightarrow j^{(l)}. \tag{5}$$

On this basis, we introduce the indexing of the stochastic parameters corresponding to the values of the impulse weighting functions $v^{(l)}[m_1,\ldots,m_l]$ from Ref. [3] as

$$\alpha^{(l)}_{j^{(l)}} \rightarrow v^{(l)}[m_1,\ldots,m_l]. \tag{6}$$

For $j^{(l)} \in [0,t_l]$, we generalize the obtained values as a vector $\alpha^{(l)} = \{\alpha^{(l)}_0,\ldots,\alpha^{(l)}_{t_l}\}$, the components of which take values in the interval $A^{(l)}(j^{(l)}) = [\alpha^{(l)}_-(j^{(l)}),\alpha^{(l)}_+(j^{(l)})]$, where (according to Ref. [4]):

$$\alpha^{(l)}_-\left(j^{(l)}\right) = b^{(l)}_- \exp\left[-j^{(l)}a^{(l)}_-\right], \alpha^{(l)}_+\left(j^{(l)}\right) = b^{(l)}_- \exp\left[-j^{(l)}a^{(l)}_+\right]. \tag{7}$$

By analogy with [5], let's perform a lexicographic ordering of the variables $\{k-m_1,\ldots,k-m_l\}$, where k there is a fixed parameter, and the indices $\{m_1,\ldots,m_l\}$ take values in the interval $[0,s]$. For each value k, we reindex the obtained sets according to Ref. [5] $\{k-m_1,\ldots,k-m_l\} \rightarrow (k,j^{(l)})$. Then the products of variables $x[k-m_r]$ for a fixed value k mentioned in expression [3] form the vector

$$x^{(l)}[k] = \left\{x^{(l)}_{k,0},\ldots,x^{(l)}_{k,t_l}\right\} = x^{(l)}[s+k], k \in [0,n]. \tag{8}$$

Taking into account the lexicographic transformations described above, expression [3] can be represented in linear form as

$$w[k] = \sum_{l=1}^{R} \langle \alpha^{(l)}, x^{(l)}[k]\rangle + \xi[k], w[k] = w[s+k], \xi[k] = \xi[s+k], k \in [0,n]. \tag{9}$$

Let the measurements of the entities "input" and "output" take place at moments $s+k, k \in [0,n]$, during the censored interval $\mathrm{T} = [s, s+n]$. In the context of the model [3], the results of the measurement of the "output" entity are embodied in vectors $w = \{w[0],\ldots, w[n]\}$, $\xi = \{\xi[0],\ldots,\xi[n]\}$. In the context of expression [8], the results of the measurement of the entity "input" are embodied in a matrix, the rows of which consist of vectors $X^{(l)} = [x^{(l)}(k), k \in [0,n]]$, $l \in [1,R]$, and a block matrix $X = [X^{(1)},\ldots,X^{(R)}]$ of dimension $(n+1)\times u$, where $u = \sum_{l=1}^{R}(t_l+1)$. We form a block vector of stochastic controlled parameters $\alpha = \{\alpha^{(1)},\ldots,\alpha^{(R)}\}$ of length R. Let's summarize the material of the paragraph by presenting a nonlinear dynamic model of the studied process [3] with stochastic initial parameters $\langle \alpha, \xi\rangle$ in a linear form as

$$w = X\alpha + \xi. \tag{10}$$

Note that although models [2,10] are structurally close, the process of formation of all their component parameters is fundamentally different (see section material).

2.2. Probabilistic characteristics of a small data model with stochastic parameters

In this section, we will focus on the analytical formalization of the probability density functions of the parameters of models [1,10]. The presentation format of these models proposed in Section 2.1 allows for two levels of formalization of the probabilistic characteristics of their parameters.

- at the level of probability density functions (*PDD* level),
- at the level of probabilities of parameter values falling into the corresponding intervals (*PHI* level).

We formalize the probabilistic characteristics of the static (*SM*) and dynamic (*DM*) variants of the small data model with stochastic parameters at the *PDD* level. Accordingly, the controlled parameters and interferences in the composition of the small data model will be interpreted as stochastic quantities, the values of which belong to the corresponding intervals on which there are probability density functions of the form

$$P_{SM}(\alpha) = \prod_{i=1}^{m} p_i(\alpha_i), \alpha_i \in \mathrm{A}_i; \tag{11}$$

$$P_{DM}(\alpha) = \prod_{l=1}^{R}\prod_{i=1}^{t_l} p^{(l)}_i\left(\alpha^{(l)}_i\right); \tag{12}$$

$$V(\mu) = \prod_{i=1}^{n}\prod_{j=1}^{m} v_{ij}\left(\mu_{ij}\right), \mu_{ij} \in \mathrm{M}_{ij}; \tag{13}$$

$$Q(\xi) = \prod_{i=1}^{n} q_i(\xi_i), \xi_i \in \Xi_i, \tag{14}$$

where expressions [13,14] characterize the measurement interferences of the "input" and "output" entities, respectively.

A natural way to evaluate functions [11]-(14) is to process the results of measurements of the "input" and "output" entities, as well

as taking into account a priori information presented in the form of probability density functions $P^0(\alpha)$, $V^0(\mu)$, $Q^0(\xi)$. Stochastic small data models generate sets of W stochastic vectors [1,10]. They correspond to the measurement results summarized by the vector y. To estimate the probability density functions [11]-(14), we characterize sets of stochastic elements of the vector w by moments $\mathrm{m}^{(k)} = \{M(w_1^{(k)}),\ldots,M(w_n^{(k)})\}$, where k is the order of the moments, and

$$M\left(w_i^{(k)}\right) = \int\limits_{\alpha\in \mathrm{A},\mu\in \mathrm{M},\xi\in\Xi_i} (F_i[X+\mu,\alpha]+\xi_i)^k P(\alpha)V(\mu)Q(\xi)d\alpha d\mu d\xi, i=\overline{1,n}. \tag{15}$$

Let's limit ourselves to $k = 1$, that is, the mathematical expectation of the components of the vector w. We apply operation [15] to expression [1], which characterizes *SM*: $M(w) = \overline{w} = \int\limits_{\alpha\in \mathrm{A},\mu\in \mathrm{M},\xi\in\Xi} (F[X+\mu,\alpha]+\xi)P(\alpha)V(\mu)Q(\xi)d\alpha d\mu d\xi$. Let's apply operation [15] to expression [10], which characterizes *DM*: $M(w) = \overline{w} = X\int\limits_{\alpha\in \mathrm{A}} \alpha P(\alpha) + \int\limits_{\xi\in\Xi} \xi Q(\xi)d\xi$.

Now let's examine the probabilistic characteristics of static and dynamic small data models with stochastic parameters at the *PHI* level. This means that both *SM*, *DM* parameters are continuous stochastic quantities whose belonging to the corresponding intervals is characterized by certain probabilities. Controlled parameters $\alpha = \{\alpha_1,\ldots,\alpha_m\}$ take values in intervals $\mathrm{A} = \{\mathrm{A}_1,\ldots,\mathrm{A}_m\}$ with probabilities $p = \{p_1,\ldots,p_m\}$, $p_j \in [0,1]$, $j = \overline{1,m}$, respectively. Similarly, measurement interferences of entity "input" μ_{ij} take values in intervals M_{ij} with probabilities v_{ij}, $v_{ij} \in [0,1]$, $i = \overline{1,n}$, $j = \overline{1,m}$, and measurement interferences of entity "output" take values in intervals Ξ_i with probabilities q_i, $q_i \in [0,1]$, $i = \overline{1,n}$. We will also mention the corresponding a priori probabilities: p_j^0, w_{ij}^0, q_i^0, $i = \overline{1,n}$, $j = \overline{1,m}$.

The studied small data models reproduce a set W of vectors w generating stochastic values of controlled parameters and interferences with probabilities p, v, q, respectively. Let us characterize the set generated by the models using first-order quasimomentum vectors:

$$\alpha = \alpha^- + L_\alpha p, \mu = \mu^- + L_\mu \otimes V, \xi = \xi^- + L_\xi q, \tag{16}$$

where $L_\alpha = diag[(\alpha_i^+ - \alpha_i^-)|i=\overline{1,m}]$, $L_\mu = diag[(\mu_{ij}^+ - \mu_{ij}^-)|i=\overline{1,n}.j=\overline{1,m}]$, $L_\xi = diag[(\xi_i^+ - \xi_i^-)|i=\overline{1,n}]$, and the sign $\otimes$ symbolizes the element-by-element multiplication operation.

Substitute quasi-average values [16] into expression [1], which characterizes *SM*, and obtains $\widetilde{w} = F[X + (\mu^- + L_\mu \otimes V), \alpha^- + L_\alpha p] + \xi^- + L_\xi q$. Substitute quasi-average values [16] into expression [10], which characterizes *DM*, and obtains

$$\widetilde{w} = X\alpha + L_\alpha + \xi^- + L_\xi q. \tag{17}$$

Therefore, when studying stochastic small data models at the *PDD* level, the corresponding probability density functions should be estimated. When choosing the *PHI* level, the vectors that characterize the corresponding probability distributions are subject to evaluation. For both levels, it is appropriate to introduce the likelihood functional as a measure of the quality of the estimates of the probability density functions. Let us focus on the analytical formalization of such a functional.

Let's determine the compatible probability density functions of the controlled parameters and the interferences of the small data model, taking into account their independence: $\Phi(\alpha,\mu,\xi) = P(\alpha)V(\mu)Q(\xi)$. Taking into account the given a priori probabilities $P^0(\alpha)$, $V^0(\mu)$, $Q^0(\xi)$, we introduce the logarithmic likelihood ratio of the form $\varphi(\alpha,\mu,\xi) = \ln\frac{P(\alpha)}{P^0(\alpha)} + \ln\frac{V(\mu)}{V^0(\mu)} + \ln\frac{Q(\xi)}{Q^0(\xi)}$. The resulting function is a deterministic function of stochastic arguments. In the context of the *PHI* level, the analogue of the function $\varphi(\alpha,\mu,\xi)$ is the function $\varphi(p, V, q) = \ln\frac{p}{p^0} + \ln\frac{V}{V^0} + \ln\frac{q}{q^0}$, where p^0, V^0, q^0 are the a priori probabilities.

We define the desired likelihood functional based on Shannon's information entropy [33] as

$$H[P(\alpha),V(\mu),Q(\xi)] = -L[P(\alpha),V(\mu),Q(\xi)] =$$
$$= \int\limits_{\alpha\in \mathrm{A},\mu\in \mathrm{M},\xi\in\Xi} \varphi(\alpha,\mu,\xi)\Phi(\alpha,\mu,\xi)d\alpha d\mu d\xi = \int\limits_{\alpha\in \mathrm{A}} P(\alpha)\ln\frac{P(\alpha)}{P^0(\alpha)}d\alpha + \int\limits_{\mu\in \mathrm{M}} V(\mu)\ln\frac{V(\mu)}{V^0(\mu)}d\mu + \int\limits_{\xi\in\Xi} Q(\xi)\ln\frac{Q(\xi)}{Q^0(\xi)}d\xi.$$

The resulting functional is interpreted both as a measure of the distance between probability density functions and as a measure of the degree of invariance of functions $P(\alpha)$, $V(\mu)$, $Q(\xi)$ concerning the observer. Taking into account the fact of independence of parameters α, μ, ξ established by expressions [11]-(14), we redefine the functional $L[P(\alpha),V(\mu),Q(\xi)]$ as

$$L[P(\alpha),V(\mu),Q(\xi)] = -\sum_{j=1}^{m}\int\limits_{\alpha\in \mathrm{A}} p_j(\alpha_j)\ln\frac{p_j(\alpha_j)}{p_j^0(\alpha_j)}d\alpha_j - -\sum_{i=1}^{n}\sum_{j=1}^{m}\int\limits_{\mu_{ij}\in \mathrm{M}_{ij}} v_{ij}(\mu_{ij})\ln\frac{v_{ij}(\mu_{ij})}{v_{ij}^0(\mu_{ij})}d\mu_{ij} - \sum_{i=1}^{n}\int\limits_{\xi_i\in\Xi_i} q_i(\xi_i)\ln\frac{q_i(\xi_i)}{q_i^0(\xi_i)}d\xi_i.. \tag{18}$$

Based on the analytical form of the functional [18], we define the likelihood function as

$$L(p,V,q) = -\sum_{j=1}^{m} p_j \ln\frac{p_j}{p_j^0} - \sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij}\ln\frac{v_{ij}}{v_{ij}^0} - \sum_{i=1}^{n} q_i \ln\frac{q_i}{q_i^0}. \tag{19}$$

The quality of estimating the probability density functions using the functional [18] or the components of the probability vector [19] increases as the estimates approach the maximum value of [18]. Let's develop this concept into the statement of the

corresponding optimization problem with the objective function

$$L[P(\alpha), V(\mu), Q(\xi)] \to \max, \tag{20}$$

where the functional $L[P(\alpha), V(\mu), Q(\xi)]$ is described by expression [19].

Among the constraints of the optimization problem [20], we note both the condition that the probability density functions must belong to the space

$$\mathrm{D} = \mathrm{P} \cup \mathrm{V} \cup \mathrm{Q}, \tag{21}$$

where P, V, Q are the spaces of the probability densities of the controlled parameters and measurement interferences of the entities "input" and "output" of the small data model, respectively, and the condition that maintains a balance between the k^{-1}-th power of the k-th moment of vector w (the result of modelling) and vector y (the result of measurements).

Considering that the formalization of the vector w is determined by the type of small data model, the second constraint of the optimization problem [20] is represented by the corresponding expressions:

$$SM: \left(M\left\{F^{(k)}[(X+\mu),\alpha] + \xi^{(k)}\right\}\right)^{\frac{1}{k}} = y, \tag{22}$$

$$DM: \left(M\{X\alpha + \xi\}\right)^{\frac{1}{k}} = y. \tag{23}$$

2.3. Structural features of probability density functions of small data model parameters

Let us characterize the situation when the model SM is of a power-law type:

$$w(t) = \sum_{l=1}^{R} \sum_{j=1}^{m} \alpha_j^l x_j^{(l)}(t) + \xi(t), \tag{24}$$

where $\alpha = \{\alpha_j\}$ is a stochastic vector of controlled parameters formed by independent components that take values in intervals $\mathrm{A}_j = [\alpha_j^-, \alpha_j^+]$ with probability density functions $p(\alpha) = \{p_j(\alpha_j)\}$, $j = \overline{1,m}$.

The measurement of the entities "input" and "output" is carried out at moments $\{t_i\}$, $i = \overline{1,n}$. The result of measuring the entity "input" is a set of matrices $\{X^{(l)}\}$, $l = \overline{1,R}$: $X^{(l)} = \begin{pmatrix} x_1^{(l)}(t_1) & \dots & x_m^{(l)}(t_1) \\ \vdots & \ddots & \vdots \\ x_1^{(l)}(t_n) & \dots & x_m^{(l)}(t_n) \end{pmatrix} = \begin{pmatrix} x_{1,1}^{(l)} & \dots & x_{1,m}^{(l)} \\ \vdots & \ddots & \vdots \\ x_{n,1}^{(l)} & \dots & x_{n,m}^{(l)} \end{pmatrix}$. The result of the measurement of the "output" entity is a stochastic vector $w = \{w(t_i)\}$, $i = \overline{1,n}$.

Taking into account what has been introduced, let's present the model [24] as

$$w = \sum_{l=1}^{R} X^{(l)} \alpha^{(l)} + \xi, \tag{25}$$

where $\alpha^{(l)} = \{\alpha_j^{(l)}\}$, $j = \overline{1,m}$; $\xi = \{\xi(t_i)\} = \{\xi_i\}$ is a vector of measurement interferences of the "output" entity, whose independent components take values from intervals $\Xi_i = [\xi_i^-, \xi_i^+]$ with probability density functions $\{q_i(\xi_i)\}$, $i = \overline{1,n}$. A priori information is characterized as $P^0(\alpha) = Q^0(\xi) = const$.

Let us formalize the entropy estimation of the probability density functions $P(\alpha) = \{p_j(\alpha_j)\}$, $j = \overline{1,m}$, and $Q(\xi) = \{q_i(\xi_i)\}$, $i = \overline{1,n}$, in the context of the model [25] and investigate its structural properties. We will use the formulation of the optimization problem [20] as a basis. Therefore, we obtain the objective function

$$L[P(\alpha), Q(\xi)] = -\sum_{j=1}^{m} \int\limits_{\alpha_j \in \mathrm{A}_j} p_j(\alpha_j) \ln p_j(\alpha_j) d\alpha_j - \sum_{i=1}^{n} \int\limits_{\xi_i \in \Xi_i} q_i(\xi_i) \ln q_i(\xi_i) dq_i \to \max \tag{26}$$

and the constraints

$$1 - \int\limits_{\alpha_j \in \mathrm{A}_j} p_j(\alpha_j) d\alpha_j = 0, j = \overline{1,m}; \tag{27}$$

$$1 - \int\limits_{\xi_i \in \Xi_i} q_i(\xi_i) d\xi_i = 0, i = \overline{1,n}; \tag{28}$$

$$\Phi_i[p(\alpha),q(\xi)]=\sum_{l=1}^{R}\sum_{j=1}^{m}x_{il}^{(l)}\int\limits_{\alpha_j\in A_j}\alpha_j^l p_j(\alpha_j)+\int\limits_{\xi_i\in\Xi_i}\xi_i q_i(\xi_i)d\xi_i=y_i,\ i=\overline{1,n}. \tag{29}$$

The structural properties of functions $P(\alpha)$ and $Q(\xi)$ are regulated by their class. To determine the latter in the context of *SM*, the optimization problem [20–22] should be solved, and in the context of *DM*, the optimization problem [20,21,23] should be solved. In both cases, these are stochastic linear programming problems with probabilistic equality constraints [34,35]. Having implemented in an analytical form the standard approach to solving such problems (see the authors' previous work on this topic [33]) we obtained optimal probability density functions classified as continuous differential functions, namely:

$$p_j^*(\alpha_i)\sim\beta_j^{(1)}\exp\left(-\sum_{l=1}^{R}\beta_{jl}^{(2)}\alpha_j^l\right),q_i^*(\xi_i)\sim\beta_i^{(3)}\exp\left(-\theta_i^{(1)}\xi_i\right), \tag{30}$$

where $\beta_j^{(1)}=\exp(-1-\theta_j^{(2)})$, $\beta_{jl}^{(2)}=\sum_{i=1}^{n}\theta_j^{(1)}x_{ij}^{(l)}$, $\beta_i^{(3)}=\exp(-1-\theta_i^{(3)})$, $j\in[1,m]$, $i\in[1,n]$; $\theta^{(1)}$, $\theta^{(2)}$, $\theta^{(3)}$ are the Lagrange multipliers, and $\theta^{(1)}$ corresponds to constraint [28], and $\theta^{(2)}$, $\theta^{(3)}$ are correspond to constraints [27,28].

For a linear stochastic small data model, the estimation function [30] is always exponential. Empirical data on the "input" and "output" entities affect only the form and not the structural properties of the estimation function [30]. For the nonlinear stochastic small data model, the structural properties of the estimation function [30] are more diverse but do not leave the class of nonlinear continuous functions.

Now let's characterize the situation when the model *DM* is power-law (based on the linear form of the latter defined by expression [17]). The problem of optimal estimation of probability density functions of controlled parameters [12] and interferences [14] in this case is characterized by the objective function

$$L[P(\alpha),Q(\xi)]=-\int\limits_{\alpha\in A}P(\alpha)\ln P(\alpha)d\alpha-\int\limits_{\xi\in\Xi}Q(\xi)\ln Q(\xi)d\xi\to\max \tag{31}$$

and constraints

$$1-\int\limits_{\alpha\in A}P(\alpha)d\alpha=0,1-\int\limits_{\xi\in\Xi}Q(\xi)d\xi=0; \tag{32}$$

$$\Phi[P(\alpha),Q(\xi)]=X\int\limits_{\alpha\in A}\alpha P(\alpha)d\alpha+\int\limits_{\xi\in\Xi}\xi Q(\xi)d\xi=y. \tag{33}$$

Note that the functional [33] is an $(n+1)$-dimensional function. The optimization problem [31]-(33) is of the same type as the optimization problem [26]-(29). Having implemented in analytical form the standard approach to solving such problems, we obtained optimal probability density functions classified as continuous-differential functions, namely:

$$P^*(\alpha)=\exp\left(-1-\theta^{(2)}-\langle\theta^{(1)},X\alpha\rangle\right),Q^*(\xi)=\exp\left(-1-\theta^{(3)}-\theta^{(1)}\otimes\xi\right), \tag{34}$$

where $\theta^{(1)}$ is a vector of Lagrange multipliers of length $(n+1)$, the components of which satisfy the constraint [33]; $\theta^{(2)}$, $\theta^{(3)}$ are vectors of factors whose components satisfy the constraint [32].

From the analytical form of expressions [34], we can conclude that optimal probability density functions with Lagrange multiplier parameters belong to the exponential type. This conclusion is due to the type of functions [31]-(33).

Also worth investigating is the version of the implementation *SM*, where the "input" entity was measured without interference, i.e.:

$$\widetilde{w}=XL_\alpha p+L_\xi q+\Lambda(\alpha^-,\xi^-), \tag{35}$$

where $\Lambda(\alpha^-,\xi^-)=X\alpha^-+\xi^-$. It is also assumed that information about the a priori values of the controlled parameters [35] is available: p^0, q^0. We analytically formalize estimates of probability density functions of these parameters for selected classes of vectors p, q.

The sought estimates are determined as a result of solving the optimization problem with the objective function

$$L(p,q)=-\sum_{j=1}^{m}p_j\ln\frac{p_j}{p_j^0}-\sum_{i=1}^{n}q_i\ln\frac{q_i}{q_i^0}\to\max \tag{36}$$

and constraints aimed at probabilities normalizing and maintaining a balance between the modelled and measured values of the "output" entity: $\sum_{j=1}^{m}p_j=1$, $\sum_{i=1}^{n}q_i=1$; $\sum_{j=1}^{m}x_{ij}L_\alpha^j p_j+L_\xi^i q_i+\Lambda_i=y_i$, $i\in[1,n]$, $n<m$.

The optimization problem [36] can be classified as a stochastic problem of linear programming, the solution of which is formalized in terms of Lagrange functions in compliance with the formulated optimality conditions:

$$H\left(p,q,\theta^{(1)},\theta^{(2)},\theta^{(3)}\right)=L(p,q)+\theta^{(1)}\left(1-\sum_{j=1}^{m}p_j\right)+\theta^{(2)}\left(1-\sum_{i=1}^{n}q_i\right)+\sum_{i=1}^{n}\theta_i^{(3)}\left[y_i-\sum_{j=1}^{m}x_{ij}L_\alpha^j p_j-L_\xi^i q_i-\Lambda_i\right]. \tag{37}$$

In the context of [37], we obtain the following expressions for the sought optimal probabilities p^*, q^*:

$$0 \le p_j^*\left(\theta^{(3)}\right) = \frac{p_j^0 \exp\left(-\sum_{i=1}^{n} \theta_i^{(3)} x_{ij} L_\alpha^j\right)}{\sum_{j=1}^{m} p_j^0 \exp\left(-\sum_{i=1}^{n} \theta_i^{(3)} x_{ij} L_\alpha^j\right)} \le 1, j = \overline{1,m}, \tag{38}$$

$$0 \le q_i^*\left(\theta^{(3)}\right) = \frac{q_i^0 \exp\left(-\theta_i^{(3)} L_\xi^i\right)}{\sum_{i=1}^{n} q_i^0 \exp\left(-\theta_i^{(3)} L_\xi^i\right)} \le 1, i = \overline{1,n}. \tag{39}$$

The values of the Lagrange multipliers $\{\theta_i^{(3)}\}$, $i = \overline{1,n}$, necessary for the calculation of expressions [38,39] are calculated according to $\Phi_i(\theta^{(3)}) = \frac{1}{y_i - \Lambda_i} \sum_{j=1}^{m} x_{ij} L_\alpha^j p_j^*(\theta^{(3)}) + L_\xi^i q_i^*(\theta^{(3)}) = 1$.

Finally, we analytically formalize the estimates of the probability density functions of the controlled parameters of the model [35] for selected classes of vectors p, q considering that

$$0 \le p_j \le 1, j = \overline{1,m};\ 0 \le q_i \le 1, i = \overline{1,n}. \tag{40}$$

$$L(p,q) = -\sum_{j=1}^{m}\left[p_j \ln\frac{p_j}{p_j^0} + (1-p_j)\ln(1-p_j)\right] - \sum_{i=1}^{n}\left[q_i \ln\frac{q_i}{q_i^0} + (1-q_i)\ln(1-q_i)\right] \to \max, \tag{41}$$

where $p_j^0 = p_j^0/(1-p_j^0)$, $q_i^0 = q_i^0/(1-q_i^0)$, and constraints, which include the conditions [40] and the condition focused on maintaining the balance between the modelled and measured values of the "output" entity:

$$\sum_{j=1}^{m} x_{ij} L_\alpha^j p_j + L_\xi^i q_i + \Lambda_i = y_i, i \in [1,n]. \tag{42}$$

We formalize the solution of the stochastic linear optimization problem (41) with constraints [40] and (42) analytically in the form of expressions

$$0 \le p_j^*\left(\theta^{(3)}\right) = \frac{p_j^0}{\left(1-p_j^0\right)\exp\left(\sum_{i=1}^{n} \theta_i^{(3)} x_{ij} L_\alpha^j\right)} \le 1, j = \overline{1,m}, \tag{43}$$

$$0 \le q_i^*\left(\theta^{(3)}\right) = \frac{q_i^0}{q_i^0 + (1-q_i^0)\exp\left(-\theta_i^{(3)} L_\xi^i\right)} \le 1, i = \overline{1,n}. \tag{44}$$

The values of the Lagrange multipliers $\{\theta_i^{(3)}\}$, $i = \overline{1,n}$, necessary for the calculation of expressions (43), and (44) are calculated according to (45).

$$\Phi_i\left(\theta^{(3)}\right) = \frac{1}{y_i - \Lambda_i} \sum_{j=1}^{m} \theta_i^{(3)} x_{ij} p_j^*\left(\theta^{(3)}\right) + L_\xi^i q_i^*\left(\theta^{(3)}\right) = 1. \tag{45}$$

The solution of the system of equation (45) concerning the exponential Lagrange multipliers $g_i = \exp(-\theta_i^{(3)})$, $i \in [1,n]$, is expressed as $g_i^{k+1} = g_i^k \Phi_i(g^k)$, $g_i^0 > 0 \forall i \in [1,n]$.

3. Results

We will demonstrate examples of estimating the parameters of static and dynamic stochastic small data models.

Let there be a static small data model of the form [35] with five stochastic controlled parameters $\alpha = \{\alpha_j\}$, $j = \overline{1,5}$, which take values from the same intervals $A = A_1 = \ldots = A_5 = [0,00; 10,00]$. The available information about the a priori values of the controlled variables is summarized by the vector $\alpha^0 = \{1,00; 2,00; 2,00; 4,00; 1,00\}$. The interference affecting the modelled "output" entity $y = \{y_1, y_2\}$ is characterized by the vector $\xi = \{\xi_1, \xi_2\}$, the components of which belong to the corresponding intervals $\Xi_1 = [-3,00; 3,00]$, $\Xi_2 = [-6,00; 6,00]$. The results of the measurements of the "input" and "output" entities are represented by the values summarized in the matrix $X = \begin{pmatrix} 1,80 & 2,10 & 3,30 & 2,00 & 1,50 \\ 4,10 & 3,80 & 3,00 & 2,80 & 1,90 \end{pmatrix}$ and the vector $y = (21,10; 32,80)$.

Based on the initial data, the quasi-moments [16] for the controlled parameters and interference are described by expressions $\alpha_j = 10p_j, j = \overline{1,5}$; $\xi_1 = 6q_1 - 3$, $\xi_2 = 12q_2 - 6$. Substitute the obtained expressions into the model [35] and obtain $w(p,q) = XL_\alpha p + L_\xi q = \vec{1}$, where $XL_\alpha = \begin{pmatrix} 0,75 & 0,87 & 1,37 & 0,83 & 0,62 \\ 1,06 & 0,98 & 0,77 & 0,72 & 0,45 \end{pmatrix}$, $L_\xi = \begin{pmatrix} 0,25 & 0,00 \\ 0,00 & 0,31 \end{pmatrix}$, $\vec{1} = (1,00 \quad 1,00)$.

Suppose there are five sets of a priori data: $p_1^0 = (1{,}00; 1{,}00; 1{,}00; 1{,}00; 1{,}00)$, $p_2^0 = (0{,}10; 0{,}20; ,0{,}30; 0{,}30; 0{,}10)$, $p_3^0 = (0{,}30; 0{,}40; 0{,}10; 0{,}05; 0{,}15)$, $q_1^0 = (0{,}20; 0{,}80)$, $q_2^0 = (1{,}00; 1{,}00)$. The a priori set $\{p_1^0, q_2^0\}$ is characterized by a uniform distribution of stochastic values of controlled parameters and interference. The a priori set $\{p_2^0, q_1^0\}$ is characterized by a non-uniform distribution of stochastic values of controlled parameters and interference. The a priori set $\{p_3^0, q_2^0\}$ is characteristic of the combined distribution, in which the stochastic values of the controlled parameters are unevenly distributed, and the stochastic values of the interference are uniformly distributed.

Let's formulate the problem of estimating the probability density functions of parameters α, ξ for selected classes of vectors p, q based on the optimization problem [36] (with normalization of probability values p, q). We obtain the statement of the optimization problem $T1$:

$$L(p,q) = -\sum_{j=1}^{5}\left(p_j \ln\left(\frac{p_j}{p_j^0}\right)\right) - \sum_{i=1}^{2}\left(q_i \ln\left(\frac{q_i}{q_i^0}\right)\right) \to \max,$$

$$\sum_{j=1}^{5} p_j = 1, \sum_{i=1}^{2} q_i = 1, p_j, q_i > 0 \forall j \in [1,5], i \in [1,2].$$

$$0{,}75p_1 + 0{,}87p_2 + 1{,}37p_3 + 0{,}83p_4 + 0{,}62p_5 + 0{,}25q_1 = 1,$$

$$1{,}06p_1 + 0{,}98p_2 + 0{,}77p_3 + 0{,}72p_4 + 0{,}45p_5 + 0{,}31q_1 = 1.$$

Let us formulate the problem of estimating the probability density functions of parameters α, ξ for selected classes of vectors p, q based on the optimization problem (41) (with interval values of probabilities p, q of the form [40]). We obtain the statement of the optimization problem $T2$:

$$L(p,q) = -\sum_{j=1}^{5}\left(p_j \ln\left(\frac{p_j}{p_j^0}\right)\right) - \sum_{i=1}^{2}\left(q_i \ln\left(\frac{q_i}{q_i^0}\right)\right) \to \max,$$

$$0 \le p_j \le 1, j = \overline{1,5};\ 0 \le q_j \le 1, i = \overline{1,2}.$$

$$0{,}75p_1 + 0{,}87p_2 + 1{,}37p_3 + 0{,}83p_4 + 0{,}62p_5 + 0{,}25q_1 = 1,$$

$$1{,}06p_1 + 0{,}98p_2 + 0{,}77p_3 + 0{,}72p_4 + 0{,}45p_5 + 0{,}31q_1 = 1.$$

Entropy L reaches its maximum at point $(\widehat{p}_j = 0{,}36p_j^0; \widehat{q}_i = 0{,}36q_i^0), j = \overline{1,5}, i = \overline{1,2}$. Fig. 1 visualized the extreme values of $\widehat{p}$, $\widehat{q}$ depending on the corresponding a priori data sets p^0, q^0: $\{\widehat{p}, \widehat{q}\} = f(\{p^0, q^0\})$.

Based on the presented in Fig. 1 values, we calculate the optimal estimates $p^* = \{p_j^*\}, j = \overline{1,5}$; $q^* = \{q_i^*\}, i = \overline{1,2}$, for $T1$ the problem using expressions [38,39] taking into account the following combinations of a priori data sets: $AS1 : \{p_1^0, q_1^0\}$, $AS2 : \{p_2^0, q_1^0\}$, $AS3 : \{p_3^0, q_1^0\}$, $AS4 : \{p_1^0, q_2^0\}$, $AS5 : \{p_2^0, q_2^0\}$, $AS6 : \{p_3^0, q_2^0\}$, $AS = \{AS_l\}, l = \overline{1,6}$. The result of calculating the dependence $\{p^*, q^*, L^*\} = f(AS)$ is presented graphically in Fig. 2, where the entropy estimate L^* was determined by expression [37].

Corresponding to those presented in Fig. 2 values of $\{p^*, q^*, L^*\}$, the values of $\{\alpha^*, \xi^*\}$, where $\alpha^* = \{\alpha_o^*\}, j = \overline{1,5}$; $\xi^* = \{\xi_i^*\}, i = \overline{1,2}$:

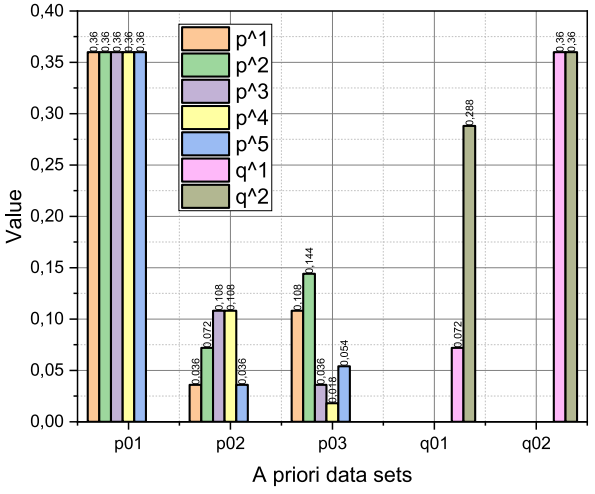


Fig. 1. Visualization of dependence $\{\widehat{p}, \widehat{q}\} = f(\{p^0, q^0\})$.

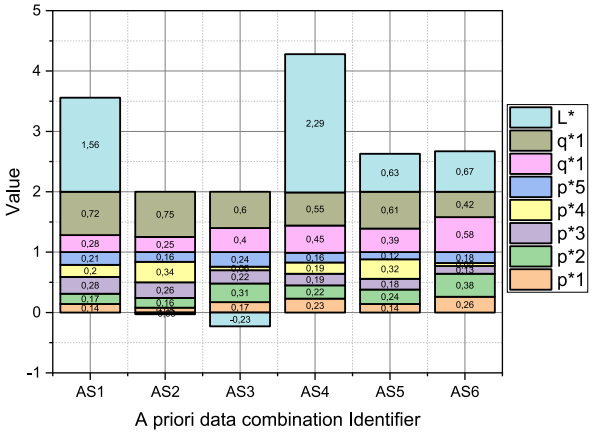


Fig. 2. Visualization of dependence $T1 : \{p^*, q^*, L^*\} = f(AS)$.

$\{\alpha^*, \xi^*, \delta\} = f(AS)$, are presented in Fig. 3. The error δ was calculated using the expression $\delta = \|\alpha^0 - \alpha^*\| / (\|\alpha^0\| + \|\alpha^*\|)$.

Based on the presented in Fig. 1 values, we calculate the optimal estimates p^*, q^* for $T2$ the problem using expressions (43), and (44) taking into account combinations of a priori data sets generalized by the AS. The result of calculating the dependence $\{p^*, q^*, L^*\} = f(AS)$ is presented graphically in Fig. 4, where the entropy estimate L^* was determined by expression (41).

Corresponding to those presented in Fig. 4 value of $\{p^*, q^*, L^*\}$, the values $\{\alpha^*, \xi^*, \delta\}$ are presented in Fig. 5.

Now we will give an example of estimating the density functions of the probability distribution of a dynamic small data model (on the example of a non-linear (power-law) type of the latter). We specify the analytical form of the studied model as $w[k] = \sum_{i=0}^{2} v^{(1)}[i]x[k-i] + \sum_{i,j=0}^{2} v^{(2)}[i,j]x[k-i]x[k-j] + \xi[k]$, $k \geq 2$, where $v^{(1)}[i] = v^{(2)}[i,j]$ if $(i,j) > 2$.

Such a model summarizes nine parameters: $\alpha_0 = \alpha_0^{(1)} = v^{(1)}[0]$, $\alpha_1 = \alpha_1^{(1)} = v^{(1)}[1]$, $\alpha_2 = \alpha_2^{(1)} = v^{(1)}[2]$, $\alpha_3 = \alpha_0^{(2)} = v^{(2)}[0,0]$, $\alpha_4 = \alpha_1^{(2)} = v^{(2)}[0,1] + v^{(2)}[1,0]$, $\alpha_5 = \alpha_2^{(2)} = v^{(2)}[0,2] + v^{(2)}[2,0]$, $\alpha_6 = \alpha_3^{(2)} = v^{(2)}[1,1]$, $\alpha_7 = \alpha_4^{(2)} = v^{(2)}[1,2] + v^{(2)}[2,1]$, $\alpha_8 = \alpha_5^{(2)} = v^{(2)}[2,2]$.

The values of the constants [7] are as follows: $a_-^{(1)} = a_+^{(1)} = a_-^{(2)} = a_+^{(1)} = 0{,}08$, $b_-^{(1)} = 0{,}50$; $b_+^{(1)} = 1{,}00$; $b_-^{(2)} = 1{,}00$; $b_+^{(2)} = 2{,}00$. The intervals [6] to which the components of the vector of controlled parameters $\alpha = \{\alpha_1, \ldots, \alpha_9\}$ belong are defined as $\alpha_1 \in [0{,}50; 1{,}00]$, $\alpha_2 \in [0{,}46; 0{,}92]$, $\alpha_3 \in [0{,}42; 0{,}85]$, $\alpha_4 \in [1{,}00; 2{,}00]$, $\alpha_5 \in [0{,}92; 1{,}84]$, $\alpha_6 \in [0{,}85; 1{,}70]$, $\alpha_7 \in [0{,}85; 1{,}70]$, $\alpha_8 \in [0{,}79; 1{,}58]$, $\alpha_9 \in [0{,}72; 1{,}44]$.

Let two sessions of measurement of “input" and “output" entities be implemented. We formalize blocks [9] as

$$X^{(1)} = \begin{pmatrix} x[2] & x[1] & x[0] \\ x[3] & x[2] & x[1] \end{pmatrix} = \begin{pmatrix} x_{00} & x_{01} & x_{02} \\ x_{10} & x_{11} & x_{12} \end{pmatrix},$$

$$X^{(2)} = \begin{pmatrix} x^2[2] & x[2]x[1] & x[2]x[0] & x^2[1] & x[1]x[0] & x^2[0] \\ x^2[3] & x[3]x[2] & x[3]x[1] & x^2[2] & x[2]x[1] & x^2[1] \end{pmatrix} = \begin{pmatrix} x_{00} & x_{01} & x_{02} & x_{03} & x_{04} & x_{05} \\ x_{10} & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \end{pmatrix}.$$

Accordingly, the general matrix $X = [X^{(1)}, X^{(2)}]$ will look like this:

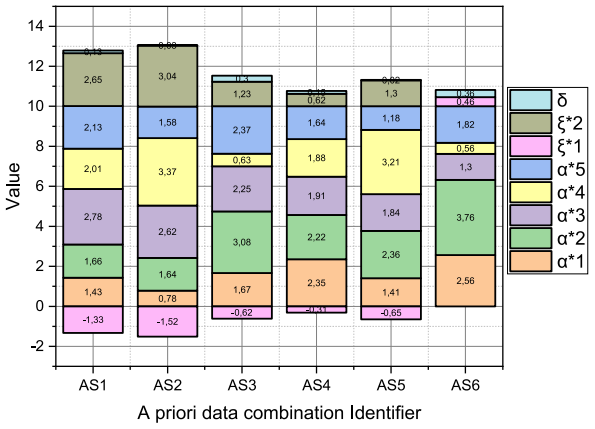


Fig. 3. Visualization of dependence $T1 : \{\alpha^*, \xi^*, \delta\} = f(AS)$.

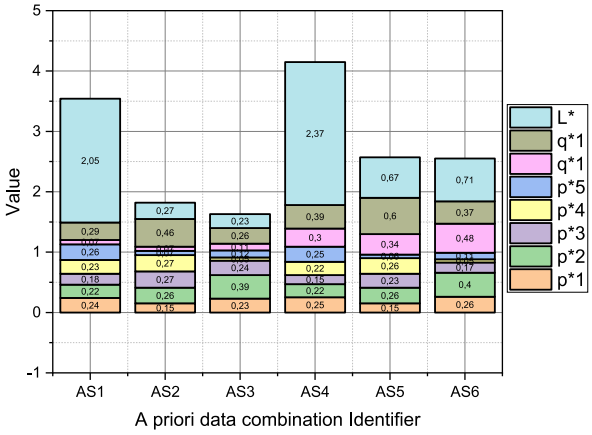


Fig. 4. Visualization of dependence $T2:\{p^*,q^*,L^*\}=f(AS)$.

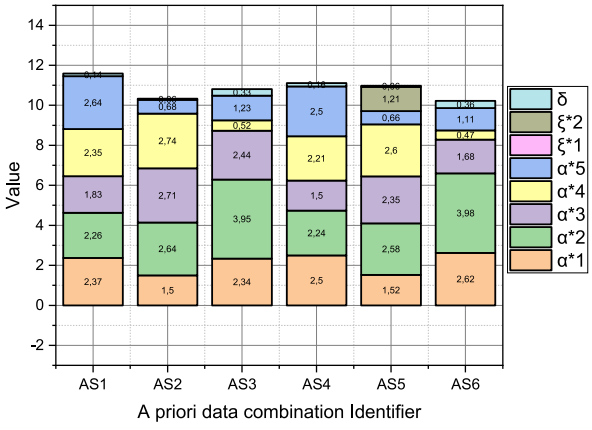


Fig. 5. Visualization of dependence $T2:\{\alpha^*,\xi^*,\delta\}=f(AS)$.

$$X=\begin{pmatrix}3,90 & 1,90 & 2,80 & 0,90 & 1,60 & 5,20 & 3,60 & 1,90 & 4,20\\ 9,30 & 3,90 & 1,90 & 3,80 & 8,50 & 4,90 & 0,90 & 1,60 & 2,60\end{pmatrix}.$$

Let us characterize the interference vectors $[0]=\xi_0\in[-3,00;3,00]$ $\xi[1]=\xi_1\in[-6,00;6,00]$. The measured values of the "output" entity are as follows: $y[2]=y_0=18,52$; $y[3]=y_1=43,35$.

For the initial values described above, we analytically characterize the estimates of the probability density functions of the model parameters in the form of vectors $p^*=\{p_j^*(\alpha_j)\}$, $j=\overline{0,8}$; $q^*=\{q_i^*(\xi_i)\}$, $i=\overline{0,1}$, as $p_j^*(\alpha_j)=\exp(-1-\theta_j^{(1)}-\sum_{l=0}^{1}\theta_l^{(2)}x_{lj}\alpha_j)$, $q_i^*(\xi_i)=\exp[-1-\theta_i^{(3)}-\theta_i^{(2)}\xi_i]$, where the Lagrange multipliers $\{\theta_j^{(1)},\theta_i^{(2)},\theta_l^{(3)}\}$ are determined by the expressions

$$\Theta_j^{(1)}\left(\theta^{(1)},\theta^{(2)}\right)=\int_{\alpha_j^-}^{\alpha_j^+}\exp\left(-1-\theta_j^{(1)}-\sum_{l=0}^{1}\theta_l^{(2)}x_{lj}\alpha_j\right)d\alpha_j=1,$$

$$\Theta_i^{(2)}\left(\theta^{(3)},\theta^{(2)}\right)=\int_{\xi_i^-}^{\xi_i^+}\exp\left(-1-\theta_i^{(3)}-\theta_i^{(2)}\xi_i\right)d\xi_i=1,$$

$$\Phi_i\left(\theta^{(1)},\theta^{(2)},\theta^{(3)}\right)=\sum_{j=0}^{8}x_{ij}\int_{\alpha_j^-}^{\alpha_j^+}\exp\left(-1-\theta_j^{(1)}-\sum_{l=0}^{1}\theta_l^{(2)}x_{lj}\alpha_j\right)d\alpha_j+\int_{\xi_i^-}^{\xi_i^+}\exp\left(-1-\theta_i^{(3)}-\theta_i^{(2)}\xi_i\right)d\xi_i=y_i.$$

The *Symbolic Math Toolbox* of the *Matlab* software package was used for analytical calculations of the above integrals. The *Levenberg-Marquardt* method from the *Matlab* software package (https://www.mathworks.com/help/optim/ug/equation-solving-

algorithms.html) was used to solve the obtained nonlinear equations. As a result, the following values of the Lagrange multipliers were found:

$$\theta^{(1)} = (-10,048; -6,842; -13,769; -1,545; 11,275; -37,219; -32,991; -15,080; -29,719), \theta^{(2)} = (26, .646; 14, 724), \theta^{(3)} = (9,982; -2,792).$$

Fig. 6 visualized calculated curves $p_j^*(\alpha_j^{(1)}), j = \overline{0,2}; p_j^*(\alpha_j^{(2)}), j = \overline{3,8}; q_i^*(\xi_i), i = \overline{0,1}$.

4. Discussion and future work

Let's start the discussion with the analysis of the results visualized in Figs. 2–5. They represent the results of calculating the optimal estimates for $T1$, $T2$ problems for the common initial data presented at the beginning of Section 3 and the common calculated data summarized in Fig. 1. The specificity of the problem $T1$ was that optimal estimates $\langle p^*, q^* \rangle$ for the data presented in Fig. 1 and defined combinations of a priori data sets AS were obtained as a result of solving the optimization problem [36] with constraints, including oriented to the normalization of probabilities: $\sum_{j=1}^{m} p_j = 1$, $\sum_{i=1}^{n} q_i = 1$. The specificity of the problem $T2$ was that the optimal estimates $\langle p^*, q^* \rangle$ for the data presented in Fig. 1 and the determined combinations of a priori data sets AS were obtained as a result of solving the optimization problem (41) with restrictions, including regulating the interval method of determining the corresponding probabilities: $0 \le p_j \le 1, j = \overline{1,m}$; $0 \le q_i \le 1, i = \overline{1,n}$, (see expression 6.10). Comparative analysis of the presented in Figs. 2, Figs. 4 and 3, Fig. 5 results allow us to state that the estimates of controlled parameters of the small data model obtained for interval probabilities are characterized by a higher value of the conditional entropy maximum than the same type of estimates obtained for normalized probabilities.

We substantiate the empirically discovered fact that under certain conditions estimates [38,39] may differ from estimates (43) and

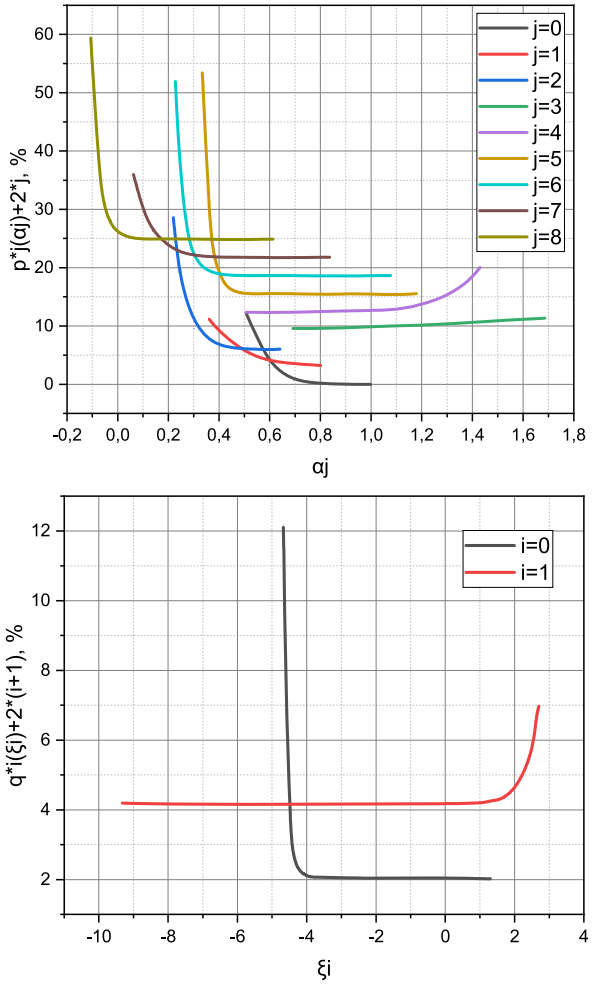


Fig. 6. Calculated dependence of $p_j^* = f(\alpha_j), j = \overline{0,8}; q_i^* = f(\xi_i), i = \overline{0,1}$.

(44) in the Shannon entropy metric [36]. We define the entropy $H(p,q)$ as $H(p,q) = H(z), z = (p,q) \in R_+^{m+n}$, then we denote the optimal estimates [38,39] as z_1^*, and the optimal estimates (43), (44) as z_2^*. We define the global maximum of entropy as $\widehat{z} = \text{argmax}H(z)$ on the set

$$Z = \{z : 0 \le z \le 1\} \supset \widetilde{Z} = \{z : \langle p, 1\rangle, (q, 1)\}. \tag{46}$$

The entropy defined by expression [36] is geometrically characterized as a concave function with a single (respectively, global) maximum at the point $\widehat{z}$. The entropy value at an arbitrary point z depends on the distance to the point $\widehat{z}$. We define the distance from point z_1^* to point $\widehat{z}$ as $\Delta(z_1^*, \widehat{z})$, and the distance from point z_2^* to point $\widehat{z}$ as $\Delta(z_2^*, \widehat{z})$. Taking into account the fact that the function [36] is concave on the set (46), we can write:

$$\Delta\left(z_1^*, \widehat{z}\right) \ge \Delta\left(z_2^*, \widehat{z}\right). \tag{47}$$

This inequality turns into strict equality only for $z_1^* = \widehat{z}$. Here is another interpretation of expression (47): for $z \in (R_+^{(m+n)} \setminus \widetilde{Z})$ we have $H(z_1^*) \le H(z_2^*)$. Deviation of L^* values from Fig. 2 from symmetrical values from Fig. 4 are fully included in the presented theoretical argumentation and not only testify in favour of the adequacy of the mathematical apparatus presented in Section 3, but also allow us to predict the promising direction of further research (analytical formalization of the dependence between the normalized or interval presentation of probabilities p, q and the variability of L^* values). However, we can already say that etalon parameters and a priori probabilities are interrelated: with a “successful" choice of them (a set $AS2 : \{p_2^0, q_1^0\}$ with an uneven distribution of stochastic values of the controlled parameters and interference) a “better" approximation to the etalon parameters is obtained by the value of the relative squared error ε, than with an unsuccessful selection of the set (for example, when choosing a set $AS3 : \{p_3^0, q_1^0\}$ characterized by a combined distribution, in which the stochastic values of the controlled parameters are unevenly distributed, and the stochastic interference values are uniformly distributed).

We will conclude the discussion by commenting on the results presented in Fig. 6. The task of restoring the dependence between data sets under specific hypotheses regarding data properties is fully characterized by the proposed dynamic small data model [3] based on the Volterra polynomial with stochastic weight functions. The evaluation of the probability densities of the weight functions of the controlled parameters and interference, represented in Fig. 6., fully confirmed the hypotheses formulated in Section 3 regarding the class of these functions and their dynamic properties. The results of the experiment confirmed that for the linear stochastic dynamic small data model, the estimation function [30] is always exponential. Empirical data on the entities “input" and “output" affect only the form and not the structural properties of the evaluation function [30]. For the nonlinear stochastic small data model, the structural properties of the estimation function [30] are more diverse but do not leave the class of nonlinear continuous functions.

The direction of ***further research*** is an analytical formalization of the dependence between the normalized or interval presentation of probabilities p, q and the variability of L^* values. In addition, this article presents a mathematical apparatus for estimating dynamic small data, as well as tools for implementing this evaluation. It would be logical to devote one of the following studies to the evaluation of real small data using the formulated methods. Our focus is on dynamic data on real physical processes [36–39].

5. Conclusions

Formalization of dependencies between data sets, taking into account specific hypotheses about data properties, is a constantly relevant task. Previously, the authors presented the material [33], in which they proposed an interference-resistant concept for solving this problem taking into account small data phenomena. As part of this concept, stochastic small data models were investigated, the probabilistic characteristics (parameter’s distribution density functions) of which are maximizing the values of the corresponding functionals defined in terms of Shannon entropy. Approaches for evaluating the mentioned functions and studying their properties were also proposed. This article is devoted to the development of the concept mentioned above.

The mathematical apparatus presented in the article is based on the principle of maximization of information entropy on sets determined as a result of a small number of censored measurements of “input" and “output" entities in the presence of noise. These data structures became the basis for the formalization of linear and nonlinear dynamic and static models of small data with stochastic parameters, which include both controlled and noise-oriented input and output measurement entities. For all variants of the above-mentioned small data models, the tasks of determining the optimal estimates of the probability density functions of the parameters (taking into account the variants of both the normalized and the interval representation of the corresponding probabilities) were carried out. Formulated optimization problems are reduced to the forms canonical for the stochastic linear programming problem with probabilistic constraints. This made it possible to present solutions in an analytical form. We note that the formalization of the mentioned optimization problems with an orientation towards maximizing the information entropy guarantees obtaining the best solutions in the conditions of the maximum uncertainty of both the stochastic controlled parameters and measurement noises of the models created.

The functionality of the proposed mathematical apparatus is demonstrated in examples of evaluation of probability density functions of parameters of linear and nonlinear (power-law) stochastic static and dynamic small data models. The obtained results are included in the theoretical basis of the work and testify in favour of its adequacy.

Funding

This work was funded by Researchers Supporting Project number (RSP2024R503), King Saud University, Saudi Arabia (funder: Dr.

TorkiAltameem).

Data availability statement

Most data is contained within the article. All the data are available on request due to restrictions, e.g., privacy or ethics.

CRediT authorship contribution statement

Kovtun Viacheslav: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Torki Altameem:** Validation, Resources, Data curation. **Mohammed Al-Maitah:** Validation, Resources, Data curation. **Wojciech Kempa:** Validation, Resources, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are grateful to all colleagues and institutions that contributed to the research and made it possible to publish its results.

References

[1] L.P. Fávero, P. Belfiore, R. de Freitas Souza, Overview of data science, analytics, and machine learning, Data Science, Analytics and Machine Learning with R (2023) 3–6, https://doi.org/10.1016/b978-0-12-824271-1.00034-2.
[2] G. Revathy, S.A. Alghamdi, S.M. Alahmari, et al., Sentiment analysis using machine learning: progress in the machine intelligence for data science, Sustain. Energy Technol. Assessments 53 (2022) 102557, https://doi.org/10.1016/j.seta.2022.102557.
[3] D. Ranke, C.A.R. Perini, J.-P. Correa-Baena, In data science we trust: machine learning for stable halide perovskites, Matter 4 (2021) 1092–1094, https://doi.org/10.1016/j.matt.2021.03.007.
[4] I.V. Izonin, R.O. Tkachenko, O.L. Semchyshyn, An ensemble method for the analysis of small biomedical data based on a neural network without training, in: Èlektronnoe Modelirovanie, vol. 45, National Academy of Sciences of Ukraine (Co. LTD Ukrinformnauka) (Publications), 2023, pp. 65–76, https://doi.org/10.15407/emodel.45.06.065, no. 6.
[5] K. Yeturu, Machine learning algorithms, applications, and practices in data science, Handb. Stat. (2020) 81–206, https://doi.org/10.1016/bs.host.2020.01.002.
[6] V. Jalajakshi, A.N. Myna, Importance of statistics to data science, Global Transitions Proceedings 3 (2022) 326–331, https://doi.org/10.1016/j.gltp.2022.03.019.
[7] H. Hassani, C. Beneki, E.S. Silva, et al., The science of statistics versus data science: what is the future? Technol. Forecast. Soc. Change 173 (2021) 121111 https://doi.org/10.1016/j.techfore.2021.121111.
[8] R.S. Jones, J.M. Rosenberg, Characterizing whole class discussions about data and statistics with conversation profile analysis, J. Math. Behav. 67 (2022) 100996, https://doi.org/10.1016/j.jmathb.2022.100996.
[9] S. Watanabe, Mathematical theory of bayesian statistics where all models are wrong, Handb. Stat. (2022) 209–238, https://doi.org/10.1016/bs.host.2022.06.001.
[10] E.A. Manziuk, et al., Semantic alignment of ontologies meaningful categories with the generalization of descriptive structures, in: PROBLEMS IN PROGRAMMING, No. 3–4, National Academy of Sciences of Ukraine (Co. LTD Ukrinformnauka) (Publications), Dec. 2022, pp. 355–363, https://doi.org/10.15407/pp2022.03-04.355.
[11] R.C. Mittal, Mathematics and Statistics behind Machine Learning, Unpublished, 2022, https://doi.org/10.13140/RG.2.2.15915.92969.
[12] C. Ley, R.K. Martin, A. Pareek, et al., Machine learning and conventional statistics: making sense of the differences, Knee Surg. Sports Traumatol. Arthrosc. 30 (2022) 753–757, https://doi.org/10.1007/s00167-022-06896-6.
[13] S. Gocheva-Ilieva, Special issue “statistical data modeling and machine learning with applications.”, Mathematics 9 (2021) 2997, https://doi.org/10.3390/math9232997.
[14] T. Sun, Y. Liu, S. Gao, et al., Distribution-based maximum likelihood estimation methods are preferred for estimating Salmonella concentration in chicken when contamination data are highly left-censored, Food Microbiol. 113 (2023) 104283, https://doi.org/10.1016/j.fm.2023.104283.
[15] C. Lubeigt, F. Vincent, L. Ortega, et al., Approximate maximum likelihood time-delay estimation for two closely spaced sources, Signal Process. 210 (2023) 109056, https://doi.org/10.1016/j.sigpro.2023.109056, 2023.
[16] T.C. Fung, Maximum weighted likelihood estimator for robust heavy-tail modelling of finite mixture models, Insur. Math. Econ. 107 (2022) 180–198, https://doi.org/10.1016/j.insmatheco.2022.08.008.
[17] F. Liu, A comparison between multivariate linear model and maximum likelihood estimation for the prediction of elemental composition of coal using proximate analysis, Results in Engineering 13 (2022) 100338, https://doi.org/10.1016/j.rineng.2022.100338.
[18] F. Li, K. Li, K. Lu, et al., Random noise suppression and parameter estimation for magnetic resonance sounding signal based on maximum likelihood estimation, J. Appl. Geophys. 176 (2020) 104007, https://doi.org/10.1016/j.jappgeo.2020.104007.
[19] Z. Chen, L.G. Epstein, A central limit theorem for sets of probability measures, Stoch. Process. their Appl. 152 (2022) 424–451, https://doi.org/10.1016/j.spa.2022.07.003.
[20] Z. Shao, B. Liang, H. Gao, Extracting independent and identically distributed samples from time series significant wave heights in the yellow sea, Coast Eng. 158 (2020) 103693, https://doi.org/10.1016/j.coastaleng.2020.103693.
[21] T. Łuczak, K. Mieczkowska, M. Šileikis, On maximal tail probability of sums of nonnegative, independent and identically distributed random variables, Stat. Probab. Lett. 129 (2017) 12–16, https://doi.org/10.1016/j.spl.2017.04.024.
[22] B. Avanzi, G. Boglioni Beaulieu, P. Lafaye de Micheaux, et al., A counterexample to the existence of a general central limit theorem for pairwise independent identically distributed random variables, J. Math. Anal. Appl. 499 (2021) 124982, https://doi.org/10.1016/j.jmaa.2021.124982.
[23] V. Kovtun, K. Grochla, V. Kharchenko, M.A. Haq, A. Semenov, Stochastic forecasting of variable small data as a basis for analyzing an early stage of a cyber epidemic, in: Scientific Reports, vol. 13, Springer Science and Business Media LLC, 2023, https://doi.org/10.1038/s41598-023-49007-2 no. 1.
[24] F.S. Al-Duais, R.S. Al-Sharpi, A unique Markov chain Monte Carlo method for forecasting wind power utilizing time series model, Alex. Eng. J. 74 (2023) 51–63, https://doi.org/10.1016/j.aej.2023.05.019.

[25] A.m Tariq, J. Yan, F. Mumtaz, Land change modeler and CA-Markov chain analysis for land use land cover change using satellite data of peshawar, Pakistan, Phys. Chem. Earth, Parts A/B/C 128 (2022) 103286, https://doi.org/10.1016/j.pce.2022.103286.
[26] S. Gundlach, O. Junge, L. Wienbrandt, A. Comparison of Markov chain Monte Carlo software for the evolutionary analysis of Y-chromosomal microsatellite data, Comput. Struct. Biotechnol. J. 17 (2019) 1082–1090, https://doi.org/10.1016/j.csbj.2019.07.014.
[27] A.H. Amshi, R. Prasad, Time series analysis and forecasting of cholera disease using discrete wavelet transform and seasonal autoregressive integrated moving average model, Scientific African 20 (2023) e01652, https://doi.org/10.1016/j.sciaf.2023.e01652.
[28] V. Kovtun, et al., Research of pareto-optimal schemes of control of availability of the information system for critical use, In Proc. 1st International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2020), CEUR-WS 2623 (2020) 174–193.
[29] S. Copiello, Peer and neighborhood effects: citation analysis using a spatial autoregressive model and pseudo-spatial data, Journal of Informetrics 13 (2019) 238–254, https://doi.org/10.1016/j.joi.2019.01.002.
[30] V. Kovtun, E. Zaitseva, V. Levashenko, K. Grochla, O. Kovtun, Small stochastic data compactification concept justified in the entropy basis, Entropy 25 (12) (2023) 1567, https://doi.org/10.3390/e25121567. MDPI AG.
[31] T. Ransford, N. Walsh, Norms of polynomials of the Volterra operator, J. Math. Anal. Appl. 517 (2023) 126626, https://doi.org/10.1016/j.jmaa.2022.126626.
[32] P.K. Singh, S. Saha Ray, An efficient numerical method based on lucas polynomials to solve multi-dimensional stochastic itô-volterra integral equations, Math. Comput. Simulat. 203 (2023) 826–845, https://doi.org/10.1016/j.matcom.2022.06.029.
[33] E. Palezzato, M. Torielli, K-lefschetz properties, sectional matrices and hyperplane arrangements, J. Algebra 590 (2022) 215–233, https://doi.org/10.1016/j.jalgebra.2021.10.014.
[34] O. Bisikalo, V. Kharchenko, V. Kovtun, et al., Parameterization of the stochastic model for evaluating variable small data in the Shannon entropy basis, Entropy 25 (2023) 184, https://doi.org/10.3390/e25020184.
[35] T.D. van Pelt, J.C. Fransoo, A note on “linear programming models for a stochastic dynamic capacitated lot sizing problem.”, Comput. Oper. Res. 89 (2018) 13–16, https://doi.org/10.1016/j.cor.2017.06.015.
[36] H. Tempelmeier, M. Kirste, Hilger, T. Linear programming models for a stochastic dynamic capacitated lot sizing problem, Comput. Oper. Res. 91 (2018) 258–259, https://doi.org/10.1016/j.cor.2017.11.010.
[37] N. Abbas, K.U. Rehman, W. Shatanawi, K. Abodayeh, Mathematical model of temperature-dependent flow of power-law nanofluid over a variable stretching riga sheet, Waves Random Complex Media 1–18 (2022), https://doi.org/10.1080/17455030.2022.2111029.
[38] T.A.M. Shatnawi, N. Abbas, W. Shatanawi, Comparative study of casson hybrid nanofluid models with induced magnetic radiative flow over a vertical permeable exponentially stretching sheet, AIMS Mathematics 7 (2022) 20545–20564, https://doi.org/10.3934/math.20221126.
[39] A. Nazir, N. Abbas, Shatanawi, W. On stability analysis of a mathematical model of a society confronting with internal extremism, Int. J. Mod. Phys. B 37 (2022), https://doi.org/10.1142/s0217979223500650.
[40] T.A.M. Shatnawi, N. Abbas, W. Shatanawi, Mathematical analysis of unsteady stagnation point flow of radiative casson hybrid nanofluid flow over a vertical riga sheet, Mathematics 10 (2022) 3573, https://doi.org/10.3390/math10193573.