BELARUSIAN STATE UNIVERSITY

RESEARCH INSTITUTE FOR APPLIED PROBLEMS
OF MATHEMATICS AND INFORMATICS

COMPUTER DATA ANALYSIS AND MODELING:

STOCHASTICS AND DATA SCIENCE

Proceedings of the Fourteenth International Conference
Minsk, September 24–27, 2025

Minsk
BSU
2025

UDC 519.2

Computer Data Analysis and Modeling: Stochastics and Data Science : Proc. of the Fourteenth Intern. Conf., Minsk, Sept. 24–27, 2025. — Minsk : BSU, 2025. — 307 p.
ISBN .

This collection of papers includes proceedings of the Fourteenth International Conference "Computer Data Analysis and Modeling: Stochastics and Data Science" organized by the Belarusian State University and held in September 2025 in Minsk. Papers are reviewed by qualified researchers from Belarus and Russia.

The papers are devoted to the topical problems: robust data analysis; multivariate analysis; statistical analysis of time series and random fields; probabilistic and statistical analysis of discrete data; stochastics: theory and applications; statistical methods for signal and image processing; econometric modeling and financial mathematics; survey analysis and official statistics; data science: theory, visual analytics, software and applications.

For specialists who work in the fields of mathematical statistics and its applications, computer data analysis, data science and statistical software development.

UDC 519.2

ISBN

PROGRAM COMMITTEE

Honorary Chair

A. Blokhin — Vice-Rector of the Belarusian State University (Minsk, Belarus)

Co-Chairs

Prof. Dr. Yu. Kharin (Minsk, Belarus)
Prof. Dr. A. Zubkov (Moscow, Russia)

Conference Secretary

Prof. V. Malugin (Minsk, Belarus)

Members

Prof. A. Dudin (Minsk, Belarus)
Prof. A. Egorov (Minsk, Belarus)
Prof. Sh. Formanov (Tashkent, Uzbekistan)
Prof. A. Kharin (Minsk, Belarus)
Prof. G. Khatskevich (Minsk, Belarus)
Prof. V. Krasnoproshin (Minsk, Belarus)
Prof. B. Lemeshko (Novosibirsk, Russia)
Prof. V. Lotov (Novosibirsk, Russia)
Prof. V. Mkhitarian (Moscow, Russia)
Prof. V. Mukha (Minsk, Belarus)
Prof. Yu. Pavlov (Petrozavodsk, Russia)
Prof. O. Sharipov (Tashkent, Uzbekistan)
Prof. N. Troush (Minsk, Belarus)
Prof. A. Tuzikov (Minsk, Belarus)

Local Organizing Committee

Chair

Yu. Orlovich, Dean of the Faculty of Applied Mathematics and Computer Science

Co-Chair

A. Kharin, Head of the Department of Probability Theory and Mathematical Statistics

Members

E. Krasnogir, V. Kutsapalava, M. Maltsew, V. Malugin, U. Palukha, I. Pirshtuk, V. Voloshko

To 25 years Anniversary
of the Research Institute for Applied Problems
of Mathematics and Informatics at BSU

PREFACE

The Fourteenth International Conference "Computer Data Analysis and Modeling: Stochastics and Data Science" (CDAM'2025) organized by the Belarusian State University on September 24–27, 2025, is devoted to the topical problems in computer data analysis and modeling. Methods of computer data analysis and modeling are widely used in variety of fields: computer support of scientific research; decision making in economics, business, engineering, medicine end ecology; statistical modeling of complex systems of different nature and purpose. In the Republic of Belarus computer data analysis and modeling have been developed successfully for more than 30 years. Scientific conferences CDAM were held in September 1988, December 1990, December 1992, September 1995, June 1998, September 2001, September 2004, September 2007, September 2010, September 2013, September 2016, September 2019 and September 2022 in Minsk.

The Proceedings of the CDAM'2025 contain 65 papers. The topics of the papers correspond to the following scientific problems: robust data analysis; multivariate analysis; statistical analysis of time series and random fields; probabilistic and statistical analysis of discrete data; stochastics: theory and applications; statistical methods for signal and image processing; econometric modeling and financial mathematics; survey analysis and official statistics; data science: theory, visual analytics, software and applications.

The Organizing Committee of the CDAM'2025 makes its acknowledgements to the Belarusian State University, the Research Institute for Applied Problems of Mathematics and Informatics for financial and technical support.

Yuriy Kharin
Andrey Zubkov

CONTENTS

ON THE ASSESSMENT OF THE IMPACT OF THE STRUCTURAL COMPLEXITY OF REGIONAL ECONOMIES ON GRP

M.Y. Afanasiev[1], A.A. Gusev[2]

[1,2] *Central Economics and Mathematics Institute of Russian Academy of Sciences*
Moscow, RUSSIA
e-mail: [1]mi.afan@yandex.ru, [2]gusevalexeyal@yandex.ru

Current scientific discussions are focused on identifying professions and types of economic activity that will become most in demand in the future and determine priority areas for diversification of regional economies. Analysis of such trends is important for forecasting the dynamics of GRP. The purpose of this work is to construct an integral index of structural complexity on the basis of four basic indices of economic complexity of regional economies, calculated by the authors on the basis of data on the structure of employment, the structure of the distribution of enterprises and the structure of GRP.

Keywords: regional economy, econometrics, economic complexity, integral index, GRP

1 Introduction

The transition from an economy based on the export of raw materials to a high-tech economy and the strengthening of the economic security of regions involves an increase in the complexity of production structures and economic systems. Recommendations for the diversification of national and regional economies can be based on approaches focused on increasing economic complexity [4–7]. For example, the paper [3] proposes an approach to the selection of diversification areas based on recommendations for the development of sectors, aimed at increasing the economic complexity of the regional economy. The accumulated experience allows us to approach the problem of a generalized assessment of the complexity of regional economies using the integral index of structural complexity. To construct the integral index, various structures of regional economies have been formed: the structure of GRP according to data on production volumes by type of economic activity (TEA); the structure of employment of regions by occupational groups; the structure of employment of regions by TEA; the structure of the distribution of enterprises by TEA. Based on the concept of economic complexity, the complexity of each structure is estimated and a corresponding basic index of economic complexity is constructed [1,2].

Figure 1 show 0-1 matrices describing the structures of strong TEAs and professional groups of regions for four basic indices according to 2022 data. The rows of the matrices correspond to regions, and the columns correspond to TEAs or occupational groups. A dark cell of the matrix means that the corresponding element of the matrix is equal to 1, that is, the TEA (or occupational group) is strong in the region. The rows are ordered from top to bottom in descending order of the regions complexity scores. The columns are ordered from left to right in ascending order of TEA (or occupational

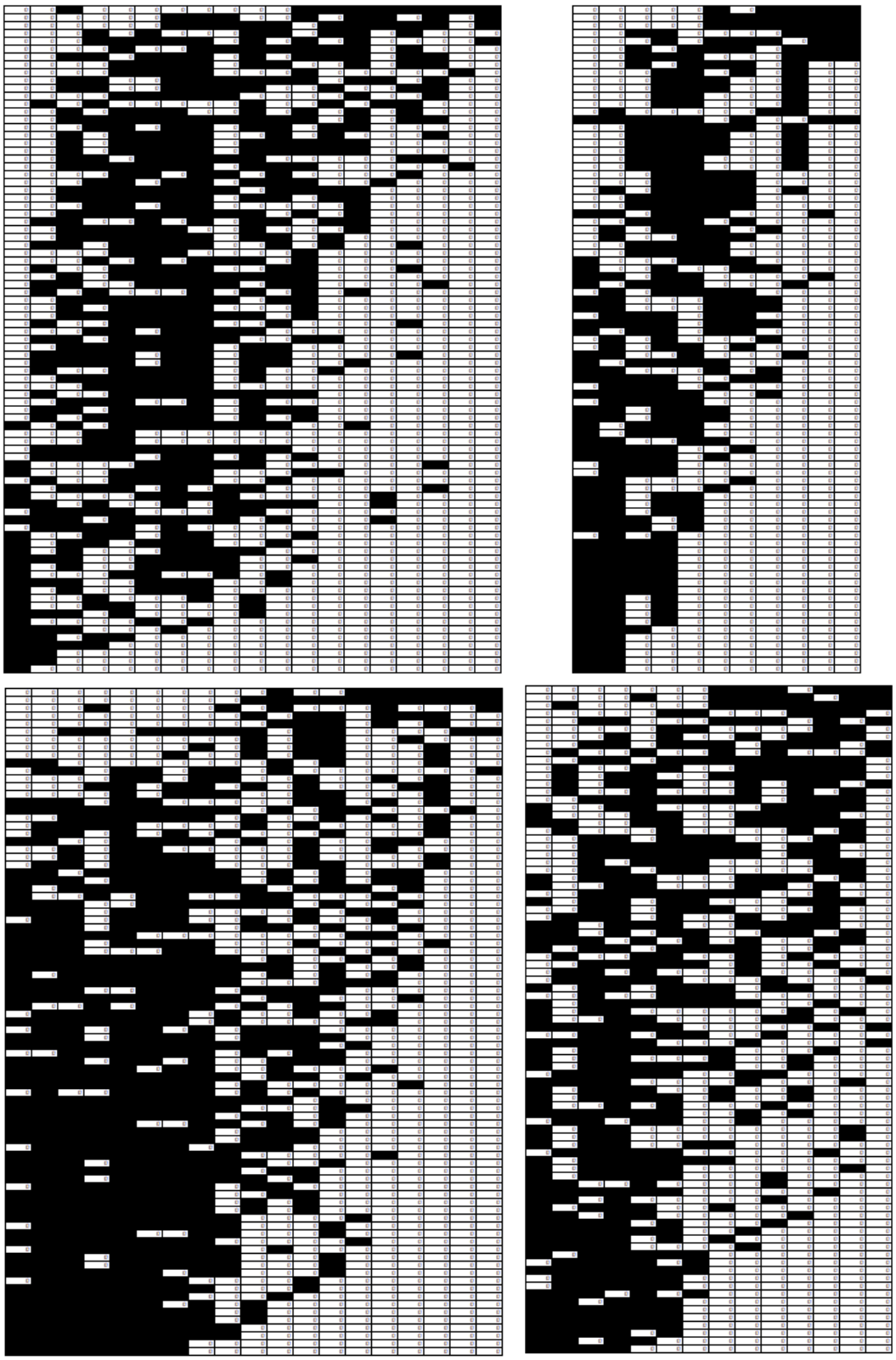


Figure 1: Matrices 0-1: region-TEA by GRP structure (left top); region-occupational group on professional employment (right top); region-TEA by number of enterprises (left bottom); region-TEA by number of employed (right bottom)

groups) difficulty scores. Figure 1 combined with the difficulty scores provide insight into the respective regional structures of strong TEAs, as well as the relationship between regional difficulty scores on the one hand and TEAs on the other.

2 Model

Four approaches were used to construct integral indices of structural complexity: component analysis, averaging, the principle of equal correlation with basic indices, and the principle of proximity to the standard. To assess the impact of structural complexity on GRP, the parameters of the production function with a built-in integral index are estimated:

$$\begin{aligned}\ln(VRP_{kt}) = (\alpha + \alpha_1 \cdot t) \cdot \ln(L_{kt}) + (\beta + \beta_1 \cdot t) \cdot \ln(K_{kt}) \\ + (s + s_1 \cdot t) \cdot nINT_{kt} + (r + r_1 \cdot t) \cdot d_k + c \cdot t + \text{const} + \varepsilon_{kt}, \qquad (1)\end{aligned}$$

where VRP_{kt} is GRP of the region k in constant prices, L_{kt} is number of employees, K_{kt} is cost of fixed assets, $nINT_{kt}$ is value of the components of the normalized integral index of structural complexity, const is constant, ε_{kt} is regression error, t is time, $d_k \in \{0, 1\}$ equals 1 for a group of regions with a general specialization, on the basis of the aggregate of which the natural rent is estimated, r is valuation of natural resource rent.

The integral index, built on the principle of proximity to the standard, shows the best statistical characteristics when assessing the production function (1) according to the data of 2019 and 2022 and is significant at the 5% level. Estimates of natural resource rent and GRP elasticity in terms of the complexity structure of the index make it possible to predict the impact of a wide range of federal and regional projects on the growth of regional economies.

References

1. Afanasiev M.Yu., Gusev A.A., Nanavyan A.M. (2023) Assessment of the professional structure of the employed population in Russian regions based on the concept of economic complexity. *Economic and social changes: facts, trends, forecast*. Vol. **16**, No. **6**. pp. 91-107.

2. Afanasiev M.Yu., Gusev A.A., Nanavyan A.M. (2025) *Integral index of the complexity of employment structures in Russian regions* (in print).

3. Afanasiev M.Yu., Ilyin N.I. (2022) New guidelines for choosing priority areas of economic diversification based on the system of situational centers. *Economics and mathematical methods*. Vo. **58**, No. **4**. pp. 29–44.

4. Afanasiev M.Yu., Kudrov A.V. (2021) Economic complexity and nesting of structures of regional economies. *Economics and Mathematical Methods*. Vol. **57**, No. **3**. pp. 67–78.

5. Lyubimov I.L. [et. al.] (2017). The complexity of the economy and the possibility of export diversification in Russian regions. *Journal of the New Economic Association*. Vol. **2**, No. **34**. P. 94–122.

6. Hausmann R. [et. al.] (2014) *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. The MIT Press.

7. Hidalgo C.A., Hausmann R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*. Vol. **106**, No. **26**. P. 10570–10575.

DISTRIBUTIONS OF LENGTHS OF EXCURSIONS OF A BROWNIAN BRIDGE

V.I. Afanasyev[1]

[1]*Steklov Mathematical Institute of Russian Academy of Sciences*
Moscow, RUSSIA
e-mail: [1]viafan@mi-ras.ru

The paper shows distributions of the lengths of excursions of a Brownian bridge, as well as the one-dimensional distributions of the meanders and inverse meanders of the Brownian bridge.

Keywords: Brownian bridge, excursion of Brownian bridge, meander of Brownian bridge

Let $\{W_0(t),\ t \in [0,1]\}$ be a standard Brownian bridge. Denote $\beta(s)$ the time of the last attainment by W_0 of state 0 before $s \in (0,1)$ and denote $\gamma(s)$ the time of the first attainment by W_0 of state 0 after $s \in (0,1)$. The interval $(\beta(s), \gamma(s))$ is called the excursion of the Brownian bridge straddling s; the intervals $(\beta(s), s)$ and $(s, \gamma(s))$ are called the left and right sides of the excursion, respectively.

The role of a Brownian bridge in mathematical statistics is well known. According to Donsker's theorem, an empirical process converges in distribution in the space $D[0,1]$ to a Brownian bridge. From this theorem, in particular, follows the fundamental theorem of mathematical statistics on convergence in distribution of Kolmogorov statistics to the maximum of absolute value of a Brownian bridge.

Statisticians are interested in distribution of the length of excursions of a Brownian bridge, i.e., the random variables $\Delta(s) = \gamma(s) - \beta(s)$, as well as distribution of the lengths of their left and right parts, i.e., the random variables $\Delta_1(s) = s - \beta(s)$ and $\Delta_2(s) = \gamma(s) - s$. Note that $\Delta(s) = \Delta_1(s) + \Delta_2(s)$. We indicate the main result, limiting ourselves to the value $s = 1/2$ and by setting $\Delta_1(1/2) = \Delta_1$, $\Delta_2(1/2) = \Delta_2$ and $\Delta(1/2) = \Delta$.

Theorem 1. *For* $0 < a_1 < a_2 \le 1/2$ *and* $0 < b_1 < b_2 \le 1/2$

$$\mathbf{P}(\Delta_1 \in (a_1, a_2), \Delta_2 \in (b_1, b_2)) = \frac{1}{\pi} \int_{a_1}^{a_2} \int_{b_1}^{b_2} \frac{dadb}{(a+b)^{3/2} \sqrt{(1-2a)(1-2b)}}.$$

From Theorem 1, as corollaries, we obtain the following results.

Theorem 2. *The random variable* Δ *is absolutely continuous and its distribution density has the form: for* $x \in (0, 1/2)$

$$p_\Delta(x) = \frac{1}{\pi x^{3/2}} \arcsin \frac{x}{1-x},$$

and for $x \in (1/2, 1)$

$$p_\Delta(x) = \frac{1}{2x^{3/2}}.$$

Theorem 3. *For* $0 < a \leq 1/2$

$$\mathbf{P}\left(\Delta_1 \leq a\right) = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{6a-1}{2a+1}.$$

Theorem 4. *For* $0 < a \leq 1/2$

$$\mathbf{P}\left(\Delta_2 \leq a\right) = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{6a-1}{2a+1}.$$

In addition to the lengths of a Brownian bridge excursion and their parts, it is interesting to consider the Brownian bridge at these intervals. For each $s \in (0,1)$, we introduce the following concepts:

excursion of a Brownian bridge:

$$W_0^{(ex)}\left(t;s\right) = \frac{\left|W_0\left(\beta\left(s\right) + t\left(\gamma\left(s\right) - \beta\left(s\right)\right)\right)\right|}{\sqrt{\gamma\left(s\right) - \beta\left(s\right)}}, \quad t \in [0,1]\,;$$

meander of a Brownian bridge:

$$W_0^{(me)}\left(t;s\right) = \frac{\left|W_0\left(\beta\left(s\right) + t\left(s - \beta\left(s\right)\right)\right)\right|}{\sqrt{s - \beta\left(s\right)}}, \quad t \in [0,1]\,;$$

inverse meander of a Brownian bridge:

$$W_0^{(inv.me)}\left(t;s\right) = \frac{\left|W_0\left(s + t\left(\gamma\left(s\right) - s\right)\right) - W_0\left(s\right)\right|}{\sqrt{\gamma\left(s\right) - s}}, \quad t \in [0,1]\,.$$

First, we will indicate the auxiliary results.

Theorem 5. *For* $0 < a_1 < a_2 \leq 1/2$ *and* $0 < b_1 < b_2$

$$\begin{aligned} & \mathbf{P}\left(\Delta_1 \in (a_1, a_2)\,,\ W_0\left(\frac{1}{2}\right) \in (b_1, b_2)\right) \\ = & \frac{1}{\pi} \int\limits_{b_1}^{b_2} b e^{-b^2} db \int\limits_{a_1}^{a_2} \frac{e^{-b^2/(2a)}}{a^{3/2}\sqrt{1-2a}} da. \end{aligned}$$

Theorem 6. *For* $0 < a_1 < a_2 \leq 1/2$ *and* $0 < b_1 < b_2$

$$\begin{aligned} & \mathbf{P}\left(\Delta_2 \in (a_1, a_2)\,,\ W_0\left(\frac{1}{2}\right) \in (b_1, b_2)\right) \\ = & \frac{1}{\pi} \int\limits_{b_1}^{b_2} b e^{-b^2} db \int\limits_{a_1}^{a_2} \frac{e^{-b^2/(2a)}}{a^{3/2}\sqrt{1-2a}} da. \end{aligned}$$

Set for $x > 0$

$$F(x) = \sqrt{2} \int_0^x u e^{-3u^2/4} I_0\left(u^2/4\right) du,$$

where $I_0(\cdot)$ is a modified Bessel function of the first kind of order 0, i.e.

$$I_0(z) = \sum_{m=0}^{\infty} \frac{1}{(m!)^2} \left(\frac{z}{2}\right)^{2m}, \qquad z \in \mathbf{C}.$$

We assume that $s = 1/2$ and we set

$$W_0^{(me)}(t; 1/2) = W_0^{(me)}(t), \qquad W_0^{(inv.me)}(t; 1/2) = W_0^{(inv.me)}(t).$$

As corollaries of Theorems 5 and 6, we indicate one-dimensional distributions of the introduced processes at time $t = 1$.

Theorem 7. *For $x > 0$*

$$\mathbf{P}\left(W_0^{(me)}(1) \leq x\right) = F(x).$$

Theorem 8. *For $x > 0$*

$$\mathbf{P}\left(W_0^{(inv.me)}(1) \leq x\right) = F(x).$$

APPLICATION OF R LANGUAGE TO DECOMPOSE THE GENDER GAP IN AVERAGE HOURLY WAGES IN THE REPUBLIC OF BELARUS

N.V. Agabekova[1], A.G. Bendega[2]
[1,2]*Belarusian State Economic University*
Minsk, BELARUS
e-mail: [1]agabnin@mail.ru, [2]sasha.bendegaweightlifting@mail.ru

The paper describes the capabilities of the R programming language to decompose the gender gap in average hourly wages of workers by major occupation groups using the Oaksaki-Blinder method.

Keywords: decomposition, R package, gender pay gap, Oaksaki-Blinder model

1 Introduction

Target 8.5 of Sustainable Development Goal 8 in the Republic of Belarus is "By 2030, achieve full and productive employment and decent work for all women and men, including for young people and persons with disabilities, and equal pay for work of equal value" [1]. One of the indicators used to assess the achievement of this goal is the gender gap in the average hourly wages of employees, calculated on the basis of a sample survey of organizations on the wages of employees by personnel category and occupation group (Form 6-t (occupations)). However, the unadjusted gender gap is not a criterion of discrimination, as it does not take into account a number of factors, such as occupational segregation, level of education, etc. Therefore, there is a need to analyze the gender gap in order to identify the influence of various factors on its value.

This paper describes the application of the R programming language to decompose the gender gap in the average hourly wages of workers by occupation group.

2 Model

The Oaksaki-Blinder method is widely used to decompose the gender gap. At the first stage, the equations for male and female subsamples are estimated separately. In general, the equation has the following form:

$$\ln W = B^0 + \sum B \cdot X + e, \tag{1}$$

where W are values of average hourly wages of men and women, B^0 is a constant term, B is a regression coefficient, X are professional characteristics of employees.

The overall difference in mean outcomes between gender groups is decomposed into the effect of differences in the mean characteristics of men and women (composition effect) and the effect of returns to characteristics as differences in regression coefficient estimates (wage structure effect).

We utilize the Oaxaca package of the R programming language for the purpose of decomposition [2].

Neumark's formula is applied to decompose the difference in average hourly wages:

$$\begin{aligned}\ln \overline{W_m} - \ln \overline{W_f} = &\sum (\overline{X}^i_m - \overline{X}^i_f) \cdot B^i_t + \sum (B^i_m - B^i_t) \cdot \overline{X}^i_m \\ &+ \sum (B^i_t - B^i_f) \cdot \overline{X}^i_f + (B^0_m - B^0_t) + (B^0_t - B^0_f),\end{aligned} \tag{2}$$

where W_m, W_f are values of average hourly wages of men and women, $\overline{X}^i_m$, $\overline{X}^i_f$ are individual professional characteristics of men and women (in the studied example: the average work experience in the organization, the proportion of men and women with a certain level of education, the distribution of men and women in minor groups of occupations), B^0_t, B^0_f, B^0_m are constant terms of the pooled, female and male equations, B^i_t, B^i_f, B^i_m are regression coefficients of the pooled, female and male equations.

In the specified formula (2), the first term reflects the explained part of the gender gap, that is the difference in wages due to gender differences with the same impact on the analyzed characteristics. The rest terms correspond to the unexplained part of the gender gap. The regression coefficients of the pooled equation are used as reference coefficients. It allows us to obtain more objective estimates and avoid bias associated with the choice of a reference equation.

The data base of the calculations is an impersonal primary statistical database of a sample survey of organizations on employee salaries by staff categories and occupation groups for October 2024 (Form 6-t (professions)).

At the first stage, the average hourly wage of employees is calculated and the sample survey observations are extrapolated by aggregated employee weights. To do this we use the following code: (Figure 1):

The employee's weight is rounded to the nearest integer value and the required number of rows for each observation is duplicated accordingly.

Further, the data obtained is divided into the major occupation groups (where the gender gap is more than 5 percent) and dummy variables for minor occupation groups and educational level are created within each major group.

At the second stage, the gender gap is decomposed according to formula (2). The command written in the R programming language for calculating the major group "Specialists-professionals" is shown in Figure 2:

The results of the gender gaps decomposition are presented in Table 1.

The decomposition explains from 32.1% (for the major group "Employees engaged in the provision of office administrative and support services, services to consumers, preparation, processing of information and accounting") to 89.2% (for the main group "Unskilled workers") of the gender gap in the average hourly wage. It can be noted that in the groups with the highest gender gaps, only a small part of the gender gap remains unexplained (for the major group "Specialists", the unexplained part accounts for 28.4% of the gender gap, for the major group "Unskilled workers" – 10.8% of the gender gap). The insufficiently explained part of the gender gap in certain major groups indicates that additional factors may be included in the proposed models.

```r
#экстраполяция данных
new$okrugl <- round(new$AGGRW_1, 0)
fun1 <- function(row) {
  n <- row$okrugl
  dfnew <- data.frame(
    MALGR= rep(row$MALGR, n),
    POL = rep(row$POL, n),
    GR = rep(row$GR, n),
    SEN = rep(row$SEN, n),
    OBRGR = rep(row$OBRGR, n),
    lnzp = rep(row$lnzp, n)
  )
  return(dfnew)
}
new2 <- new %>% rowwise () %>% do(fun1(.))
```

Figure 1: Extrapolation of sample observation data

```r
results_twoMAL <-oaxaca(lnzp ~ SEN + MALGR212 + MALGR213 + MALGR214 + MALGR215 + MALGR216 +
                          MALGR221 + MALGR222 + MALGR225 + MALGR226 + MALGR231 +
                          + MALGR232 + MALGR233 + MALGR234 + MALGR235 + MALGR241 +
                          MALGR242 + MALGR243 +
                          MALGR251 + MALGR252 + MALGR261 + MALGR262 + MALGR263 + MALGR264 +
                          MALGR265 +OBRGR2 + OBRGR3 + OBRGR4
                | women)
```

Figure 2: Decomposition of the gender gap in the average hourly wage of the major group "Specialists-professionals"

Table 1: Gender gaps in average hourly wages by major occupation groups, explained part and contribution of factors, %

Major gr.	**Gender gap**	**The explained part of the gender gap**			
		Total	**Work exp.**	**Educ.**	**Minor gr.**
Specialists-professionals	36,5	69,4	-5,0	2,1	72,3
Specialists	41,3	71,6	0,1	-0,3	71,8
Employees in office/admin services, consumer services, information processing	31,2	32,1	9,1	2,8	20,2
Skilled workers in industry, construction	11,5	46,6	-0,1	1,7	45,0
Operators, machinists, assemblers	15,6	45,1	2,2	-1,1	45,1
Unskilled workers	40,3	89,2	2,6	0,6	86,0

The greatest contribution to explaining the gender gap is made by the uneven distribution of men and women in minor groups of occupations (varies from 20.2% of the difference in the logarithms of the average hourly wages of men and women in the major group "Employees engaged in the provision of office administrative and support services, services to consumers, preparation, processing of information and accounting" to 86.0% in the major group "Unskilled workers").

The conducted research proves that the unequal distribution of men and women on various grounds (especially occupational segregation) has a significant impact on the gender gap in average hourly earnings, and therefore an uncorrected gender gap cannot be used as a criterion of discrimination. The presence of an unexplained part indicates that additional factors may be included in the model (for example, the size of the enterprize, type of economic activity, etc.).

The National Statistical Committee of the Republic of Belarus has carried out an experimental calculation of the adjusted gender gap, which, according to data for October 2024, amounted to 6% (unadjusted – 26.8%). The adjusted gender gap makes it possible to eliminate the influence of individual characteristics of employees, while taking into account such factors as: educational level, occupational segregation, the influence of industry factors, work experience in the organization [3]. The resulting value of the adjusted gender gap confirms the absence of significant differences in the average hourly wages of men and women in the Republic of Belarus.

References

1. *National platform for reporting indicators of Sustainable Development Goals (SDGs)*. [Electronic resource]. Mode of access: `https://sdgplatform.belstat.gov.by/target/8`. Date of access: 05.05.2025.

2. Hlavac, M. (2022). *oaxaca: Blinder-Oaxaca Decomposition in R*. R package version 0.1.5. [Electronic resource]. Mode of access: `https://CRAN.R-project.org/package=oaxaca`. Date of access: 20.04.2025.

3. *Gender gap in average hourly wages*. [Electronic resource]. Mode of access: `https://www.belstat.gov.by/upload-belstat/upload-belstat-pdf/oficial_statistika/2025/info-gen-zp.pdf`. Date of access: 05.05.2025.

ANALYSIS OF RHYTHMIC PATTERNS OF THE TIME SERIES BASED ON THE STATISTICAL MODEL OF NBD

N.P. Alexeyeva[1], I.A. Samarin[2], A.A. Sotov[3]

[1,2] *Faculty of Mathematics and Mechanics, Saint Petersburg State University*
[3] *Faculty of Oriental Studies, Saint Petersburg State University*
Saint Petersburg, RUSSIA
e-mail: [1]nina.alekseeva@spbu.ru, [2]igor_060401@list.ru,
[3]a.sotov@yahoo.co.uk

The paper presents two methods for parameterizing quasi-periodic cycles in price returns time series (we shall call such patters, rhythmic structures). Both methods involve categorizing time series by discrete scaling of returns. The first method is based on the property of the negative binomial distribution which describes occurrence of named entities in natural texts (used in corpus linguistics and NLP). The classification problem is solved according to the parameters of the most frequently used words. The second method is based on reducing the dimensionality of incidence matrices of text sequences to a dictionary using neural networks. In both cases, it is possible to show the difference in the rhythmic structure of text financial sequences related to different industries.

Keywords: time series, leap categorization, n-gram occurrence, negative binomial distribution, neural networks

1 Introduction

In time series analysis, quasi-periodic cycles (rhythmic patterns) are classified as random, and if an adequate model is selected, the parameter estimates can be used in problems of classifying and systematizing time series.

To measure the rhythmic characteristics of a time series, a linguistic model of the negative binomial distribution (NBD) of word occurrence in texts is proposed [1]. To do this, it is necessary to categorize the time series relative to its returns, compile a dictionary of all possible patterns, and select from them those whose occurrence is consistent with NBD.

Estimates of the NBD parameters of the most commonly used patterns can be used as a set of time series characteristics.

Previously, for educational purposes, the model was successfully tested in the task of classifying electroencephalograms of patients with cervical dystonia and epilepsy; in this work, the dynamics of stock quotes on the Moscow Exchange stock market, recorded at equal intervals from April 15, 2013 to September 3, 2024, were considered as experimental data.

For the $N = 80$ sequences longer than 3,000 data points, the numerical return values were replaced with letter codes as follows: a (if returns are no less 2%), b (from 1% to less than 2%, c (from 0.5% to less than 1%), d (from 0.25% to less than 0.5%),

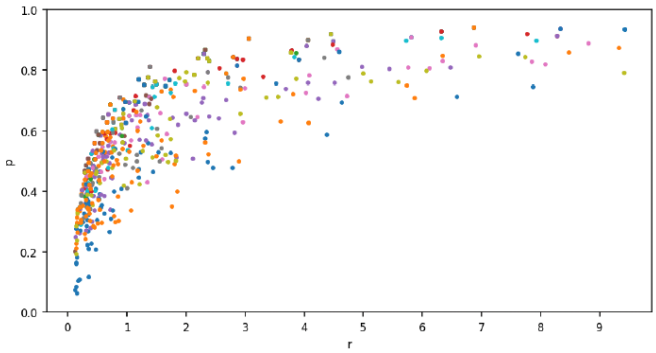


Figure 1: Two-dimensional plot of parameter r, p estimates

e (between -0.25% and 0.25%), f (from -0.5% to less than -0.25%), g (from -1% to less than -0.5%), h (from -2% to less than -1%), i (no more -2%).

We will recall the resulting categorical sequence of returns as the financial text sequence, which can be divided into k parts.

We define an n-gram as a sequence of n symbols in the text sequence and construct a sample $x_1, \ldots, x_k$, where x_i shows how many times the n-gram occurs in the i-th part, $i = 1, \ldots, k$, $k = 31$. For example, the 3-gram aii, which means an extremely large increase followed by two extremely large returns falls.

We can calculate the average values of occurrence of aii, for example, in the text sequence of Sakhalin-energo and TGK-2 are equal to 0.3 and 0.45, respectively, but the greatest interest is in their distribution laws, which are consistent with NBD.

So, we consider the empirical distributions of the most frequent n-grams, $n = 3$ or $n = 4$, and check their consistency with NBD.

2 The negative binomial distribution model of n-gram occurrence

Let there be a sample $x_1, \ldots, x_k$ of occurrence of some n-gram in the text financial sequence. By assumption, this sample is distributed according to the negative binom $\beta_-(\cdot|r, p)$ with probabilitis

$$p_j = \frac{\Gamma(r+j)}{\Gamma(r)\Gamma(j+1)} p^r (1-p)^j, \;\; j = 0, 1, 2, \ldots.$$

The parameters are estimated using the maximum likelihood method [1]. For example, in the case of Sakhalin-energo we have estimates $\hat{r} = 0.18, \hat{p} = 0.4$, and in the case of TGK-2 estimates $\hat{r} = 6, \hat{p} = 0.93$. From this we can conclude that in the first case this 3-gram aii is generally less common, but is concentrated in a small time interval, and in the second case it is found in a more sparse form.

Figure 1 shows a two-dimensional diagram of parameter estimates for 3-grams from the text financial sequences under consideration. Note that the two-dimensional dia-

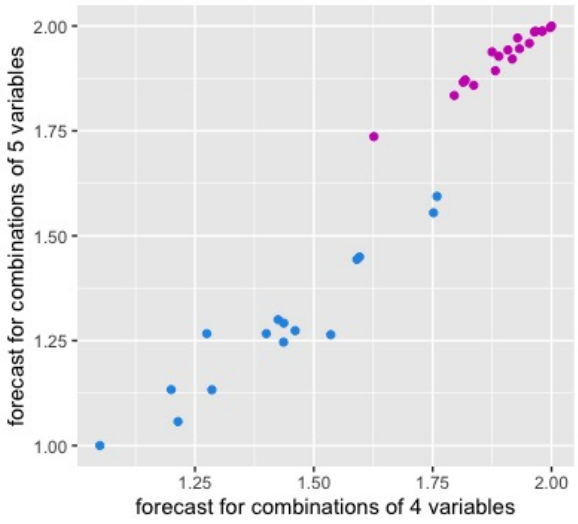


Figure 2: Illustration of the division of the oil and gas and energy industries according to the rhythm of financial instability

grams of parameter estimates in the case of analyzing the distribution of occurrences of words in ordinary texts look similar.

As a result of using this model, we obtain a set of time series characteristics that can be used to classify and systematize them. In the case of financial series, it is interesting to check whether there is a difference by industry.

3 Classification of text financial sequences by industries

The text financial sequences under consideration belong to six industries: banks,ferrous and non-ferrous metallurgy, oil and gas complex, retail, energy. Based on the NBD parameters of the most representative 3-grams: aii, aaa, iaa, eee, iii, iai, iia, aia, aai, a linear classifier with an accuracy of 0.93 can be constructed.

For this purpose, we use the partial classification method [2]. Linear classifiers are built based on the most informative combinations of 4 or 5 variables, and forecasting is performed based on averaging the obtained partial forecasts (projective classifier).

Figure 2 shows a two-dimensional diagram of projective classifiers obtained from partial classifiers by four and five variables, by which text financial sequences from the oil and gas and energy complexes can be divided.

This separability can be obtained in another way. If for each text financial sequence we construct a matrix of its occurrences in the dictionary, then using the Word2Vec neural network with the Continuous Bag of Words and Skip-Gram architectures [3], [4], [5] we perform a reduction in the dictionary dimension, then it can be found that for some components from the obtained vector representation, the accuracy of classifying

the dynamics of oil and gas and energy complex quotes reaches the same 92 percent as in the parametric method with the NBD model.

References

1. Alexeyeva N.P., Sotov A. (2013). The negative binomial model of word usage. *Electronic Journal of Applied Statistical Analysis*. Vol. **6**, Num. **1**,84-96.

2. Alexeyeva N.P., Al-Juboori F.S.Sh. (2022). About the full prediction approximation by a lot of partial predictions in case of incomplete data. Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy. Vol. **9**, No. **4**. pp. 575–589.

3. Olah C. Understanding LSTM Networks (2015). — URL: `https://colah.github.io/posts/2015-08-Understanding-LSTMs.8`

4. Mikolov T. Chen K. Corrado G. Dean J. (2013). Efficient Estimation of Word Representations in Vector Space. — CoRR.

5. Pang B., Lee L. Sentence Polarity Dataset. (2002). — URL: `https://www.cs.cornell.edu/people/pabo/movie-review-data`

A NOTE ON EXIT TIMES FOR NONLINEAR AUTOREGRESSIVE PROCESSES

A. Aliev[1], A. Dzhalilov[2], R. Fontana[3]

[1]*Great Bay University*
Guandong, CHINA
[2,3]*Turin Polytechnical University in Tashkent*
Tashkent, UZBEKISTAN
e-mail: [1]aliyev95.uz@mail.ru, [2]adzhalilov21@gmail.com

We study the exit times from a bounded interval for a nonlinear autoregressive process of order one, denoted by $\mathbb{X}(f) := \{X_n(f),\, n = 1, 2, \dots\}$ where the process is defined by the recurrence relation (1) with a continuous, contractive function $f : \mathbb{R}^1 \to \mathbb{R}^1$, a small positive noise parameter $\varepsilon > 0$, and a sequence $\{\xi_n\}$ of independent and identically distributed standard normal random variables. Klebaner and Liptser (see [1, 2]) applied the large deviation principle to obtain key asymptotic estimates for the exit times from the interval $[-1, 1]$ for linear AR(1) processes. Building on their results, G. Hognas and B. Jung [3] derived upper bounds for exit times in the case of AR(1) processes driven by several piecewise-linear maps on $[-1, 1]$. In the present work, we extend the results of Hognas and Jung by considering a broader class of piecewise continuous maps f. We show that, for this class, the asymptotic behavior of the exit times depends critically on both the slopes and the locations of the breakpoints of f.

Keywords: autoregressive process, exit time, Markov chain, large deviation principle

1 Introduction

The study of exit times plays an important role in understanding the behavior of stochastic processes, particularly in applications involving the evolution of populations (see [4], [5]), finance analysis [6] surveillance analysis (see [7]) and many others. In this work, we investigate the exit times of piecewise linear autoregressive processes driven by Gaussian distributed noise. These processes, which generalize classical linear autoregressive models, exhibit regime-dependent dynamics, making their statistical properties and first-passage times particularly interesting. Let $\mathbb{X}(f) := \{X_n(f),\, n = 1, 2, \dots\}$ be a nonlinear autoregressive $(AR(1))$ process defined recursively by

$$X_{n+1}^{(\varepsilon)}(f) = f(X_n^{(\varepsilon)}) + \varepsilon\xi_{n+1},\ n \geq 0,\ X_0^{(\varepsilon)} = x_0 \in (-1, 1), \tag{1}$$

where the contractive function $f : \mathbb{R}^1 \to \mathbb{R}^1$ is continuous, $\varepsilon > 0$ is a small positive parameter, and ξ_n is an i.i.d. sequence of standard normal random variables (innovations). We introduce the exit time from the interval $[-1, 1]$ (see [1], [3], [8]):

$$\tau^{(\varepsilon)}(f) := \min\{k \geq 1 :\ \mid X_k^{(\varepsilon)}(f) \mid \geq 1\}.$$

In the study of exit times for autoregressive models, various techniques have been employed, including martingale methods, large deviation principles (LDP), and other

probabilistic approaches. Martingale techniques, in particular, have been used to derive analytical approximations for the distribution and expectation of exit times in AR(1) processes. These methods have been especially effective in analyzing Ornstein-Uhlenbeck processes within a continuous-time framework. In [1], Klebaner and Liptser established LDP for a general nonlinear autoregressive process defined recursively by

$$X_{n+1}^{(\varepsilon)}(g) := g(X_n^{(\varepsilon)}, \dots, X_{n-m-1}^{(\varepsilon)}, \varepsilon\xi_{n+1}),$$

where $\varepsilon > 0$, $m \geq 1$ and $g(x_1, \dots, x_m, y)$ is a continuous function. Furthermore, consider the linear autoregressive AR(1) sequence $\mathbb{X}(\lambda) := \{X_n(\lambda),\ n = 1, 2, \dots\}$ defined by

$$X_{n+1}^{(\varepsilon)}(\lambda) = \lambda X_n^{(\varepsilon)}(\lambda) + \varepsilon\xi_{n+1},\ n \geq 1,\ X_0^{(\varepsilon)} = x_0,$$

where $\{\xi_n\}$ is a sequence of i.i.d. random variables, $\xi_1 \sim \mathcal{N}(0, 1)$, and λ is a nonrandom constant. Applying the LDP to the linear AR(1) process $\mathbb{X}(\lambda)$ Klebaner and Liptser [1] derived an upper bound for the expectation of exit times.

Theorem 1 (see [1]). *Let $\tau^{(\varepsilon)}(\lambda)$ be the exit time from the interval $[-1, 1]$ of the random process $\mathbb{X}(\lambda)$. If $\mid \lambda \mid < 1$, then*

$$\limsup_{\varepsilon\to 0} \varepsilon^2 \log E\tau^{(\varepsilon)}(\lambda) \leq \frac{1}{2}(1 - \lambda^2).$$

Numerous authors have investigated different aspects of the exit time problem for linear autoregressive processes (see, e.g., [9], [10]). A key open problem is extending these results from linear contractive functions to nonlinear functions. In [4], Allen et al. examined examples involving nonlinear contractive functions. The simplest nonlinear extension is the case of piecewise linear functions, which serves as a natural bridge between linear and fully nonlinear models. Hognas and Jung [3] studied $AR(1)$ processes with contractive functions f that are continuous, increasing, and have a fixed point at $x = 0$. Notably, the problem of obtaining an upper bound for $E\tau^{(\varepsilon)}(\lambda)$ can be reduced to finding the infimum of certain sums (see [1])

$$S_N(y_0, y_1, \dots, y_N) := \sum_{i=0}^{N} (y_i - f(y_{i-1}))^2.$$

2 Main results

In [3], Hognas and Jung established several key results concerning sums of the form $S_N(f)$. Utilizing these results for specific piecewise linear functions, they obtained upper bounds for $\limsup_{\varepsilon\to 0} \varepsilon^2 \log E_{x_0}\tau^{(\varepsilon)}(f)$. Additionally, they computed the minimum values of $S_N(f)$ for specific piecewise linear functions f (see [3] for further details). In this work, we extend their analysis to $AR(1)$ processes, considering piecewise linear functions f with 3 breakpoints within the interval $[-1, 1]$.

We consider the class $\mathbb{M}$ of all functions $f : \mathbb{R}_+ \to \mathbb{R}_+$ that are nondecreasing, continuous and $f(0) = 0$. We introduce the following subclasses of $\mathbb{M}$ (see Figure 1).

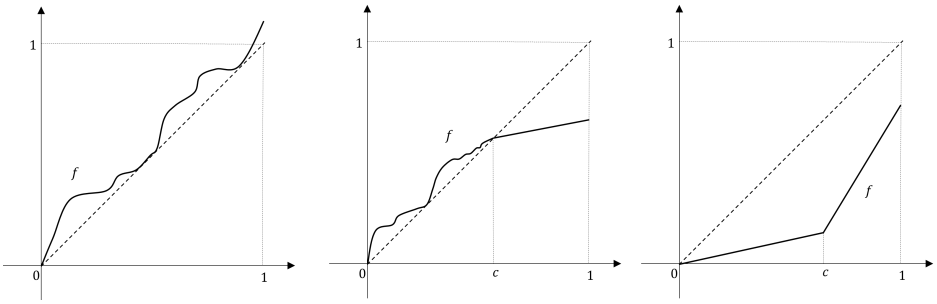


Figure 1: From left to right, the images display the set of functions $\mathbb{C}^{(up)}$, $\mathbb{C}^{(int)}$ and $\mathbb{C}^{(low)}$

- Let $\mathbb{C}^{(up)} \subset \mathbb{M}$ be a class of functions f such that $f(x) \geq x$ on $[0,1]$;
- Let $f \in \mathbb{C}^{(int)} \subset \mathbb{M}$ be a class of functions f such that $f(x) \geq x$ on $[0,c] \subset [0,1]$. Moreover, f is linear and $f(x) < x$ on outside of $[c,1]$;
- $\mathbb{C}^{(low)} \subset \mathbb{M}$ denotes a class of functions f such that $f(x) \leq x$ for $\forall x \in [0,1]$, and f is piecewise linear on $[-1,1]$ with 1 breaks. More explicitly,

$$f(x) = \begin{cases} \alpha_1 x, & x \in [0,c), \\ \alpha_2 x + \beta, & x \in [c,1), \end{cases} \quad c \in (0,1].$$

Denote by $\mathbb{M}^c$ set of functions $\varphi_{f,g} : \mathbb{R} \to \mathbb{R}$ defined for $f, g \in \mathbb{M}$ by

$$\varphi_{f,g}(x) = \begin{cases} -f(-x), & x \leq 0, \\ g(x), & x > 0. \end{cases}$$

We formulate our main results.

Theorem 2. *Let $\varphi_{f,g} \in \mathbb{M}^{(c)}$. Consider the random process $\mathbb{X}(\varphi_{f,g})$ defined by (1). The following statements hold.*

- *If $f \in \mathbb{C}^{(up)}$ and $g \in \mathbb{C}^{(up)}$, then for any $|x_0| \leq 1$,*

$$\lim_{M\to\infty} \lim_{\varepsilon\to 0} \varepsilon^2 \log P\left(\max_{1\leq k\leq M} |X_k^{(\varepsilon)}| \geq 1 \mid X_0^{(\varepsilon)} = x_0 \right) = 0;$$

- *In the case $f, g \in \mathbb{C}^{(int)} \cup \mathbb{C}^{(low)}$, there exist non-positive constants $K_1(f)$, $K_1(g)$, $K_2(f)$, $K_2(g)$ such that*

$$\lim_{M\to\infty} \lim_{\varepsilon\to 0} \inf_{|x_0|\leq 1} \varepsilon^2 \log P\left(\max_{1\leq k\leq M} |X_k^{(\varepsilon)}| \geq 1 \mid X_0^{(\varepsilon)} = x_0 \right) = \min\{K_1(f), K_1(g)\},$$

$$\lim_{M\to\infty} \lim_{\varepsilon\to 0} \sup_{|x_0|\leq 1} \varepsilon^2 \log P\left(\max_{1\leq k\leq M} |X_k^{(\varepsilon)}| \geq 1 \mid X_0^{(\varepsilon)} = x_0 \right) = \min\{K_2(f), K_2(g)\}.$$

Theorem 3. *Assume that* $\varphi_{f,g} \in \mathbb{M}^{(c)}$ *and* $\mathbb{X}(\varphi_{f,g})$ *is defined by* (1). *The following statements hold.*

- *If* $f \in \mathbb{C}^{(up)}$ *and* $g \in \mathbb{C}^{(up)}$, *then for any* $|X_0^{(\varepsilon)}| = |x_0| \leq 1$,

$$\lim_{\varepsilon \to 0} \varepsilon^2 \log E\tau^{(\varepsilon)}(f) = 0;$$

- *In the case* $f, g \in \mathbb{C}^{(int)} \cup \mathbb{C}^{(low)}$, *there exist non-positive constants* $K_1(f)$, $K_1(g)$, $K_2(f)$, $K_2(g)$ *such that*

$$\limsup_{\varepsilon \to 0} \varepsilon^2 \log E\tau^{(\varepsilon)}(\varphi) \leq -\min\{K_1(f), K_1(g)\},$$
$$\liminf_{\varepsilon \to 0} \varepsilon^2 \log E\tau^{(\varepsilon)}(\varphi) \geq -\min\{K_2(f), K_2(g)\}.$$

References

1. Klebaner F., Liptser R. (1996) Large deviations for past-dependent recursions, *Probl. Inf. Transm.* **4** pp. 23-34.

2. Klebaner F., Liptser R. (2011) Asymptotic analysis of ruin in the constant elasticity of variance model, *Theory Probab. Appl.* Vol. **55**, Num. **2**, pp. 291-297.

3. Hognas G., Jung B. (2025) Exit times for some nonlinear autoregressive processes, *Modern Stoch. Theory Appl.* pp. 1-18, https://doi.org/10.15559/25-VMSTA277.

4. Allen L.J.S., Fagan F., Hognas G., Fagerholm H. (2005) Population extinction in discrete-time stochastic population models with an Allee effect, *J. Difference Equ. Appl.*, **11**, pp .4-5, 273-293.

5. Brockwell P.J., Davis R.A. (2016) Introduction to Time Series and Forecasting. 3rd edition. Springer Texts in Statistics, Springer International Publishing Switzerland.

6. Fabrizio Lillo, Giulia Livieri, Anton Solomko, Stefano Marmi, Sandro Vaienti, Analysis of bank leverage via dynamical systems and deep neural networks, *SIAM Journal on Financial Mathematics*, Vol. **14**, Num. **2**, pp. 598-643, 2023,

7. Frisen M., Sonesson Cr. (2006). Optimal Surveillance Based on Exponentially Weighted Moving Averages, *Sequential Analysis*,**25**, pp. 379-403.

8. Jung B. (2013) Exit times for multivariate autoregressive processes, *Stoch. Proc. Appl.* **123**, pp. 3052-3063.

9. Basak G.K., Ho K-W.R., (2004) Level-crossing probabilities and first-passage times for linear processes, *Adv. Appl. Probability* **36**, pp. 643-666.

10. Baumgarten C., (2014) Survival probabilities of autoregressive processes, *ESAIM: Probability and Statistics* **18**, pp. 145170.

ADAPTIVE CHI-SQUARE TEST FOR GOODNESS-OF-FIT

D.E. Andreev[1]

[1]*Lomonosov Moscow State University*
Moscow, RUSSIA
e-mail: [1]danila.andreev@math.msu.ru

This work proposes a modification of the goodness-of-fit chi-square test. We find the limiting distribution of the corresponding test statistic. A comparative analysis of the test's power is also carried out against well-known tests.

Keywords: Goodness-of-fit, adaptive chi-squared test, weighted sums of chi-squared random variable

1 Introduction

The classical goodness-of-fit chi-square test was first introduced by Karl Pearson in 1900. He also established the limiting distribution of the corresponding test statistic, which made the method especially convenient for practical use in the pre-digital era. The idea of the test is to divide the real line into m intervals and count the number of observations in each interval. The test statistic is then computed as

$$\chi_n = \sum_{i=1}^{m} \frac{(\mu_i - np_i^0)^2}{np_i^0},$$

where μ_i is the number of observations in the i-th interval, n is the sample size, and p_i^0 is the theoretical probability of the i-th interval under the null hypothesis.

One major drawback of this approach is its sensitivity to the choice of the partitioning scheme, that can change the power of the test. Thus, it's important to construct chi-square tests that adapt the partitioning to the data. In particular, the approach proposed by Heller, Heller, and Gorfine [1] offers an interesting direction in this context.

In their work, the authors suggested using all possible partitions of the support. The problem with this approach is that the test statistic does not have a known limiting distribution. Because of this, a permutation method was used to calculate the p-value. This approach is slow to run.

Our approach also uses a set of partitions, but the test statistic has a limiting distribution. This makes it possible to compute p-values much faster.

2 Model

In this paper, we consider the following testing problem. Suppose we have a sample of i.i.d. random variables $X_1, \ldots, X_n$ with cumulative distribution function F. Let the null hypothesis be $H_0\colon F = F_0$ and the alternative be $H_1\colon F \neq F_0$.

We propose the following adaptive version of the chi-square test. We first divide the real line into N initial intervals such that each interval has equal probability under F_0. These intervals are then grouped into k cells, where each cell is a union of several adjacent intervals. The length of a cell is defined as the number of original intervals it contains.

Let p denote the probability of falling into each of the N intervals under the null hypothesis.

For each possible grouping into k cells, we compute the chi-square statistic, treating each cell as a single interval. The final test statistic is defined as the sum of all chi-square statistics over all such groupings.

Let us now formalize this. Let V be the set of all possible partitions. Then the statistic can be written as:

$$S_n = \sum_{\mathrm{v} \in V} \sum_{j^{\mathrm{v}}=1}^{k} \frac{(\nu_{j_1^{\mathrm{v}}} + \cdots + \nu_{j_l^{\mathrm{v}}} - lpn)^2}{lpn},$$

where l is the length of the cell indexed by j^{v} in the partition v. Here, $\nu_{j_r^{\mathrm{v}}}, r \in \{1, \ldots, l\}$ denotes the number of observations in the r-th interval of the cell indexed by j^{v}.

This statistic can also be expressed in the form:

$$S_n = \frac{1}{np} \sum_{i,j=1}^{N} \alpha_{ij}(\nu_i - np)(\nu_j - np),$$

which can be written in a matrix form as $S_n = X_n A X_n^T$, where

$$X_n = \left(\frac{\nu_1 - np}{\sqrt{np}}, \ldots, \frac{\nu_N - np}{\sqrt{np}} \right),$$

and $A = \{\alpha_{ij}\}$ is an $N \times N$ matrix. The values α_{ij} depend on the parameters N and k. The matrix is too complicated to present here, so we will describe it during the report.

We now state the main theorem:

Theorem 1. *Under the null hypothesis, the following convergence in distribution holds:*

$$S_n = X_n A X_n^T \xrightarrow{d} Z^T M Z, \quad n \to \infty,$$

where Z is a column vector of dimension $N-1$ with a standard multivariate normal distribution, and M is a symmetric positive definite matrix of size $N-1$.

The structure of the well-known matrix M will be discussed in the report.

The limiting random variable can be represented as follows:

$$Z^T M Z \sim \sum_{i=1}^{N-1} \lambda_i W_i^2,$$

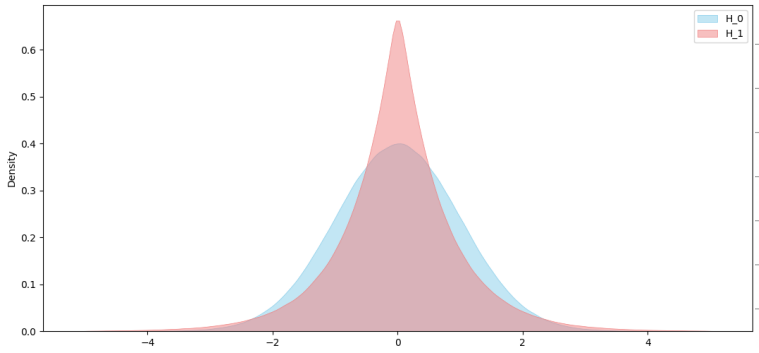
where $W_i \sim \mathcal{N}(0,1)$ are independent standard normal random variables, and $\lambda_i, i = 1, \ldots, N-1$, are the eigenvalues of the matrix M.

Several methods for approximating such weighted sums of chi-squared random variables are discussed in [2]. These methods were used to evaluate the power of the proposed test.

3 Examples of Results

We examine the proposed test for different values of the hyperparameter N. It is compared with the classical chi-squared test, the Kolmogorov-Smirnov test, and the Cramer-von Mises test.

As an illustrative example, consider the case $H_0 : F = F_0$, where $F_0 \sim \mathcal{N}(0, 1)$, and the actual data come from the distribution Laplace$(0, 1/\sqrt{2})$. The left plot on Figure 1 shows the probability density functions, and the table on the right presents the power values.



Test	n=100	n=150
adaptive_with_N=6	0.514	0.7164
adaptive_with_N=7	0.5496	0.7483
adaptive_with_N=8	0.5277	0.7336
adaptive_with_N=9	0.5286	0.7416
adaptive_with_N=10	0.5229	0.723
adaptive_with_N=11	0.5192	0.7268
adaptive_with_N=12	0.5139	0.7189
adaptive_with_N=13	0.4989	0.7079
adaptive_with_N=14	0.4948	0.7131
chi2_classic	0.4795	0.6684
kstest	0.2782	0.4203
cramervonmises	0.2031	0.369

Figure 1: Results: density functions (left) and comparison of test power (right)

References

1. Heller, R., Heller, B., Gorfine, M. (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *J. Machine Learning Research*. Vol. **17**, No. **1**. P. 978–1031.

2. Bodenham, D.A., Adams, N.M. (2016). A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*. Vol. **26**. P. 917–928.

ASSESSMENT OF THE STATE OF AN AC OVERHEAD LINE BY THE RELINEARIZATION METHOD

A.B. Balametov[1], T.M. Isayeva[2]

[1]*Azerbaijan Scientific-Research and Design-Prospecting Power Engineering Institute*
[2]*Azerbaijan State Oil and Industry University*
Baku, AZERBAIJAN
e-mail: [1]balametov.azniie@gmail.com, [2]tarana.isa@gmail.com

Information about the current mode of the electric power system (EPS) is received by the dispatch control centers in the form of telemetry and tele-signals by the SCADA complexes, and from phasor measurement units (PMU) devices. Tele-measurements include information about the mode parameters, and the state of the switching equipment. Since the system of equation of state is nonlinear, the problem of state estimation is traditionally solved using iterative methods. This article presents the results of solving the state estimation problem using a method based on the Kipnis-Shamir re-linearization method, which allows solving it by non-iterative method. The results of the solution are given on the example of a power transmission line with 500 kV voltage.

Keywords: measurements, state estimation, PMU, polynomial equations, relinearization

1 Introduction

The main trend in development of the modern electric power industry is a control intellectualization. Traditionally, information about the current EPS mode $Y = [P_{ij}, Q_{ij}, I_{ij}, P_i, Q_i]$ entered the dispatcher control centers in the form of telemetries and tele-signals [1, 2, 3, 4].

Since for monitoring the state of the entire power system as a whole, the telemetries from SCADA (Supervisory Control and Data Acquisition) are insufficient and contain errors, for specification of telemetries and calculating unmeasured parameters the state estimation (SE) methods are used.

To monitoring, analysis and operational control of the EPS after SE, the calculation of the steady state (current state) of the electric power systems (EPS) is performed.

In modern conditions the EPS control requires real-time execution of SE of large and complex power systems.

SCADA complexes receive and process distant information once a second, without synchronizing measurements in astronomical time. New measuring equipment - PMU (phasor measurement units) - has been applied with invention of satellite communication systems. Unlike SCADA, PMU measurements are $Y = [U_i, I_{ij}, \delta_i, \varphi_{ij}]$.

Measuring systems for monitoring, control and protection of the power system (WAMS) consisting of PMU devices allow of obtaining a more real state of the power system [5, 6].

State estimation of the entire EPS based on the PMU measurements only is currently impossible due to the high cost of the corresponding equipment; therefore, they are usually installed at the most critical facilities.

The mathematical basis of the problem of the SE of EPS is the least square method.

2 Traditional state estimation

At classical formulation of the SE problem, the criterion

$$\varphi(x) = (\bar{y} - y(\hat{x}))^T R_y^{-1} (\bar{y} - y(\hat{x})) \to min,$$

is minimized, where $x = (\delta, U)$ is the state vector, consisting of magnitudes U and phase angles δ of voltages of all nodes of the EPS circuit, except the basic node phase; $y = f(x)$ – measured mode parameters; $z = f(x)$ – unmeasured mode parameters; R_y – is a diagonal matrix, the elements of which are the measurement dispersions [1, 2, 3, 4].

The state equations are nonlinear, therefore, the SE problem is solved by the iterative method, for example, the weighted least-squares method.

At each iteration, corrections:

$$\Delta x_i = \left[H_i^T \cdot R_y^{-1} \cdot H_i\right]^{-1} \cdot H_i^T R_y^{-1} \left[\bar{y} - y(x_i)\right],$$

and the next approximation $x_{i+1} = x_i - \Delta x_i$ are calculated, where H is measurements Jacobi matrix.

The following initial information can be used as initial approximations of the state vector:

- measurements;
- pseudo-measurements;
- rated values of voltage magnitude and zero values of voltage phases.

Then all unmeasured mode parameters are calculated through the state vector.

To solve the SE problem, the method of test equations (TE) was developed and implemented in the form of program in [2]. The TEs are steady state equations. These equations include measured mode variables and variables calculated through measured ones only.

When PMU and SCADA measurements are used together, the SE problem retains all the disadvantages inherent in traditional state estimate;

- problems in validation due to significant difference in accuracy of PMU and SCADA measurements;
- bad conditionality of the Jacobi matrix and due to this fact, a slowing down of the convergence of the iterative process.

3 Using of phasor measurements in solving of the EPS SE problem

Installation of PMU in the EPS nodes allows of using new high-accuracy measurements. At that, the redundancy of measurements increases, which contributes to the detection of glaring errors in telemetries and improves the quality of the state estimation. The main types of measurements received from the PMU are magnitudes and phases of nodal voltages (U_i, δ_i) and currents (I_{ij}, φ_{ij}) in outgoing lines.

The initial information for the SE problem is SCADA and PMU measurements, physical and calculated PMU and PM of power flows. The measurement accuracy of the "calculated" PMU is almost equal to the measurement accuracy of the physical PMU. Accuracy of pseudo-measurements of power flows is significantly higher than the accuracy of telemetries in SCADA. This is due to the high accuracy of the PMU measurements. The TEs method by using PMU measurements also allows of checking the quality of SCADA measurements.

For example, when a PMU is installed in separate node, each PMU installed in the node can provide measuring the magnitudes and phase of the voltage in that node and the magnitudes and phases of currents in the outgoing lines. Independent voltage measurements in one node can be used for the validation of these measurements [1, 2].

Iterative methods work well for state estimation, but these methods require an initial approximation and can encounter convergence problems if the initial approximation is too far from the actual state of the system. Large dimension of circuits, complexity and need for the high-speed performance require the development and applying of special algorithms and computational procedures for the SE.

Traditional state estimation methods do not meet the speed requirements. PMU measurements are carried out with a high sampling rate. Therefore, it is possible to estimate the state of individual elements of the EPS (power plants, substations, electrical network zones) in the "rate of process" with very high accuracy.

Linear state estimation of EPS based on PMU measurements is performed in one iteration [6].

In this case, the state vector and the measurement vector are, respectively, equal to

$$x = \begin{bmatrix} \dot{U}_i = U'_i + j \cdot U''_i \\ \dot{U}_j = U'_j + j \cdot U''_j \end{bmatrix}, \bar{y} = \begin{bmatrix} \dot{U}_i = U'_i + j \cdot U''_i \\ \dot{U}_j = U'_j + j \cdot U''_j \\ \dot{I}_{ij} = I'_{ij} + j \cdot I''_{ij} \\ \dot{I}_{ji} = I'_{ji} + j \cdot I''_{ji} \end{bmatrix},$$

The measurement vector is related to the EPS state vector as $\bar{y} = H \cdot x$,where

$$H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \dot{Y}_{ij} + \dot{Y}_{i0} & -\dot{Y}_{ij} \\ -\dot{Y}_{ij} & \dot{Y}_{ij} + \dot{Y}_{j0} \end{bmatrix},$$

is the Jacobi matrix.

The linear state vector is calculated as:

$$x = \left[B^T R^{-1} B\right]^{-1} B^T R^{-1} z.$$

4 State estimation based on re-linearization method

The SE considered in [7] is based on the Kipnis-Shamir relinearization method [7]. In this method measurement equations, which are the voltage value at the node of the power line and the equations for the driving power of the node are formulated using rectangular coordinates of the bus voltages. At such formulation nonlinear measurement equations become quadratic voltage polynomials [8]. Then the method uses two transformations of the original system to the high-dimensional equations system with the quadratic variables to solve using non-iterative method. At accurate measurements this method gives the same results as the weighted least squares method.

The initial data required for the method is the system topology, information about the mode parameters and measurements from the system.

If the transmission line parameters are expressed using the π-model and the measurements are voltage magnitudes and linear flows, then the measurement equations have the form [8, 9]:

$$U_i^2 = U_{iR}^2 + U_{iI}^2; U_j^2 = U_{jR}^2 + U_{jI}^2;$$

$$P_{i,j} = g_{i,j}\left(U_{iR}^2 + U_{iI}^2 - U_{iR}U_{jR} - U_{iI}U_{jI}\right) + b_{i,j}\left(U_{iI}U_{jR} - U_{iR}U_{jI}\right);$$

$$Q_{i,j} = b_{i,j}\left(U_{iR}^2 + U_{iI}^2 - U_{iR}U_{jR} - U_{iI}U_{jI}\right) + g_{i,j}\left(U_{iR}U_{jI} - U_{iI}U_{jR}\right) + bs\left(U_{iR}^2 + U_{iI}^2\right);$$

$$g_{ij} = \frac{R_{ij}}{Z_{ij}^2}, b_{ij} = \frac{X_{ij}}{Z_{ij}^2}, Z_{ij}^2 = R_{ij}^2 + X_{ij}^2,$$

where i is the sending node; node j is the active power receiving node, R_{ij}, X_{ij} and bs are the active resistance of line, reactance and conductivity to earth, respectively.

Since these equations are linear with respect to the quadratic voltage terms (U_{iR}^2; U_{iI}^2; $U_{iR}U_{jR}$, *etc.*), they can be represented in matrix form

$$A_\xi \xi = C, \tag{1}$$

where C, ξ, A_ξ are the vector of measured values, the vector of quadratic voltage variables, and the matrix of coefficients for ξ, respectively. The vector ξ consists of the quadratic variables of the real and imaginary parts of the voltages denoted by $x_i x_j$, where the indexes i and j are not associated with the numbers of the nodes.

5 First transformation of variables

Transformation of variables is performed, and system (1) is rearranged in the following form

$$[AB]\begin{bmatrix} Y \\ Z \end{bmatrix} = C$$

where A contains linearly independent columns A_ξ, and B contains the remaining columns A_ξ, Y is vector of elements ξ corresponding to A, and Z is vector of elements corresponding to B.

Let us denote the quadratic variables $x_i x_j$ in Y as $y_1; \ldots; y_{N_y}$, and by N_y we denote the total number of variables Y. Quadratic variables $x_i x_j$ in Z we denote as $z_1; \ldots; z_{N_z}$ in the order and through N_z the total number of Z-variables.

In addition, all quadratic variables containing the imaginary component of the balancing node and the corresponding columns of the matrix are excluded from the system, in the balancing node since a zero imaginary component is specified.

The Y variables in the rearranged system can now be expressed in terms of Z variables and measurement values C:

$$Y = d + D \cdot Z, \quad d = \left(A^T A\right)^{-1} A^T C, \quad D = -\left(A^T A\right)^{-1} A^T B.$$

6 Second transformation of variables

At this stage combinations of paired products of quadratic variables are formed according to certain rules. Correct paired products meet the condition:

$$s_{ij} s_{pq} = (x_i x_j)(x_p x_q) = (x_i x_p)(x_j x_q) = s_{ip} s_{jq}.$$

These pair product ratios are used to impose additional constraints on the unknowns so that a correct solution can be obtained. For the paired products to be valid, the s_{ij} and s_{pq} must exist among the set of quadratic variables Y and Z, and the s_{ij} and s_{pq} can not be the same pair as s_{ip} and s_{jq}. For each correct paired product, one equation can be generated in the form: $s_{ij} s_{pq} - s_{ip} s_0$. More details on this can be found in [8].

7 Simulation

The simulation was carried out for 350 km long 500 kV power transmission line (PTL) (see Figure 1). In [10], simulation of the mode of such line was carried out by using the equations of the line with distributed parameters.

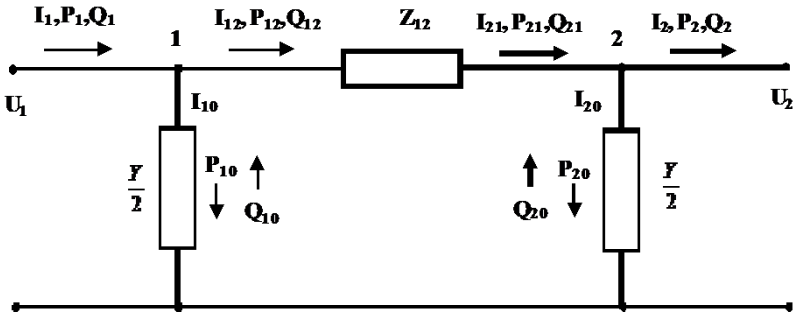


Figure 1: Two-node π-scheme of the PTL

OHL parameters in relative units $R = 0.0046; X = 0.04186; g = 0.014; b = 3.2$. The measurements are: U_1 = 520.06 kV; $U2$ = 490 kV; P_{12} = 935.18 MW; Q_{12} = 80.07 MVAr. The results of calculation of the steady state of the PTL are presented in the table 1.

Table 1: Results of steady-state of the PTL

Bus	Voltage	Ang	Generation		Load	
	Mag (pu)	(deg)	P, MW	Q, MVAr	P, MW	Q, MVAr
1	1.042	0	935.18	80.07	0	0
2	0.973	-22.64	0	0	900	50
Branch	From Bus	To Bus	From Bus Injection1		To Bus Injection 2	
			P, MW	Q, MVAr	P, MW	Q, MVAr
1	1	2	935.18	80.07	-900	-50

Total power losses are: 35.184 MW 362.76 MVAr.
The bus admittance matrix is

$$YBUS = \begin{pmatrix} 2.30242 & -22.03152i & -2.29542 & +23.66652i \\ -2.29542 & +23.66652i & 2.30242 & -22.03152 \end{pmatrix}.$$

Forming of measurement matrixes and coefficients C, A, B, d, D.
Passing to the original designations, we get:

$$xSE_1 := \sqrt{ynew_1} = 1.042,$$

$$xSE_3 := \delta 1 = 0,$$

$$xSE_4 := -\sqrt{z_1} = -0.3655,$$

$$xSE_2 := \sqrt{ynew_2} = 0.90174.$$

8 Conclusion

The advantage of the SE based on the re-linearization method is that it does not have the disadvantages of the convergence problem.

Compared to traditional weighted least squares method, the non-iterative method requires more measurements for observability. The traditional method can solve cases with 3 or more measurements. This is because the non-iterative method tries to compute the solution, while the least squares method repeats in the direction of the solution.

The results of applying the method on 500 kV PTL is presented.

References

1. Schweppe F.C., Wildes J. (1970). Power system static state estimation. Part 1: exact model. *IEEE Trans. On Power Systems*. NUum. **1**, pp. 120-125.

2. Gamm A.Z. (1976). Statistical methods for assessing the state of electric power systems. *M.: Nauka*. pp. 17-24.

3. Gamm A.Z. [et. al.] (1983). *Evaluation of the state in power engineering*. Moscow: Nauka.

4. Gamm A.Z., Glazunova A.M., GrishinYu.A., Kolosok I.N., Korkina E.S. (2009). Development of algorithms for assessing the state of the electric power system. *Electricity*. Num. **6**, pp. 2-9.

5. Phadke A.G. (2002). Synchronized Phasor Measurements. A Historical Overview. *IEEE/PES Transmission and Distribution Conference*. Vol. **1**, pp. 476-479

6. Phadke A.G., Thorp J.S. (2008). *Synchronized Phasor Measurements and Their Application, Springer Science + Business Media*.

7. Kipnis A., Shamir A. (1999). Cryptanalysis of the HFE public key cryptosystem. *Advances in Cryptology - CRYPTO '99*. pp. 19–30.

8. Fardanesh B. (2012). Method and systems for power systems analysis: a non-iterative state solver/estimator for power systems operation and control. *U.S. Patent 20 050 160 128*.

9. Jiang XT [et al.] (2013). Power system state estimation using a non-iterative direct state calculation method. *Presented at CURENT annual site visit and industry conference, Knoxville, TN*.

10. Balametov A., Khalilov E., Isayeva T.M. On the increasing of accuracy of power transmission lines modes mathematical modeling. *Proceedings of the 6th International Conference on Control and Optimization with industrial Applications, Baku, Azerbaijan*. pp. 98-101.

SHORT-TERM FORECASTING AND NOWCASTING OF REAL GDP USING COMBINED FORECASTS BASED ON MIDAS REGRESSION MODELS

N.D. Bazhanova[1], V.I. Malugin[2]

[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]nadia.bazhanova@gmail.com, [2]malugin@bsu.by

A model with mixed data sampling MIDAS and combined forecasts based on them has been developed for short-term forecasting and naucasting of the real GDP growth rates of the Belarusian economy. A comparative analysis of the forecast accuracy for the constructed models indicates their effectiveness.

Keywords: high-frequency data, regression model MIDAS, real GDP short-term forecasting and nowcasting, combined forecasts, Belarusian economy

1 The problem and purpose of research

Key macroeconomic indicators, like Gross Domestic Product (GDP), are produced quarterly by the National Statistical Committee of Belarus, while timely measures such as the Composite Index of Economic Sentiment (CIES) are published monthly. The quarterly GDP estimate is released 90 days after the quarter ends.

Classical regression models require data of the same frequency. Data alignment can be achieved by aggregating high-frequency variables or interpolating low-frequency ones, though the latter is rarely used. Aggregation may lead to a loss of information about the explanatory variable dynamics, reducing model accuracy.

This study employs MIDAS (Mixed Data Sampling Model) [2] to incorporate high-frequency data for forecasting macroeconomic indicators in Belarus [3, Malugin, 2024].

The objective is to develop MIDAS regression models for short-term forecasting and nowcasting of real GDP growth rates in Belarus, using monthly economic indicators. A comparative analysis of forecast accuracy is conducted based on the constructed mixed data models and combined forecasts.

The research addresses the following tasks: 1) testing the stationarity of annual GDP growth rates and seasonally adjusted CIES; 2) constructing econometric models U-MIDAS, MIDAS-GETS, and MIDAS-GETSIS; 3) generating combined forecasts using traditional methods; 4) comparing the forecasting capabilities of the models for predicting growth rates in the Belarusian economy based on CIES.

2 Mixed frequency models

During the research, the following specifications of MIDAS models were considered: 1) U-MIDAS model; 2) MIDAS-GETS, implementing the 'general to specific' approach; 3) MIDAS-GETSIS, a modification that incorporates dummy variables.

The U-MIDAS model (unrestricted mixed data sampling) is defined as [4]:

$$y_t = \mu + \alpha_1 y_{t-1} + \sum_{k=1}^{K} z_{k,t-1} + \sum_{l=1}^{L} \gamma_l d_{l,t} + \sum_{i=0}^{k} \sum_{j=0}^{m_i} \beta_j^{(i)} x_{tm_i-j}^{(i)} + \varepsilon_t, \ t = 1, \ldots, T, \quad (1)$$

where y_t and y_{t-1} are the low-frequency time series of the endogenous variable and its lag; $z_{k,t-1}$ represents the low-frequency time series of exogenous economic variables; $x_t^{(i)}$ denotes the higher-frequency explanatory factors; m_i is the number of observations of the explanatory variable for one value of the dependent variable. The U-MIDAS model is linear in the parameters $\left\{\beta_j^{(i)}\right\}$ and can be estimated using Ordinary Least Squares (OLS) method.

The MIDAS-GETS and MIDAS-GETSIS models implement the "general to specific" approach [1]. A repetitive algorithm sequentially removes one variable at a time from the model and checks for significance using diagnostic tests (normality, autocorrelation of residuals, etc.). If a variable is found to be insignificant, it is removed, leading to the formation of final model combinations. After calculating all possible combinations, the final models are compared using an information criterion for selection. The best model is then chosen. In the case of the MIDAS-GETSIS model, if the analyzed data contain selections and structural breaks, modeling from 'general to specific' using saturation indicators allows for the assessment of metrics by dividing them into specific samples and adding dummy variables.

3 Model construction and forecast accuracy evaluation

The following time series are used: GGDP — annual growth rates of GDP of the Republic of Belarus (RB), calculated by the income method, at quarterly frequency (in %); CESI_SA — seasonally adjusted CIES (Figure 1).

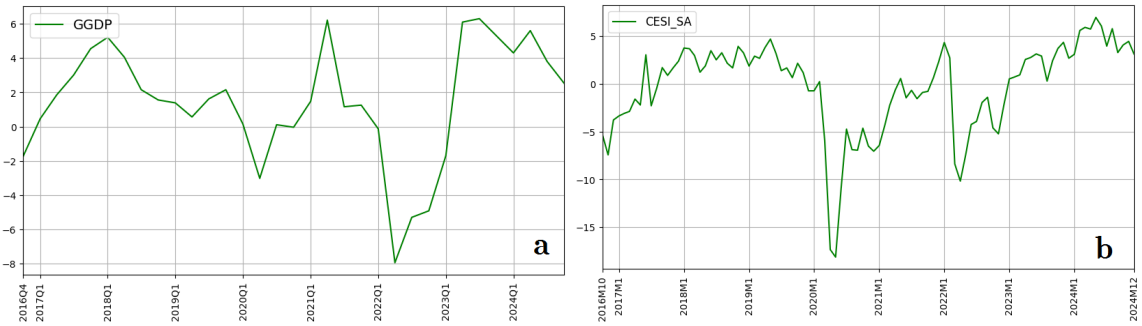


Figure 1: a) GGDP; b) CESI_SA

The stationarity of the time series was tested using the Break Point Unit Root (BPUR-test) [5]. The results for various specifications of the tested model, differing in the inclusion of a constant c and a linear trend t (Table 1), indicated that both GGDP and CESI_SA are stationary at a significance level of 0.05.

Table 1: Results of time series testing using the BPUR test

Time series	Model includes	ADF-statistics	Critical values at the sig. level $\varepsilon = 0.05$	Moments of struct. changes	Type of model
GGDP	c	-6.134	-4.444	2022Q3 (AO)	TS
CESI_SA	c, t	-5.843	-4.860	2019M10 (AO)	TS

3.1 Results of the MIDAS model constriction and combined forecasts

In this study, forecasts for future periods were made using the proposed models with an expanding window, evaluated from 2016Q4 to 2023Q2. Starting in 2023Q2, the evaluation period was extended by one quarter, with forecasts calculated for 2023Q3 to 2024Q4. Six point forecasts of GDP growth rates were compared with actual data, and forecast accuracy metrics, including MAE and RMSE, were calculated.

The U-MIDAS, MIDAS-GETS, and MIDAS-GETSIS models utilized annual GDP growth rates for RB as the endogenous variable. Each model included quarterly regressors: a constant, the lagged variable GGDP(-1), and the dummy variable DUM2022Q2. The high-frequency variable used was CESI_SA, with the MIDAS-GETSIS model excluding the dummy variable.

When selecting a model for further use, it is crucial to consider the model with the minimum accuracy metrics. However, over time, another method may prove more accurate, making combined forecasts an optimal solution. This is particularly relevant given changing data and external factors that impact forecasting accuracy.

Combined forecasts were constructed using various weighting approaches. Specifically, the following methods were used [6]: equal weights (MIDAS-AVER); weights from OLS (MIDAS-OLS); inverse values of the loss function (MIDAS-MSE); and inverse values of the ranks of the loss function (MIDAS-RANKS).

Figure 2 shows the actual values and forecast results of real GDP growth rates generated by the MIDAS models and the combined forecasts.

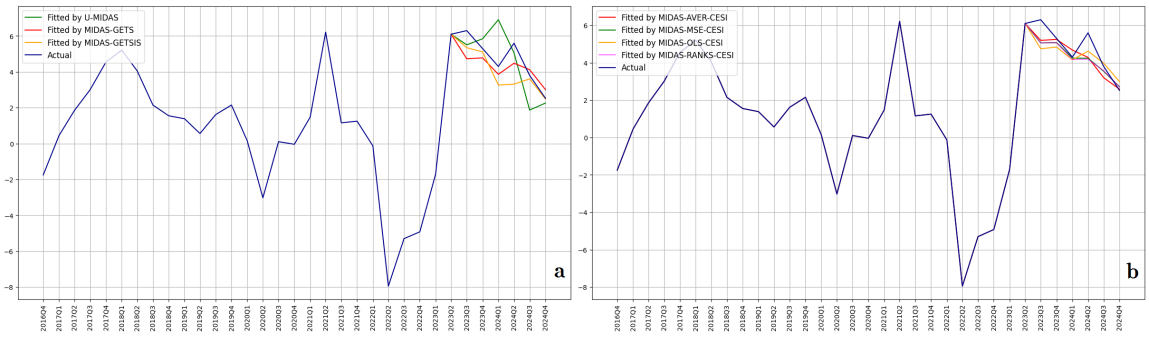


Figure 2: Actual and forecasted values of GGDP based on: a) MIDAS models; b) combined forecasts

3.2 Comparative analysis of forecast accuracy

Based on the comparative analysis of forecast accuracy metrics (Table 2), the combined forecast MIDAS-MSE shows superior accuracy compared to the other models.

Table 2: Results of evaluating the accuracy of MIDAS models and combined forecasts

Model	RMSE	MAE
U-MIDAS	1.403011	1.112751
MIDAS-GETS	0.867813	0.742120
MIDAS-GETSIS	1.100931	0.780526
MIDAS-AVER	0.759032	0.585382
MIDAS-OLS	0.801509	0.628978
MIDAS-MSE	0.768115	0.552095
MIDAS-RANKS	0.782078	0.569598

4 Conclusion

The following main results were obtained: 1) various modifications of the MIDAS regression model were constructed based on mixed-frequency data, intended for short-term forecasting and nowcasting of the real GDP growth rate of RB on a quarterly basis based on monthly economic indicators; 2) different combined forecasts were created using standard combination methods; 3) a comparative analysis of the accuracy of autonomous and combined forecasts of the target indicator was performed, identifying optimal conditions for addressing the target problems.

The obtained results, including forecast models and methods for combining forecasts, should be utilized as components of quarterly forecasting tools for assessing the target indicator prior to the release of official values.

References

1. Bauwens, L., Sucarrat, G. (2010). General-to-specific modelling of exchange rate volatility: A forecast evaluation. *Int. J. Forecasting*. Vol. **26**, No. **4**. P. 885–907.

2. Ghysels, E., Santa-Clara, P., Valkanov, R. (2002). *The MIDAS touch: Mixed data sampling regression models*. Working paper, UNC and UCLA.

3. Malugin, V.I. (2024). Short-term forecasting and nowcasting of inflation growth rates based on mixed data models. *Bank Bulletin J.*. Vol. **1**, No. **726**. P. 23–36.

4. Foroni, C., Marcellino, M., Schumacher, C.2015Unrestricted Mixed Data Sampling (U-MIDAS): MIDAS Regressions With Unrestricted Lag PolynomialsJ. Royal Statistical Society. Series A: Statistics in Society178157–82

5. Perron, P., Vogelsang, T.J. (1989). The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis. *Econometrica*. Vol. **57**, No. **6**. P. 1361–1401.

6. Kharin, Yu.S., Malugin, V.I., Kharin, A.Yu. (2003). *Econometric Modeling*. BSU: Minsk.

A STATISTICAL STUDY OF THE SPATIAL AND TEMPORAL VARIABILITY OF TEMPERATURE AND WIND FIELDS

E. Beliauskene[1], I. Ustinova[2], L. Konstantinov[3]

[1,2,3] *Tomsk Polytechnic University*
Tomsk, RUSSIA
e-mail: [1]eam@tpu.ru, [2]igu@tpu.ru, [3]lak14@tpu.ru

The study focuses on the analysis of long-term meteorological observation data obtained from a network of stations in Central Russia, aimed at identifying spatiotemporal patterns in the distribution of temperature and wind. The results may serve as a basis for improving methods of forecasting meteorological fields, as well as, for modeling and verification of atmospheric processes.

Keywords: interpolation, time series, forecasting, statistical analysis

1 Introduction

Modern meteorology often requires the processing of large volumes of data that affect many areas of human activity, including aviation, agriculture, and emergency management. Processing such data to identify hidden patterns and forecast them is challenging due to the complexity of constructing mathematical and physical models of the atmosphere, high computational costs, and insufficient accuracy of the results obtained. According to the study [1], statistical analysis methods are best suited to address this problem [2].

The aim of this work is to develop a program capable of primary processing of obtained temperature and wind field data using statistical analysis methods to convert them into a form suitable for further development of algorithms for spatial interpolation of meteorological data.

2 Main part

The work used data from daily measurements of temperature, zonal and meridional wind components for the July and January months during the 20042014 period [3]. Measurements were made at stations located near the cities of Bologoye, Smolensk, Moscow, Sukhinichi, Ryazan, and Kursk at 00:00 and 12:00 GMT. For each weather station, vertical profiles were converted into heights in geometric meters and data arrays were generated at altitudes of 0 (ground level), 100, 200, 300 and 400 m [4]. During statistical processing, the hypothesis of normality of temperature and wind distributions was tested at all the stations and heights. Additionally, spatial correlation and autocorrelation functions were constructed to assess the dependencies between measurements taken at different spatial locations.

First, the average values of meteorological variables for each station and the daily average values for all stations at all altitudes were calculated separately for 00:00 and

12:00 hours.Then, for each station at all altitudes, the deviations were determined as the differences between the actual values of the meteorological quantities at this station and the average value of the corresponding quantity at all stations for that day.

At the next stage, we studied short-term temporal dependencies (1 to 8 days). For this purpose, we calculated the autocorrelation coefficients of the meteorological fields for all stations at all altitudes separately for 00:00 and 12:00 hours. In this way, autocorrelation functions are constructed. Figure 1 shows graphs of the autocorrelation coefficients of temperature values (July, 00:00, altitudes of 300 m and 0 m, Moscow) over 8 days, illustrating the attenuation of correlation with increasing time lag. The attenuation occurs faster at altitudes above 100 m due to the reduced influence of the surface layer.

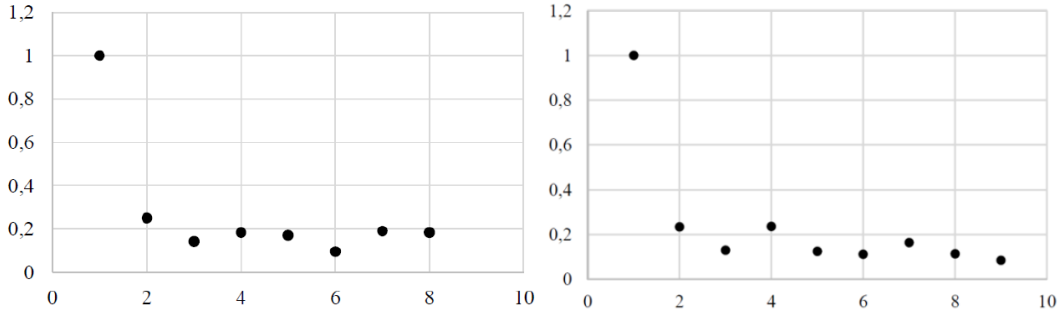


Figure 1: Autocorrelation coefficient of temperature fields: height 300 m (left picture); height 0 m (right picture)

In addition, the study showed higher autocorrelation coefficients of temperature in winter. This may be due to the fact that cold air masses can persist for a long time, providing stable low temperatures. The constructed series of autocorrelation coefficients of the zonal and meridional components of the wind show the nature of the time dependence of the wind speed on its previous values. A comparative analysis of the autocorrelation coefficients of the fields of zonal and meridional wind components at different altitudes shows significantly different dynamics of the wind regime (Figure 2). For example, at an altitude of 300 m, a faster attenuation of correlations is observed than at the level of the Earth's surface (altitude 0 m). This could be explained by the fact [5] that the relief features create more stable air flows with a 'long memory' of the process. At high altitudes, the influence of the surface weakens, air flows become more turbulent, and more quickly lose correlation with previous states.

To study the relationships between meteorological fields at various observation points, average values for each station and daily average values for all six meteorological stations were calculated. Then, daily deviations of this values from the average were calculated and the data were averaged over 6 days. Correlation matrices were also calculated. All of the above calculations were performed separately for each altitude and each time. As an example, we present the correlation coefficient values for the Moscow station (Figure 3).

Similar calculations were performed for the zonal and meridional wind components.

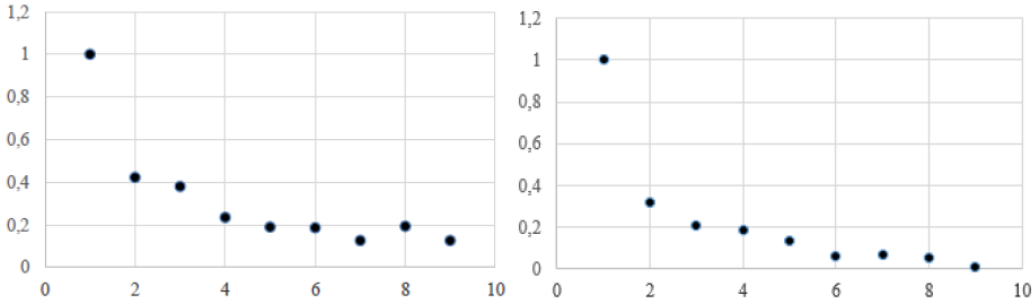


Figure 2: Autocorrelation coefficients of the zonal wind component fields, 00:00, January: height 0 m (left picture); b) height 300 m (right picture)

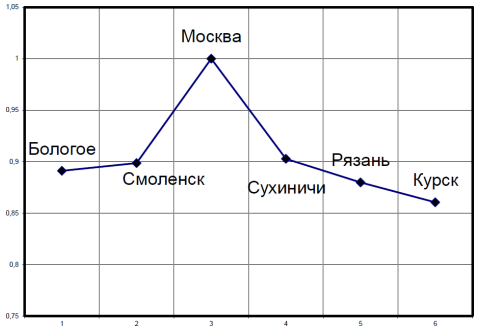


Figure 3: Spatial correlation coefficient of Moscow station, height 0 m, July, 12 hours

3 Results

To perform the presented calculations and visualize the results, a specialized software application was developed, which allows users to:

- enter time series of measurement data at different altitudes from different meteorological stations;
- calculate autocorrelation and spatial correlation functions with the ability to select maximum lag;
- visualize results in the form of graphs and diagrams;
- save numerical values of statistical characteristics in tabular format for further analysis.

The application is implemented using modern libraries for statistical data analysis and graphing, enabling comparative analysis of meteorological data collected at different stations depending on altitude, time of day, and season. This software solution significantly simplifies the study of meteorological data by allowing quick assessment of the temporal and spatial structure of the relevant fields, as well as, preparing the data for further forecasting.

4 Conclusion

The statistical analysis of meteorological fields of temperature and wind not only enhances understanding of regional climatic conditions but also offers valuable benefits in various economic sectors, including agriculture, transport, and energy. By quantifying spatial and temporal variability, these analyses support improvements in weather forecasting models and inform the development of adaptation strategies to changing weather patterns. Looking ahead, expanding the functionality of meteorological analysis software-such as adding 3D visualization of time series, implementing automatic anomaly detection, and integrating with databases for real-time updates-will further enhance the capacity to analyze and respond to complex weather dynamics. This aligns with trends in leveraging advanced data analytics and climate model outputs to support economic decision-making and resilience planning in the face of climate variability and change.

References

1. Bochenek B., Ustrnul Z. (2022). Machine Learning in Weather Prediction and Climate Analyses – Applications and Perspectives. Atmosphere. 13. [Electronic resource] Mode of access: `https://doi.org/10.3390/atmos13020180` Date of access: 27.04.2025.

2. Afanasyev V. S. (2020). Modern methods of processing and visualization of meteorological data. *Quality. Innovations. Education.* . Vol. **4**, Num. **168**, pp. 61-66.

3. University of Wyoming, Department of Atmospheric Sciences. [Electronic resource] Mode of access: `http://www.weather.uwyo.edu.` Date of access: 30.04.2025.

4. Lavrinenko A.V. (2006). *Multidimensional dynamic-stochastic models and their application in applied problems*. IOA: Tomsk.

5. Khokhlova A.V. (2023). An array of climatic characteristics of wind speed in the lower atmosphere based on aerological data. *Proceedings of the All-Russian Research Institute of Hydrometeorological Information - World Data Center. Num.* **190**, *pp. 114-120.*

MULTIPLE INSTANCE LEARNING BASED ON SAMPLE SELF-CORRECTION, FEATURE SELECTION AND ENSEMBLE CLASSIFICATION

V.B. Berikov[1], O.A. Kutnenko[2]

[1,2]*Sobolev Institute of Mathematics Siberian Branch Russian Academy of Sciences Novosibirsk, RUSSIA*

e-mail: [1]berikov@math.nsc.ru, [2]olga@math.nsc.ru

The paper considers weakly supervised learning problem, also known as multiple instance learning, or learning on multisets. A method for its solution based on the selection of informative features, filtering of the training sample and ensemble classification is developed. The results of an experimental study of the algorithm using a protein identification dataset are presented.

Keywords: weakly supervised learning, multiple instance classification, informative features, filtering of sample objects

1 Introduction

In the weakly supervised learning problem, the possible uncertainty or fuzziness of the labeling is taken into account, see review [1]. The proposed study focuses on the weakly supervised multiple instance learning problem (WSMIL) in the context of group classification, where each set, called a bag, can include different objects. The case of binary classification is considered: one of the classes is conventionally called positive and the other is called negative. A bag is labeled as positive if it contains at least one positive object (which one is unknown); otherwise the bag is labeled as negative. It is required to predict the presence or absence of positive objects for new bags.

The proposed approach to solving the problem is based on the selection of an informative feature space, selection of bags for training, and an ensemble (hybrid) approach. Automatic bag selection can be considered as a procedure for self-training of the algorithm by self-correction of the sample. The developed method was tested on the problem of protein identification. The results of applying the algorithm developed and comparing it with a number of well-known algorithms are presented.

2 Problem Description and Notation

Let there be a statistical population Γ of objects $b \in \Gamma$ described by a set of features $X = \{X_1, \ldots, X_d\}$, where d is the dimension of the feature space. Let $\vec{x} = X(b)$ ($\vec{x} \in \mathbb{R}^d$) denote the set of feature observations for object b, where $\vec{x} = (x_1, \ldots, x_d)$, $x_j = X_j(b)$, $j = 1, \ldots, d$. A metric r is given that allows one to calculate the distance between any pair of objects in the population, both in the feature space defined by X, and in any of its subspaces. Each object in the population $b \in \Gamma$ belongs to one of two classes (patterns), conventionally called positive and negative. Let us denote the set

of all possible bags by $\mathbf{B}$. Let a bag $B \in \mathbf{B}$ contain objects $b_1, \ldots, b_m$, $m = |B|$. Let $Y \in \{-1, 1\}$ denote the class corresponding to the bag. The bag is marked as positive (B^+) if $Y(B) = 1$, otherwise as negative (B^-), i.e. $Y(B) = -1$.

For an arbitrary chosen bag $B \in \mathbf{B}$, it is required to predict its membership $f(B)$ to classes. To construct the decision function, the information contained in the training sample $(B_1, Y_1) \ldots, (B_n, Y_n)$ is used, where n is the number of bags in the sample. The pair (B_i, Y_i) defines the set of objects included in the bag: $B_i = (b_{i1}, \ldots, b_{im_i})$, and its membership in the classes: $Y_i \in \{-1, 1\}$. The quality estimate (risk of incorrect classification) for the decision found can be obtained from the test sample of bags formed independently of the training sample. When solving the problem, quality metrics are used that take into account the data imbalance: $Sens = TP/(TP + FN)$ denoting the proportion of true positive predictions (sensitivity); $Spec = TN/(TN + FP)$ indicating the proportion of true negative labels (specificity); and $BA = (Sens + Spec)/2$ implying the balanced accuracy. Here TP is the number of true positive results, TN is the number of true negative labels, FP is the number of false positive predictions, FN is the number of false negative cases. We will consider the balanced accuracy (BA) as an indicator of the quality of the problem solution.

It is required to build a model that allows predicting the target feature for new bags. The problem under consideration is relevant for many applied tasks, since the annotation of the available data may not be accurate due to poor study of the problem under consideration, lack of resources for careful labeling of objects, the presence of random distortions that arise in the process of label identification, as well as due to other reasons that determine the specifics of the problem being solved.

3 The proposed method

In machine learning problems, it is often necessary to solve two additional subproblems: the selection of an informative feature space and the formation of an informative (from the point of view of a given criterion) and representative training sample. Elements of the training sample, in the case of multiple instance learning problem, are bags (sets of objects). Both the selection of informative features and the removal of noise bags (outliers) are carried out based on the analysis of the local environment of the bags. Our approach is based on the hypothesis of local compactness; the solution uses the nearest neighbor method (kNN), $k = 1$.

The following distances (where r is the Euclidean distance) are considered as the distance R between bags (groups of objects) A and B:

$$\begin{cases} R_{min}(A, B) = \min\limits_{a \in A, b \in B} r(a, b), \text{ by the "nearest neighbor" principle;} \\ R_{mc}(A, B) = \frac{1}{|A||B|} \sum\limits_{a \in A, b \in B} r(a, b), \text{ by the "average linkage" principle;} \\ R_W(A, B) = \frac{|A||B|}{|A|+|B|} \|\vec{\bar{x}}_A - \vec{\bar{x}}_B\|^2, \text{ by Ward's method.} \end{cases} \quad (1)$$

When analyzing poorly studied material, a large number of characteristics describing objects are often used. To solve the problem of constructing a subset of the most

informative characteristics, a large number of algorithms have been developed, an overview of which can be found in [2].

The presence of various types of errors in observations leads to a deterioration in the quality of the obtained patterns. The strategy of removing those training sample bags that are poorly described by the model, is sustainable if reducing sample size does not affect its representativeness. When removing outlier bags, the FRiS function (Function of Rival Similarity) [3] is used:

$$F(C, B^-|B^+) = \frac{R(C, B^+) - R(C, B^-)}{R(C, B^+) + R(C, B^-)},$$

which specifies the measure of similarity of bag C with the $^-$bags in competition with the $^+$bags. Rival similarity of bags to classes is determined by the same principle as competitive similarity between bags. As the distance from a bag to a class, we will use the distance to the nearest bag of this class. If $F(C, B^-|B^+) > 0$, then bag C is considered more similar to the negative class, otherwise to the positive one. The strategy for removing outlier bags depends on the problem peculiarities: data volume, sample balance, etc.

Using the specified distances (1), the solution variants are designed. The final decision is made by voting of decision rules. Further on, we will call the proposed method as Feature Selection and Sample Filtering (FSSF).

4 Experimental study

The efficiency of the proposed method was assessed for the problem of identifying proteins containing structures with the thioredoxin fold ($^+$bags) and those not containing these structures ($^-$bags) [4]. Figure 1 shows an example of the thioredoxin fold.

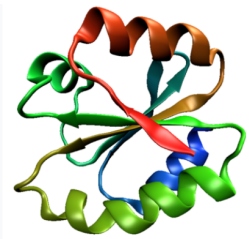


Figure 1: An example of a thioredoxin fold. The spatial topology of the alpha/beta protein fold consists of a four-stranded antiparallel beta sheet sandwiched between three alpha helices

The data table contains 26611 objects in an 8-dimensional feature space. The features describe the chemical properties of protein regions: molecular weight, hydrophobicity indices, solubility, etc. The number of bags is 193, including 25 $^+$bags and 168 $^-$bags. The bag sizes range from 35 to 189 objects. In [5], when solving this problem for classification, an SVM-based cross-validation algorithm with one $^+$bag being excluded, is proposed (the so-called jack-knife test).

The data were preliminarily normalized to the range $[0, 1]$. A random number generator selects 20 $^+$bags of the $^+$ class and 160 $^-$bags of the $^-$ class. The sample of $^-$bags is divided into 8 parts. For the distances R specified in (1), informative feature subspaces of dimension $d^* \in \{3, 4\}$ are formed for each of the 8 training samples by exhaustive search taking into account the structural features of proteins [5]. When filtering the training sample, no more than 4 outlier bags of the positive class are removed. The control is carried out in accordance with the jack-knife test. The final decision is made by the majority vote of the decision functions. The quality assessment of the solution is calculated based on the above-specified metrics.

Table 1 shows the quality estimates for the proposed algorithm and a number of known algorithms for solving the WSMIL problem. For the ease of comparison, the estimates given in [6] have been converted into comparable metrics. The comparison results show that the proposed algorithm provided a higher quality of classification. When the algorithm was running in the mode without filtering, the following estimates were obtained: $Sens = 100.00\%$, $Spec = 68.75\%$, $BA = 84.38\%$. Thus, bag filtering made it possible to improve the quality of the algorithm.

Table 1: Quality metrics for the algorithms

Values of metrics	Algorithm name						
	FSSF	Bartmip	$k_\wedge$	k_{min}	Gmil-2	Emdd	DD
$Sens$, %	100.00	75.60	83.10	85.60	75.00	64.00	87.50
$Spec$, %	71.25	76.50	78.20	78.50	75.00	63.50	33.20
BA, %	**85.62**	76.05	80.65	82.05	75.00	63.75	60.35

Table 2 shows the quality metrics of the FSSF algorithm without filtering and with filtering for all distances used in constructing the decision function separately (without forming a collective solution). The presented results show that with the specified strategy for removing positive outlier bags, the specificity of the method increases. Comparing the obtained metric values with the results in Table 1, we can conclude that when using an ensemble solution, the classification quality for the proposed method improves.

Table 2: Quality metrics for FSSF

R (1)	Metric values, without filtering			Metrics values, filtered		
	$Sens$, %	$Spec$, %	BA,%	$Sens$, %	$Spec$, %	BA, %
R_{min}	90.00	62.50	76.25	90.00	71.25	**80.63**
R_{mc}	95.00	66.87	80.94	95.00	68.13	**81.56**
R_W	95.00	60.00	**77.50**	80.00	65.63	72.81

5 Conclusion

The paper proposes a method for multiple instance weakly supervised learning using the selection of an informative feature space, filtering of training sample bags, and voting on a set of decision functions. The experimental results on protein identification dataset using the developed method are presented. The results of comparison with a number of well-known algorithms have confirmed the high efficiency of the developed algorithm. The method allows choosing the most informative sets of features, which is important for improving the quality and interpretability of the solutions, as well as self-correcting of the training sample, which makes it possible to reduce the impact of various types of errors. Further research is planned to improve the reliability and stability of weakly supervised recognition.

The work was supported by the State Contract of Sobolev Institute of Mathematics, project FWNF20220015.

References

1. Zhou Z-H. (2018). A brief introduction to weakly supervised learning. *National science review*. Vol. **5**, Num. **1**, pp. 44-53.

2. Li Y., Li T., Liu H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems*. Num. **53**, pp. 551-577.

3. Zagoruiko N.G., Borisova I.A., Dyubanov V.V., Kutnenko O.A. (2008). Methods of recognition based on the function of rival similarity. *Pattern Recognition and Image Analysis*. Vol. **18**, Num. **1**, pp. 1-6.

4. The data set Protein.csv [Electronic resource]. Mode of access: `http://www.multipleinstancelearning.com/`. Date of access: 17.09.2024.

5. Wang C., Scott S., Zhang J., Tao Q., Fomenko D., Gladyshev V. (2004) A study in modeling low-conservation protein superfamilies. *Tech. Rept. TR-UNL-CSE-2004-0003). Dept. of Comp. Sci., Univ. of Nebraska.*

6. Zhang M.L., Zhou Z.H. (2009). Multi-instance clustering with applications to multi-instance prediction. *Applied intelligence*. Num. **31**, pp. 47-68.

QUALITY OF LIFE EVALUATION: PROBLEMS, METHODS AND SURVEYS

N. Bokun[1]

[1]*Belarusian State Economic University*
Minsk, BELARUS
e-mail: [1]nataliabokun@rambler.ru

The paper describes the problems of quality of life dimension, problems of households (HH) surveys in practice of Belarusian statistics. Its contents include the next questions: quality of life, estimators and surveys, households surveys, traditional Labor Force Survey. The common estimates, design and statistical weighting of surveys are analyzed. It is proposed to use objective and subjective estimates, a combination of traditional sample methods and specialized quasir-andom, online surveys.

Keywords: quality of life, households survey, labor market, statistical weighting, employees

1 Introduction

Quality of life and the well-being of individuals and households attracts much attention of decision makers, media and the public in general. Statistical offices have experienced a growing demand for official statistics to shed light on the quality of life and the state of society "beyond GDP". The concept of well-being refers to various aspects of life that are crucial for meeting human needs, the ability to pursue one's goals, and feeling satisfied with life. While there has been much research on quality of life, there exists no uniform definition or set of indicators.

This multidimensional concept encompasses economic, social and environmental dimensions. However, tools for assessing real quality of life differ across countries, and national frameworks remain imperfect: total and regional composite indicators are not used, special subsystems of indicators (income, employment, housing, education and others) are not distinguished, and the possibilities of existing information support are not fully utilized. The directions for developing quality of life statistics are associated with calculating composite estimates and improving the methodology of household surveys.

The paper consists of the following parts:

1. Main indicators and trends
2. Composite indicators
3. Household sample survey (expenses and incomes)
4. Household survey to study employment problems

2 Main Indicators and Trends

In 2010–2024 real money incomes, wages, housing, HDI increased, the number of employed, the number of registered unemployed decreased. Since 2014 indicators of real unemployed are calculated, these parameters also decreased: 5% in 2015, 3.0% in 2024.

Table 1: Main quality of life indicators

Indicators	2010	2015	2019	2020	2021	2022	2023	2024
Life expectancy at birth, years	70.4	73.9	74.5	72.3	72.2	74.4	74.4	—
Total increase of population, ‰	-1.9	1.8	-2	-6.8	-10.1	-5.9	-4.9	-5.1
Real money incomes, % prev. year	114.8	94.3	106.1	104.7	102.1	96.4	106.4	109.7
Real wages, % previous year	115.0	97.7	106.1	104.7	102.1	96.4	111.6	113.0
Expenses on food, % total	36.8	39.1	35.7	36.8	37.6	36.8	35.4	—
Housing, m^2 per inhab.	24.6	26.5	27.8	28.9	28.9	29.4	29.9	—
Recorded crimes, per 10000 pop.	1458	1022	938	1018	943	960	930	802
Reg. unemployment, % labor force	0.7	1.0	0.2	0.2	0.1	0.1	0.1	0.1
Real unemployment, % labor force	—	5.2	4.2	4.0	3.9	3.6	3.5	3.0
GDP (PPP$) per capita	—	18096	22302	24872	27611	28426	30882	—
HDI	0.803	0.825	0.826	0.815	0.815	0.824	0.824	—

Source: Statistical yearbook, Republic of Belarus, 2024, Minsk: Belstat (in Russian), p. 15, 16, 17, 40-48, 118, 156; URL: hdr.undp.org/data-center/specific-country-data#/countries/BLR

The following quality of life trends in Belarus are observed:

- the real incomes and wages have been increasing except for the crisis 2015 and 2022 years but high share of food conception (35-37%) indicates the relatively low living standards;
- main social indicators have improved;
- the population, total labor force, total employed has been decreasing consistently, in 2024 total population decrease was 0.5%; number of employed has reduced by 0.8%;
- real unemployment rate is little variated accordingly in limits 3-4%;
- it is no composite indicator, calculated for a country and by regions;

- main statistical problems are partial and discordant set of information, data, especially connected to the health, government governance, environment, satisfaction;

One of important directions of the detail quality of life estimation is development of the sample surveys system. Nowadays it includes two surveys:

1. living standards households survey,
2. households survey to study the problem of unemployment.

3 Composite Indicators

Multidimensional concept covers all aspects of life, including composite index of subjective and objective fields such as living standards, incomes, work, employment, housing, health, education, living environment, life satisfaction. Many indicators reflect the quality of life; differences between countries are substantial (Table 2).

Table 2: Composite indicators for quality of life (QL)

Evaluation index, contributors	Dimension, indicator
Human Development index (HDI), Sen, 1993	3 dimensions: health, education, economic development
Basic Quality of life index, Diener, 1995	7 indicators to evaluate the quality of life in 77 countries
Index of social progress (ISP), Estes, 1998	41 indicators (health, education, services)
Well-being index (WBI), Mc Gillioray, 2005	As HDI except for the per capita income indicator
Social Development index (SDI), Ray, 2008	10 indicators: life expectancy, number of phone calls, power consumption
Quality of life index (QoLI), Qing, Rong, 2023	38 indicators: personal life, public life, living environment, life satisfaction
Better life index (BLI), OECD 2011, 2017, 2025	Recommended dimensions (10): subjective well-being, material living conditions, work, housing, health, knowledge and skills, physical safety, social connections, civil engagement, environmental conditions.

Source: [1–4].

From a modern prospective, composite quality of life indicator (QLI) for Belarus taking into account existing information possibilities is proposed as the next model:

$$QLI = \sqrt{QLI_o \cdot QLI_c} \tag{1}$$

$$QLI_o = \sqrt[4]{I_{ec}} \cdot I_d \cdot I_c \cdot I_e, \tag{2}$$

where QLI_o is quality of actual life index (objective estimate), QLI_c is life satisfaction index (subjective estimate), I_{ec} is economical index (GDP, retail turnover per capita, wages), I_d is demographical index (increase of the population, share of population aged 65+), I_c is social index (housing, crime rate, employment), I_e is ecological index (air polluting emissions, water discharge).

The value of each index is taken as normalized by minimax method. According to the evaluation model, QLI_o is calculated for Belarus in total and by regions. Leaders are Minsk, Minsk and Grodno regions (0.45-0.58), outsider is Mogilev region (0.120-0.150).

The main information resource for calculation of number of indicators, formation of subjective estimates is only households surveys. There are official statistics surveys or specialized mini-surveys, online-surveys.

4 Households Living Standards Survey

Households Sample Survey is conducted since January, 1995. Its main purpose is to get the information about the welfare of all population and particular demographic groups, detailed income and expenses data.

Main components of the survey are: baseline interview, four-quarterly interviews, four two-week interviews, which HH receives every quarter. More than 10 000 variables are investigated in the survey.

Survey object is households. Survey is carried out in all country regions and separately in Minsk. Annually the survey covers 0.2% or 6000 HH.

In this survey three-stage probabilistic territorial sampling is used:

1. At the first step sampling units are cities and rural soviets (village councils)
2. At the second step – local-polling districts in city and data of the soviet account in rural soviets (village councils)
3. At the third – HH

The procedure of administrative and territorial units selection repeats 1 time in 10 years, selection of polling districts and HH is carried out annually.

The methodology of weighing and extrapolation data on a general population is based on assignment of each finite unit (HH) the corresponding weight (B_i):

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \tag{3}$$

where $p_1, \ldots, p_3$ are respectively the probabilities of selecting each: city (village council); polling district in cities, village council; household within enumeration district.

Base HH weights are corrected on uninhabited apartments and non-responses by using mathematical methods.

The sample program assumes filling in some questionnaires (living conditions, personal subsidiary plots, education, health, employment), daily and quarterly questionnaires [5,6]: expenses on food and nonfood, payment of services etc.

5 Households Survey to Study the Problems of Employment

A special labor force survey has been carried out by the National Statistical Committee of the Republic of Belarus on a regular basis since 2012. The main objectives of the survey: to study the state and dynamics of demand - supply of labor, the formation of official statistical information on the number of employed, unemployed, causes and duration of unemployment, to obtain empirical statistics on labor force, employed, unemployed by sex, regions, rural, urban.

The survey is carried out quarterly, in each region and separately in Minsk. Taking into account possible non-responses, the selection share is 0.9%, or 37.2 thousand households. Sampling frame is based on the Census (2009, 2019) and includes: set of cities in each region, census enumeration districts in each selected city, villages in each selected village council, the household totality in each census enumeration district and village. The territorial probabilistic three-stage sampling is used. At the first stage, the selection units are cities, urban-type settlements, village councils, at the second - the enumeration areas, and rural settlements, at the third - households.

The methodology for statistical weighting and dissemination of data to the general population is based on assigning an appropriate weight to each holding.

Individual weights of respondents are calculated based on the results of iterative weighing: weights are calculated separately by gender, urban and rural areas; adjustments are made to the initial coefficients, first in the context of urban and rural areas, then - for five-year age groups.

Final individual weight for the respondent in each 5-year group:

$$K_i = B_B \cdot k_1, \tag{4}$$

where $B_B = \frac{S_j}{s_j}$; $k_1 = \frac{S_t}{S_\epsilon}$; S_j, s_j – population size in the j-th age and sex group based on the results of Census and survey; S_t – population size in the t-th group by urban (rural), sex (on the Census data); S_ϵ – extrapolated population size in the t-th group (by B_v).

To increase the representativeness of the data (by region, urban and rural areas, age and gender groups), it is possible to increase the number of iterations, use alternative weighing schemes [7,8].

6 Concluding Remarks

The first results of composite indicators (QLI) calculations, experience of conducting households surveys in Belarus has shown:

- To form subjective estimates and improve the representativeness by demographic groups can be extended questionnaires of official statistics surveys (individual, public life satisfaction) or used specialized face-to-face, online surveys;

- Survey problems are mainly associated with the presence of non-responses, the need of localization of sampling, regional subsamples construction; the need to use different weighing and extrapolation schemes;
- The most optimal model for selecting households is a three-stage stratified sampling; for specialized surveys is a quasirandom samples;
- The use of different weighting methods will provide very reliable information over large number of variables.

References

1. OECD. (2017). *Better Life Index – Edition 2017*. [Electronic resource]. URL: `https://stats.oecd.org/index.aspx.DatasetCode=BLI.2018-11-26` (accessed 02.02.2025).
2. *Guidelines on measurement of well-being*. Geneva.
3. Qing Y., Rong F. (2023). Quality of life assessment in 131 countries around the world (2000-2016). *Journal of Economic Statistics*. DOI:10.31085/1814-4802-2023-19-2-112-19-39, pp. 19-39.
4. World Bank. (2025). *World Development Indicators* [Database]. URL: `https://data.worldbank.org` (accessed 04.02.2025).
5. *Instructions for organization and holding Living Standards Sample Survey*. Minsk: Belstat (in Russian).
6. Bokun N. (2013). Sample Surveys of Households in Belarus: state and prospects. *Statistics in Transition*. Warsaw, pp. 110-121.
7. Bokun N. (2021). Labor market surveys in Belarus. In: *Summer School on Survey Statistics-2021: BNU Network on Survey Statistics – Virtual sessions* (03-25 September 2021). Statistics Lithuania, Vilnius, pp. 28-35.
8. *Instructions for organization and holding sample survey to study the problems of employment*. Minsk: Belstat (in Russian).

SHORT-TERM FORECASTING AND NOWCASTING OF GDP GROWTH RATES IN THE REPUBLIC OF BELARUS USING MIXED FREQUENCY VECTOR AUTOREGRESSIVE MODELS

T.A. Bout[1], V.I. Malugin[2]

Belarusian State University
Minsk, BELARUS
e-mail: [1]bout.timofey@gmail.com, [2]Malugin@bsu.by

The article presents the results of constructing vector autoregressive models based on mixed frequency data, designed for short-term forecasting and science-casting of real GDP growth rates in the Republic of Belarus based on economic indicators available with a monthly frequency of observation. A comparative analysis of the accuracy of short-term forecasts and nowcasts is carried out based on the constructed models based on mixed and aggregated data.

Keywords: mixed frequency data, short term forecasting and nowcasting, MF-VAR model, real GDP growth rates forecasting, macroeconomic indicators.

1 The relevance of the problem and the purpose of the study

The first official estimate of the real gross domestic product (GDP) is formed by the National Statistical Committee of the Republic of Belarus (NSC RB) on a quarterly frequency on the 90th day after the reporting period, i.e. with a delay of one quarter. At the same time, statistics on industry indicators and price indices are generated on a monthly basis and published in the month following the reporting month two months before the end of the current quarter. In this regard, the task of forecasting real GDP for the past, current and near future quarters based on available monthly data becomes urgent. This task of assessing the current state of the modeled process is known as the nowcasting task [1]. Obviously, the accuracy of forecasts for subsequent periods depends on the assessment of the current state.

The purpose of the study is to build vector autoregression models based on mixed data (*Mixed Frequency Vector Autoregression* – MFVAR) [2], designed for short-term forecasting one quarter ahead and tracking real GDP growth rates based on economic indicators available with a monthly frequency of observation; a comparative analysis of the accuracy of forecasts of the constructed models for mixed and aggregated data. The problem under consideration has been solved in various countries, including the Russian Federation [4]. This task has not been considered before for the Belarusian economy.

2 Description of the models

When building the models, the following tasks were solved: 1) pre-processing of time series (seasonal adjustment, logarithmization, reduction to a stationary form by reducing to growth rates); 2) selection of optimal model specifications; 3) evaluation, analysis of statistical adequacy and assessment of forecast accuracy.

The following time series of economic indicators provided by the NSC of the Republic of Belarus were used to conduct the research:

- PC_LRGDP – growth rates in logarithms of the real quarterly GDP of Belarus by sources of income use in average annual prices in 2018, million rubles, YoY (in %);
- PC_LRPP_M_SA – the growth rate in logarithms of industrial production in average annual prices in 2018, MoM (in %);
- PC_LRRET_M_SA – the growth rate in logarithms of retail turnover in average annual prices in 1995, MoM (in %);
- PC_LRINV_M_SA – growth rates in logarithms of the volume of investments in fixed assets at average annual prices in 2018, MoM (in %);
- PC_LRAGRO_M_SA – the growth rate in logarithms of the volume of agriculture in the average annual prices of 2018, MoM (in %);
- PC_LBI_BLD_M_SA – the growth rate in logarithms of the basic index of the volume of construction and installation works (January 2018 = 1), MoM (in %);
- PC_LBI_RRDH_M_SA – the growth rate in logarithms of the basic index of the volume of monetary incomes of the population (January 2018 = 1), MoM (in %);
- CESI_M_SA is a composite economic sentiment index [5].

The _SA symbols indicate a seasonally adjusted time series using the TRA-MO/SEATS method.

A constant and an impulse dummy variable dum2022q2 were also added to the model to account for the structural change in the second quarter of 2022. The MF-VAR(p) model, estimated using the least squares method, consists of 22 equations (one for the target quarterly indicator and three equations for each monthly indicator corresponding to 1, 2 and 3 months in the block). Thus, the number of estimated parameters is $22(2+22p)$, where p corresponds to the number of lags for the variables.

3 Comparative analysis of forecast accuracy

The accuracy of one-step forecasts for one quarter ahead for models based on mixed data (MF-VAR) and aggregated data (VAR) was assessed using retrospective forecasts during the model evaluation period, as well as on the basis of non-selective one-step

forecasts using the "expanding window" algorithm. According to this algorithm, forecasts for the period from the third quarter of 2022 to the fourth quarter of 2024 were made with sequential progress for one quarter. Thus, 10 quarterly forecasts were obtained for the predicted variables, on the basis of which the following forecast accuracy characteristics were calculated: RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error). The values of these characteristics for the target are shown in Table 1 for the MF-VAR and VAR models.

All the models presented in Table 1 have an optimal specification in terms of metrics. The VAR and MF-VAR models include all the described macroeconomic indicators, regardless of their significance. The VAR* model includes only the PC_LRPP_M_SA variable as significant and best in terms of metrics; the MF-VAR* model includes the PC_LRPP_M_SA, PC_LRRET_M_SA, and CESI_M_SA variables as best in terms of metrics.

Table 1: Accuracy Indicators for Forecasting Annual GDP Growth Rates of RB

Model	**RMSE**	**MAE**
Forecasting Period 2022Q3 – 2024Q4 (Retrospective Forecasts)		
VAR(5)	1.542778	1.192851
VAR*(5)	3.095332	2.563846
MF-VAR(2)	1.335281	1.064489
MF-VAR*(5)	**0.5603***	**0.3970***
Forecasting Period 2022Q3 – 2024Q4 (Expanding Window with a Step of 1)		
VAR(1)	2.495408	1.877692
VAR*(2)	2.112804	1.424611
MFVAR(1)	3.106951	2.375936
MFVAR*(2)	**1.8467***	**1.2564***

Based on the results of a comparative analysis of forecast accuracy metrics (Table 1), the MF-VAR* model has the best forecast accuracy indicators.

In Figure 1 we see a graph of one-step out-of-sample forecasts using the expanding window algorithm.

4 Conclusion

As a result of the study, it was found that the best combination of variables for predicting GDP growth in the Republic of Belarus in terms of metrics is PC_LRPP_M_SA, PC_LRRET_M_SA, CESI_M_SA, with the number of lags $p = 2$. The result also corresponds to the economic meaning of the constructed model. Indeed, industrial production (PC_LRPP_M_SA) and retail trade (PC_LRRET_M_SA) are approximations of those components that account for the largest share of the total GDP of Belarus in terms of added value; and the CESI indicator can be interpreted as the average expected value of GDP over a certain period.

Based on the results obtained, the following conclusions can be drawn:

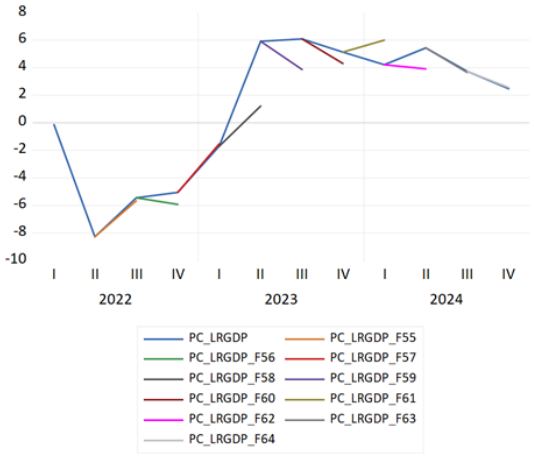


Figure 1: Forecast of non-selective values of real GDP growth in Belarus based on the MFVAR model for the best combination of variables

1) the MF-VAR model based on mixed frequency data, with the best selection of high-frequency variables, is able to make more accurate forecasts compared to the VAR model based on aggregated data in the mode of short-term forecasting and nowcasting;

2) nn the short term, the real GDP of the Belarusian economy is most influenced by such macroeconomic indicators as the volume of industrial production, the volume of retail trade and the CESI economic sentiment index.

References

1. BaBura M., Giannone D., Reichlin L. (2012). Nowcasting. *The Oxford Handbook of Economic Forecasting*, pp. 193–224.

2. Foroni C., Marcellio M. (2013). *A survey of econometric methods for mixed frequency data*. Working Paper Vol. 6, Norges Bank, 45 p.

3. Kharin Yu.S., Malugin V.I., Kharin A.Yu. (2003). *Econometric Modeling: A Study Guide*. Minsk, BSU, 318 p. (In Russian)

4. Makeeva, N.M., Stankevich I.P. (2020). Nowcasting Elements of GDP Use in Russia. *Economic Journal of the Higher School of Economics*. Vol. 10, 25 p. (In Russian)

5. Malugin V., Kruk D., Milevsky P. (2019) Economic Sentiment Index of the Belarusian Economy: Methodological, Model, and Software Tools. *Banking Bulletin Journal*, Bank Research No. 16, 31 p. (In Russian)

EXPLORING THE INTERCONNECTION BETWEEN INNOVATIVE INITIATIVES AND ESG OUTCOMES IN RUSSIAN FIRMS

V.K. Chemykhin[1]

[1]*ITMO University*
Saint-Petersburg, RUSSIA
e-mail: [1]chemykhin@yandex.ru

This study investigates the relationship between ESG performance and innovative activity in Russian companies. Using data from the RAEX ranking and the Rospatent database (2020–2024), an empirical analysis is conducted via a Bayesian HURDLE model. The results indicate that companies with higher ESG ratings demonstrate greater innovative activity, reflected in the number of registered patents. The analysis also reveals the influence of control variables, such as company size and industry affiliation, on this relationship.

Keywords: ESG, sustainable development, innovations, patents, Bayesian model, Russian companies

1 Introduction

In today's economy, sustainable development and innovation are key drivers of corporate competitiveness. The principles of ESG (environmental, social and governance factors) are increasingly integrated into business strategies, promoting long-term sustainability and financial profitability. Although innovation is not directly included in ESG metrics, there is a strong link between ESG transformation and corporate innovation [1,2,3]. The transformation of ESG encourages the adoption of new technologies to improve environmental efficiency, social conditions, and governance practices. This encourages technological advances such as the reduction of carbon emissions and the use of recycled materials. ESG-focused companies often increase R&D investments, driving the creation of new products and services that meet market and social demands. Leaders in ESG ratings show high inventive activity, exemplified by oil sector companies actively patenting sustainable technologies. High-ESG-rated firms tend to achieve better returns on R&D investments and long-term stock value growth, enabling further reinvestment in innovation [3,4]. Moreover, ESG practices indirectly boost innovation by enhancing workforce development and improving social welfare and working conditions, which raise employee skills and motivation. This study aims to analyze the relationship between ESG performance and innovation in companies, identifying groups for whom integrating sustainability metrics into innovation assessment is particularly critical. The RAEX ranking will be used to assess the impact of ESG factors on corporate innovation activity [5,6,7].

2 Data and methods

To analyze the relationship between ESG performance and innovation activity, data from the RAEX ranking (2020–2024) were aggregated into three groups based on the RAEX Europe 2023 ranking: A (A-AAA), B (B-BBB), and C (C-CCC). The top 10 percent companies were assigned to group A, the next 30 percent to group B, and the remaining 60 percent to group C. The dataset includes 160 companies for 2023, supplemented with patent and software data from the Rospatent database (20202024). Missing data were marked as Not Rated (NR). Industries with the highest data representation include Chemistry, Finance, Metallurgy and Mining, and Oil and Gas Extraction. Initially, most innovative firms belonged to group B, but after 2023, group A became dominant. The study hypothesizes that higher ESG ratings correlate with greater innovation, measured by RAEX ESG scores and annual patent counts. Control variables include company size (log of total assets), business model (industry sector), and company age. ESG ratings, size, and industry data are taken at the start of each year, with patent data recorded at year-end to ensure causality. A Bayesian HURDLE model was used to capture the two-stage patenting process: the decision to patent (binary) and the number of patents filed (zero-truncated Poisson distribution). Bayesian estimation, suitable for limited panel data, incorporates prior knowledge and provides robust posterior inferences [8,9,10,11]. Thus, the total number of patents of company i in year t depends on two factors: the probability that company i will engage in patenting activity in year t, denoted as (1- (i,t)); the parameter (i,t), which defines the distribution of the number of patents for company i in year t, conditional on the decision to engage in patenting activity. The formulas of the model are available in Appendices A and B.

3 Results

The model was estimated using the Stan programming language, employing the Hamiltonian Monte Carlo (HMC) algorithm with the No-U-Turn Sampler (NUTS). The estimation involved 5,000 iterations following 4,000 warm-up iterations, which were excluded from the final results but essential for model calibration and posterior distribution convergence. Data preprocessing and postprocessing were conducted in R, integrated with Stan via the cmdstanr package. Results are presented as posterior distributions of parameters and contrasts, where contrasts represent differences between rating groups. Posterior distributions provide estimates of both central tendencies and uncertainty. The analysis reveals that companies in rating group A (highest ESG rating) tend to have a higher number of patents compared to groups B and C, whose patent distributions appear visually similar. Notably, companies without an ESG rating exhibit significantly fewer patents, and since most unrated companies later fall into group C, the patent count for group C may be somewhat overestimated. Posterior probabilities of engaging in patenting activity (1 -) indicate similar innovation tendencies across all groups, with slightly higher probabilities for groups A and B. However, smaller sample sizes in groups A and B result in greater variability and less statisti-

cal significance compared to groups C and unrated companies. A clear trend shows increasing median expected patent counts with higher ESG ratings, with group A leading, followed by groups B, C, and unrated companies. Uncertainty is higher for group A due to its smaller sample size (13 companies). Effect estimation involved analyzing contrasts between groups for the parameter , representing the expected number of patents conditional on patenting activity. The posterior distribution of these contrasts highlights significant differences in innovation intensity across ESG rating groups. The contrast analysis reveals clear differences in patent counts across ESG rating groups. Unrated companies (NR) are expected to have the lowest patent counts, supported by positive differences between groups A, B, and NR. Group B patents less than group C, while group A patents more than group C. Comparing groups A and B shows that group A significantly outperforms group B in patent activity. Thus, among patenting companies, those with an A rating exhibit the highest innovation output. Economically, these results suggest that companies with strong ESG practices recognize the value of innovation and consistently engage in it. Companies with below-average ratings see potential for improvement both in ESG performance and innovation as part of their rating process. Analysis of the probability of engaging in patent registration indicates: no statistically significant difference between unrated companies and other groups. A significant difference between groups C and B, with group B more likely to patent than group C. While patenting activity levels are similar for companies that do patent, group C companies may refrain from patenting when patent counts are low. Thus, group B companies are more likely to patent, but group C companies patent more intensively once active. Group A companies are more likely to engage in patenting than others, though evidence is insufficient to conclusively quantify this difference. The contrasts visualizations are available in Appendices C,D.

4 Conclusion

The analysis of the relationship between ESG performance and corporate innovation, based on RAEX rankings and patent activity, yields several key findings. Companies with the highest ESG ratings (group A) exhibit greater patent activity compared to those with lower ratings (groups B and C), indicating that integrating ESG principles fosters an innovative environment and drives technology development. Interestingly, companies in group B are more likely to engage in patenting overall than those in group C. However, when companies in group C decide to patent, they tend to register a higher number of patents, suggesting that firms with lower ESG ratings may view innovation as a tool to improve their performance and competitiveness. The results confirm that ESG factors can act as innovation drivers by encouraging the development and adoption of technologies that enhance environmental efficiency, social conditions, and corporate governance. Industry affiliation and company size also influence both ESG performance and innovation, highlighting the importance of considering these factors in related analyses. Future research could explore the impact of specific ESG components on various types of innovation and uncover the mechanisms through which ESG practices stimulate innovative activity. Overall, the findings support the hypothesis of a positive link

between ESG effectiveness and innovation, suggesting that embedding ESG principles into corporate strategy can enhance long-term competitiveness and sustainability.

References

1. Guzyr, V. V. (2022). Innovative ESG transformation of firms as a global trend of sustainable development. *Economics and Innovation Management*, **1**, 33–43.

2. Tishkov, S. V. (2024). *Theory and methodology of forming innovation systems in Arctic regions.*

3. Andrey, E. (2023). ESG as an innovative tool to improve the efficiency and financial stability of financial organizations. *Procedia Computer Science*, **221**, 705–709.

4. Gazdiev, I. I. *Digital technologies as a toolkit for developing managerial decisions in corporate governance systems* [Electronic resource]. Retrieved May 8, 2025, from `https://www.dissercat.com/content/tsifrovye-tekhnologii-kak-instrumentarii-razrabotki-upravlencheskikh-reshenii-v-sisteme-korp`

5. Tikhomirov, A. A., & Kharchilava, D. Kh. (2023). The impact of ESG factors on company efficiency and investment attractiveness. *Bulletin of the Altai Academy of Economics and Law*, **5**, 312–318.

6. Bobylev, Yu. N. (Ed.). (2021). *ESG transformation textbook* [Electronic resource]. Retrieved May 8, 2025, from `https://esg-library.mgimo.ru/upload/iblock/38a/r2tczw1xi6ujkyxelg6h1jlett80gubt/BOBYLEV_Uchebnik-EUR-2021.pdf`

7. Zavyalova, E., Krotova, T. G., & Bunyakova, A. V. (2023). The impact of ESG on company competitiveness. *Law and Governance. XXI Century*, **19**(2), 62–70. (In Russian)

8. Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, **27**(8), 1–25. `https://doi.org/10.18637/jss.v027.i08`

9. Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R.* Springer.

10. Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.

11. Kshnyasev, I. A. (2010). Analysis of organism abundance: Multimodel inference as an alternative to null hypothesis testing. In *Biological systems: Stability, principles, and functioning mechanisms* (Vol. 1, pp. 348–352). Proceedings of the III All-Russian Scientific-Practical Conference, Nizhny Tagil. (In Russian)

A Probability Function of the Total Number of Patents for Company i in Year t

$$p\left(y_{i,t}|\lambda_{i,t},\theta_{i,t}\right)=\begin{cases}\theta_{i,t}, & y_{i,t}=0\\ \frac{(1-\theta_{i,t})}{\left(1-e^{-\lambda_{i,t}}\right)}\cdot\frac{(\lambda_{i,t})^{y_{i,t}}e^{-\lambda_{i,t}}}{y_{i,t}!}, & y_{i,t}>0\end{cases} \tag{1}$$

B Poisson Regression Model

$$\lambda_{i,t}=e^{\mu_{i,t}} \tag{2}$$
$$\mu_{i,t}=X_{i,t}\beta_{i,t} \tag{3}$$
$$\beta_{i,t}=\beta_0+b_i+b_t \tag{4}$$
$$b_i\sim N\left(\overline{0},\Sigma_i\right) \tag{5}$$
$$b_t\sim N\left(\overline{0},\Sigma_t\right) \tag{6}$$

Covariance matrices:

$$\Sigma_t=\begin{bmatrix}\tau_{i_1}^2 & \cdots & \rho_i\tau_{i_1}\tau_{i_K}\\ \vdots & \ddots & \vdots\\ \rho_i\tau_{i_K}\tau_{i_1} & \cdots & \tau_{i_K}^2\end{bmatrix},\quad \Sigma_i=\begin{bmatrix}\tau_{t_1}^2 & \cdots & \rho_t\tau_{t_1}\tau_{t_K}\\ \vdots & \ddots & \vdots\\ \rho_t\tau_{t_K}\tau_{t_1} & \cdots & \tau_{t_K}^2\end{bmatrix} \tag{7}$$

Notation:

- $y_{i,t}$ — number of patents of company i in year t
- $\lambda_{i,t}$ — Poisson distribution parameter for company i in year t
- $\mu_{i,t}$ — logarithm of the Poisson distribution parameter for company i in year t (used for linear modeling since $\lambda_{i,t}>0$)
- $X_{i,t}$ — $1\times K$ vector containing values of explanatory variables for company i in year t, where K is the number of explanatory variables
- $\beta_{i,t}$ — $K\times 1$ vector containing linear regression coefficients $\mu_{i,t}$ for explanatory variables of company i in year t, with coefficients varying by observation groups
- β_0 — $K\times 1$ vector containing mean (population) linear regression coefficients
- b_i — $I\times 1$ vector, where I is the total number of unique firms, containing deviations of regression coefficients from means depending on the firm (firm-specific random effects)
- b_t — $T\times 1$ vector, where T is the total number of unique years, containing deviations of regression coefficients from means depending on the year (year-specific random effects)

C Contrasts in Patent Counts Among Patenting Companies by ESG Group

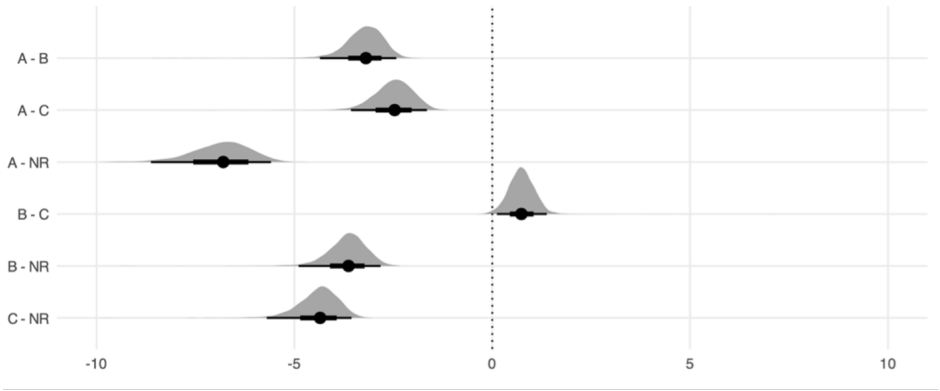


Figure 1: Contrasts in patent counts among patenting companies by ESG group

D Contrasts in the Proportion of Patenting Companies by ESG Group

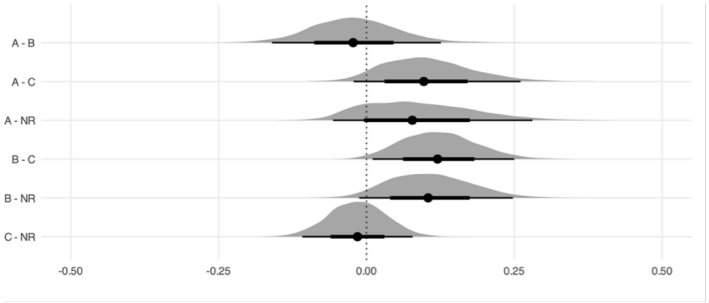


Figure 2: Contrasts in the proportion of patenting companies by ESG group

A COMPARISON OF CAUSAL SEARCH AND DOUBLE MACHINE LEARNING USING SIMULATED DATA

A.M. Chentsov[1]

[1]*MIPT, HSE, Moscow, RUSSIA*
e-mail: [1]chentsov.am@mipt.ru

The paper discusses conditional averate treatment effect estimation methods under nonlinear confounding. We use Monte-Carlo simulation with data generating processes based on convolutional neural networks with special adjustments, which allow comparison of counterfactual distributions. The generated data is subsequently analyzed through discretization and estimation of conditional distributions, as well as calculation of effects using double machine learning methods. We show that in this setup both discretized and double machine learning-based estimation perform poorly, showing very low correlation with true conditional effects.

Keywords: causal models, double machine learning, CATE

1 Introduction

Modern methods for analyzing observational data enable the estimation of heterogeneous treatment effects at any point in the covariate space, known as Conditional Average Treatment Effects (CATE), under the assumption of a known causal graph structure. However, most existing Monte Carlo-based evaluations of CATE focus on different predictive techniques, while relying on gaussian synthetic data, [1], [2], [3], which implicitly assume linear conditional mean relationships. These simplifying assumptions limit the scope of inquiry because many contemporary methods focus instead on flexible nonparametric or semiparametric estimators capable of accommodating more general forms of functional dependence. In this study, we explore data-generation mechanisms characterized by significant nonlinearities, following the structural framework outlined in Figure 1.

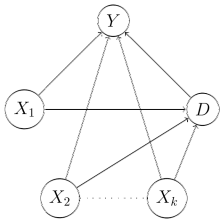


Figure 1: Directed acyclic graph model used in simulations

The treatment variable D is binary, so its conditional average treatment effect on the dependent variable Y is defined as

$$\delta(X) = \mathbb{E}[Y(X, D = 1) - Y(X, D = 0) \mid X].$$

At thirst, we estimate $\delta(X)$ by discretizing values of all variables and estimating conditional distributions of Y (this method is thereby referred as causal search). Next we estimate $\delta(X)$ by double machine learning (DML) methods [4], where the predictive models are selected through cross-validation, as discussed in [5].

2 The methods, conditions and results

Algorithm 1. Data generation with high nonlinearity and confounding.

- Gaussian $n \times k$ *i.i.d.* feature matrix is processed through a convolutional neural network with fixed depth and width parameters. Weights are initialized randomly from centered distribution[1] and applied to transform input features into intermediate representation. At the output layer, transformation according to $Z \to \frac{1}{1+e^{-Z^T w}}$ is applied to convert the outputs Z into probabilities p, where each element represents the probability for corresponding Bernoulli variable.

- Using the generated probability vector p, we sample a binary vector D according to independent Bernoulli distributions parameterized by elements of p.

- Constructing outcome variable Y through another CNN architecture with input variables Z, D. We create two additional versions of the input feature set: one augmented with $D = 1$ and another with $D = 0$, which produce the counterfactuals, denoted as $Y(X, 1)$ and $Y(X, 0)$.

Algorithm 2. Causal search estimation.

- Discretize all continuous variables based on their empirical quantiles divided into bins.

- Estimate conditional distribution of Y given observed relationships between X and D using the known dependence structure, 2.

- Estimate conditional regression functions $\mathbb{E}(Y \mid X, D = 1)$ and $\mathbb{E}(Y \mid X, D = 1)$ and evaluate them at observations, brought to the closest points on the discretization grid.

Algorithm 3. Double machine learning-based CATE estimation

- Estimate DML in a partially-linear model setup, choosing a predictor model with the smallest mean-squared prediction error in a cross-validation scheme.

- Regress the potential outcome differences on a library of tensor transformations of X to obtain CATE

- Evaluate CATE at the observation points

[1] To avoid convergence due to CLT, we use Cauchy and t(2) distributions.

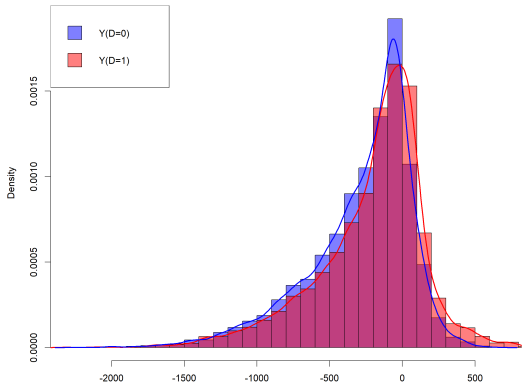


Figure 2: Counterfactual distributions of the dependent variable Y

- Obtain estimates of L^2 distance between three sets of conditional expectation functions by Monte-Carlo.

Table 1 shows the nonlinearity of the counfounders: the addition of linear controls only worsen ATE estimate (the true value being 70.0), a 5th degree polinomial of the control variables significantly improves the result.

Table 1: OLS estimation of the effect of D on Y with different controls

	(1)	(2)	(3)
	no controls	linear	polynomial deg(5)
d	11.422	−20.555	53.224***
	(19.645)	(16.335)	(8.562)
x1		61.436***	...
		(3.115)	
x2		−20.779***	...
		(3.133)	
x3		200.987***	...
		(3.072)	
Observations	10,000	10,000	10,000
R^2	0.00003	0.320	0.822
Note:			*p<0.1; **p<0.05; ***p<0.01

The results of all simulations show that the conditional average effects estimators perform poorly, achieving at best 0.2 correlation with true values for the discretized models with large number of discretization bins. On the other hand, both DML and causal search provide good estimates of averate treatment effects, usually within 10% deviation from the true value.

References

1. Jacob D. 2021. CATE meets ML – The Conditional Average Treatment Effect and Machine Learning. *Working paper*. DOI:10.48550.

2. Syrgkanis V., et. al. 2019. Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments. *Neural Information Processing Systems*.-pp. 1–15.

3. Athey S., Tibshirani J., Wager S. 2018. Generalized random forests. *Annals of Statististics*, Vol. **47**, Num. **2**, pp. 1148-1178.

4. Chernozhukov V., et. al. 2018. Double/debiased machine learning for treatment and structural parameters. *Econometric Journal* Vol.**21**, pp. C1–C68.

5. Chentsov A.M., Toropov N.I. 2024. Double machine learning estimation of effect of economy openness on deviations from uncovered interest parity. *Proceedings of MIPT* (in Russian) —Vol. **16**, Num **3**, —pp. 72–81.

DYNAMICAL BOREL-CANTELLI LEMMA FOR AUTOREGRESSIVE PROCESSES WITH LAPLACE NOISES

A.A. Dzhalilov[1], X.Sh. Abdusalomov[2]

[1]*Turin Polytechnical University in Tashkent*
[2]*National University of Uzbekistan*
Tashkent, UZBEKISTAN
e-mail: [1]adzhalilov21@gmail.com, [2]hasanboy155abs@gmail.com

Consider a deterministic dynamical system $\langle \mathbb{M}, \mathfrak{F}, \mu, T\rangle$, where μ is $T-$ invariant probability measure. The well-known dynamical Borel-Cantelli lemma states that for certain sequences of measurable subsets $A_n \subset \mathbb{M}$ and $\mu-$ almost every point x the inclusion $T_n x \in A_n$ holds for infinitely many values n. In the present paper, we study the stationary Markov process $\mathbb{X} := \{X_n,\ n \in \mathbb{N}\}$ defined as

$$X_n := X_n(\rho, \underline{\xi}) = \rho X_{n-1} + \xi_n,\ n \in \mathbb{Z},$$

where ρ is a real constant, $\underline{\xi} := \{\xi_n, n \in \mathbb{Z}\}$ is a sequence of independent, identically distributed (i.i.d.) random variables and $\xi_0 \sim Laplace(0, b)$. Let $\langle \mathbb{R}^{\mathbb{Z}},\ \mathscr{B},\ \nu\rangle$ be the probability space, where ν is a probability measure associated by stochastic process $\mathbb{X}$. Consider the shift map τ on $\mathbb{R}^{\mathbb{Z}}$. We give sufficient conditions on sequences of cylinders, that ensure the dynamical Borel-Cantelli lemma for the dynamical system $\langle \mathbb{R}^{\mathbb{Z}}, \mathscr{B}, \nu, \tau\rangle$. It also holds for AR(1) processes generated by the exponential, uniform, and Laplace distributions.

Keywords: Dynamical Borel-Cantelli lemma, autoregressive process, Gaussian distribution, Markov process

1 Introduction

The classical Borel-Cantelli lemma plays an important role in probability theory and its applications. The modern dynamical systems theory uses a special version of the Borel-Cantelli lemma (see for instance [1]. Let $\langle \mathbb{M}, \mathfrak{F}, \mu, T\rangle$ be a dynamical system with $T-$ invariant probability measure μ. The classical Poincare recurrence theorem states that for any fixed subset $A \in \mathfrak{F}$ with $\mu(A) > 0$ the equality $\mu(\{x \in A \mid\ T^n(x) \in A \text{ for infinitely many } n \in \mathbb{N}\}) = \mu(A)$ holds. We consider the sequence of nontrivial measurable subsets A_n and can still ask for the $\mu-$measure of the limsup set: $\{x \in \mathbb{M} \mid T^n(x) \in A_n \text{ for infinitely many } n \in \mathbb{N}\} = limsup T^{-n}A_n$. In the case $\Sigma\mu(T^{-n}A_n) = \Sigma\mu(A_n) < \infty$, the convergence case of the Borel-Cantelli lemma implies that the $\mu-$measure of the limsup set is zero. If $\Sigma\mu(A_n) = \infty$, and $T^{-n}A_n$ are independent, then for $\mu-$ measure of the limsup set is one. The last assertion has a limited value for deterministic dynamical systems, since one rarely deals with purely independent sets. In the case, if $\Sigma\mu(A_n) = \infty$ and the events $T^{-n}A_n$ are dependent, the situation is more complicated and more interesting. A definition, found in [2], applies to this situation. (see [2]). A sequence of measurable sets A_n $n \in \mathbb{N}$, such that

$$\Sigma\mu(T^{-n}A_n) = \Sigma\mu(A_n) = \infty, \tag{1}$$

is called **a Borel-Cantelli (BC) sequence for** T if

$$\mu(\limsup T^{-n}A_n) = \mu(\mathbb{M}).$$

The divergence case of the Borel-Cantelli lemma is not helpful for finding BC sequences since this case of the lemma requires independent sets. To obtain a BC sequence, we need to impose some restrictions. If, for a dynamical system, all sequences of subsets A_n that satisfy (1) and certain additional conditions are BC, we obtain what is called **a Dynamical Borel-Cantelli Lemma** (DBCL). The first example of such a lemma, in which only sequences of balls centered at a fixed point and with weakly monotonically decreasing radii are allowed, was proved by J. Kurzweil [3]. For a dynamical system with mixing property, the abundance of BC sequences can be interpreted as an aspect of strong chaos and stochastic behavior of the system. It is proved (see e.g. [2], [6], [7]) that for a wide class of hyperbolic or fast mixing systems, various sequences of sets have the BC property. The sets that are to be considered in this kind of problem are usually decreasing sequences of balls with the same center (see [3]) or cylinders. In [2] Chernov proved the dynamical BC lemma for Gibbs measures. In [6] Kim and Galatolo established that the Borel-Cantelli property and the waiting time problem are in general strictly connected. The main goal of this paper is to prove the dynamical BC for autoregressive AR(1) processes. Next, we introduce several important definitions. Let $\langle \mathbb{M}, \mathfrak{F}, \mu, T\rangle$ be a dynamical system with $T-$ invariant probability measure μ. Denote by $\chi_n(x)$ the indicator function of the set $B_n := T^{-n}A_n$. For every $N > 0$ we define the following two sums:

$$S_N(x) := \sum_{n=1}^{N} \chi_n(x), \;\; E_N := \sum_{n=1}^{N} \mu(A_n).$$

Definition 1. A sequence $\{A_n \subset \mathbb{M},\ n \in \mathbb{N}\}$ is said to be a **strongly Borel-Cantelli (sBC) sequence** if for μ-almost every $x \in \mathbb{M}$ we have

$$\lim_{N\to\infty} \frac{S_N(x)}{E_N} = 1.$$

We define the quantities R_{mn} which characterizes the dependence of two events B_m and B_n :

$$R_{mn} := \mu(B_m \cap B_n) - \mu(B_m)\mu(B_n) = \mu(T^{-m}A_m \cap T^{-n}A_n) - \mu(A_m)\mu(A_n).$$

A sufficient condition for a sequence $\{A_n\}$ to be an sBC sequence, in terms of R_{mn}, was first found by W. Schmidt, and the proof was later provided by Sprindzuk [8] in the context of Diophantine approximations. This condition was recently adapted to dynamical systems by D. Kleinbock and G. Margulis [4].

Assume that We suppose that the number R_{mn} satisfies the following condition:

$$\sum_{n=M}^{N}\sum_{m=M}^{N} |R_{mn}| \le CE_N,$$

for some constant $C > 0$ and for all $N > M > 1$. The last condition is called **(SP)-condition.**

Theorem 1 (see [8]). *If the sequence $\{A_n\}$ satisfies (SP), then it is an sBC sequence; moreover, for almost every $x \in \mathbb{X}$ one has*

$$\frac{S_N(x) - E_N}{\sqrt{E_N}} = \mathcal{O}\left(\sqrt{\log E_N}\right).$$

2 Main results

Let $L = [m, k]$ be a closed lattice interval, i.e. $[m, k] \subset \mathbb{Z}^1$.We call m and k the **left** and **right endpoints** oof aa finite latticeinterval$L \subset \mathbb{Z}$, respectively, and $(m + k)/2$ the **center** of L. We say that two lattice intervals $[m_1, k_1]$ and $[m_2, k_2]$ are **D-nested** for $D \geq 0$ if either $[m_1, k_1] \subset [m_2 - D, k_2 + D]$ or $[m_2, k_2] \subset [m_1 - D, k_1 + D]$. If the left endpoints of all lattice intervals $L_n,\ n \in \mathbb{N}$ lies in the interval $[0, D]$, then the intervals L_n are called **D-aligned**.

Let $\{X_n, n \in \mathbb{Z}\}$ be a sequence of real-valued random variables on the same probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. Let $\mathfrak{F}_n$ be the smallest $\sigma-$ algebra such that X_n is measurable. For $n \leq m$, we denote by $\mathfrak{F}_n^m$ the smallest $\sigma-$ algebra with respect to which $X_n, ..., X_m$ are jointly measurable.

The random process $\{X_n, n \in \mathbb{Z}\}$ is called **first order autoregressive (AR(1))** process with innovation random process $\{\xi_n, n \in \mathbb{Z}\}$ and autoregressive parameter ρ iff

$$X_n = \rho X_{n-1} + \xi_n,\ n \in \mathbb{Z}.$$

Assume that $|\rho| < 1$ and $\{\xi_n, n \in \mathbb{Z}\}$ is a sequence of independent identically distributed (i.i.d.) random variables. Let $\xi_0 \sim Laplace(0, b)$. The strong mixing property for AR(1) processes were studied by D. Andrews in [5]. The AR(1) process is strong stationary with invariant measure is $\pi(x) = \frac{\sqrt{1-\rho^2}}{2b} e^{-\frac{\sqrt{1-\rho^2}}{b}|x|}$, and the transition probability density is $p(x, y) = \frac{1}{2b} e^{-\frac{|y-\rho x|}{b}}$.

Let $\sum$ be a **symbolic space** defined as

$$\sum := (\underline{x} : \underline{x} = (...x_{i-1}, x_i, x_{i+1}, ...),\ x_i \in \mathbb{R},\ i \in \mathbb{Z}) =: \mathbb{R}^{\mathbb{Z}^1}.$$

Let $\tau : \sum \to \sum$ be **a shift map** defined as $\tau(x_i) = x_{i+1},\ i \in \mathbb{Z}$. For every $L := [m, k] \subset \mathbb{Z}$ we define the cylinder as

$$C[m, k] = \{\underline{x} :\ \underline{x} = (...x_{i-1}, x_i, x_{i+1}, ...),\ x_i \in [a_i, b_i],\ m \leq i \leq k\},$$

where $\mathscr{B}(\mathbb{R})$ is the Borel $\sigma-$ algebra of subsets $\mathbb{R}$.

Denote by $\mathscr{B}(\mathbb{R}^{\mathbb{Z}})$ the minimal $\sigma-$algebra containing all possible cylindric subsets. Denote by ν a Borel probability distribution on measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ produced by Laplas distribution $Laplace(0, b)$ i.e.

$$\nu(B) := \frac{1}{2b} \int_I e^{-\frac{|x|}{b}} dx,\ \text{for } \forall B \in \mathscr{B}(\mathbb{R}).$$

Consider the measurable space $(\mathbb{R}, \mathscr{B})$, where $(\mathscr{B})$ is the $\sigma-$ algebra of Borel subsets. We denote $\mathbb{Z} := \{\dots, -1, 0, 1, \dots\}$. Let $m, k \in \mathbb{Z}$, $m \le k$ and $G_i \in \mathscr{B}$, $m \le i \le k$. Define a cylinder as

$$C[m,k] := \{\underline{x} : \ \underline{x} = (...x_{i-1}, x_i, x_{i+1}, ...)\,,\ x_i \in G_i,\ m \le i \le k,\ \}\,,$$

Fix a number $\gamma \in (0, \frac{1}{2})$. We say that $C[m,k]$ is $\gamma-$**type cylinder** if it satisfies the following conditions:

- $G_i \subset [-A, A]$, $m \le i \le k$,
- $\max\{\sup_{-\infty<a<\infty} \nu(a + G_i),\ m \le i \le k\} \le 1 - \gamma$,

where $a + G$ denotes the set $\{a + x,\ x \in G\}$.

The main result is as follows.

Theorem 2. *Let $\gamma \in (0,1)$ and $D \ge 0$. Suppose that the cylinders $C[m^{(n)}, k^{(n)}]$,$n \ge 1$ defined on lattice intervals $[m^{(n)}, k^{(n)}] \subset \mathbb{Z}$ satisfies the following conditions:*

- *any two lattice intervals of the system $\{L_n := [m^{(n)}, k^{(n)}],\ n \ge 1\}$ are D-nested;*
- *each cylinder $C[m^{(n)}, k^{(n)}]$ is of $\gamma-$type. Then the sequence of cylinders $\{C[m^{(n)}, k^{(n)}],\ n \ge 1\}$ satisfies (SP), and hence, if in addition $\sum \mu(C_n) = \infty$, it is an sBC sequence, and (2) holds.*

References

1. Tseng J. (2008). On circle rotations and the shrinking target properties.*Discrete Contin. Dyn. Syst.*Vol. **20** ,pp 1111-1122.
2. Chernov N., Kleinbock D. (2001). Dynamical Borel-Cantelli lemmas for Gibbs measures.*Israel Journal Math.*Vol **122**,pp 1-27.
3. Kurzweil J. (1955). On the metric theory of inhomogeneous diophantine approximations.*Studia Math.*Vol **15**, pp 84-112.
4. Kleinbock D., Margulis G. (1999). Logarithm laws for flows on homogeneous spaces.*Inventiones Mathematicae.*Vol **138**.pp 451-494.
5. Andrews D. (1983). First Order Autoregressive Processes and Strong Mixing.*Cowles Foundation Discussion Papers.*Num **897**, pp 930-934.
6. Galatolo S., Kim D.H. (2008). The dynamical Borel-Cantelli lemma and the waiting time problems.*Indagationes Mathematicae* .Vol **18**.Num **3**,pp 421-434.
7. Athreya J., Margulis G. (2009). Logarithm laws for unipotent flows,I.*The Journal of Modern Dynamics.*Vol **3**,pp 359-378.
8. Sprindzuk V. (1979). Metric Theory of Diophantine Approximations. *Wiley,New York-Toronto-London.*

ON APPROXIMATE FORMULAS FOR MATHEMATICAL EXPECTATIONS OF NONLINEAR FUNCTIONALS OF RANDOM PROCESSES

A.D. Egorov[1]

[1]*Institute of Mathematics*
[1]*National Academy of Sciences of Belarus*
Minsk, BELARUS
e-mail: [1]egorov@im.bas-net.by

Algorithm for calculations of mathematical expectations of nonlinear functionals from the solution to an linear 2-dimensional equation of Skorohod with first-order chaos in coefficients and some linear Ito equations is described. An approach based on the using of multiple Stieltjes integrals for constructing of approximate formulas is used. Numerical examples illustrating the application of the obtained formulas are given.

Keywords: stochastic differential equations, Skorohod equation, mathematical expectations of functionals from solutions, approximate formulas

1 Introduction

In [1] approximations of mathematical expectations of nonlinear functionals from random processes of the form, which are complex functionals of the form $F(X_{(\cdot)}(Y))$, where $X_{(\cdot)} \equiv X$ is a solution of stochastic differential equation (SDE), are considered. Approximations are based on the construction of approximate formulas that are exact for functional polynomials of X. The possibility of constructing formulas with the specified accuracy was determined by the specific type of moments of process X. In work [2], in the case when X is a solution of the linear Ito equation, an approach to constructing an approximate formula that is exact for polynomial of the third degree of the solution was obtained, independent of the previously considered restrictions. This approach was then applied in [3, 4, 5, 6] in cases when X is a solution of the linear Ito equations with a leading Poisson process, the Ito-Levy equation and with the functional X admitting chaotic expansions in multiple integrals. In the paper [8] in the case when X is the solution of the linear Skorohod equatin, a new approach to constructing an approximate formula of the third degree of accuracy is proposed, based on the use of multiple Stieltjes integrals, which in many cases seems simpler. This report is devoted to application this approach for calculating functionals of the solution of linear Skorohod SDE with first-order chaos in the coefficients and some of those considered in [2, 3, 4, 5]. Numerical examples are considered that illustrate the application of the obtained formulas.

2 Main results

We use in this report the next formulae from [5], exact for functional polynomial $P(X_{(\cdot)}) = F_0 + \sum\limits_{k=1}^{3} \int\limits_{[0,1]^k} f_k(t_1,\dots,t_k)X(t_1,\dots,t_k)dt_1\cdots dt_k,\ F_0 = const,\ f_k(t_1,\dots,t_k)$ are real functions:

$$E[F(X_{(\cdot)})] \approx F(0)+\sum_{j=1}^{2} A_j \Lambda F(c_j M_1(\cdot))+0.5 \int\limits_{[0,1]^2} M_2(u_1,u_2)d^2_{u_1,u_2}\Delta F(1_{[0,\cdot]}(u_1)+1_{[0,\cdot]}(u_2))-$$

$$\frac{1}{6}\int\limits_{[0,1]^3} M_3(u_1,u_2,u_3)d^3_{u_1,u_2,u_3}\Lambda F\Big(\sum_{j=1}^{3} 1_{[0,\cdot]}(u_j)\Big) \equiv J(F(X_{(\cdot)})),$$

where the multiple Stylites integrals are in the right part of the equality, $\Delta F = 0.5(F(x)+F(-x))$, $\Lambda F(x) = 0.5(F(x)-F(-x))$; $M_k(u_1,\dots,u_k)$, $k=1,2,3$, are moments; A_j, c_j are constants satisfying $A_1c_1 + A_2c_2 = 1,\ A_1c_1^3 + A_2c_2^3 = 0$.

We consider the cases when the moments can be evaluated in a form permitting their calculation with sufficient exactness. So if X is a solution of the stochastic differential we confine oneself to cases when a solution of the stochastic equation can be found explicitly. In the report approximate evaluation of the mathematical expectation of nonlinear functionals from solutions of SDE Ito, driven by Wiener, Poisson, Ito-Levy processes, and Skorohod SDE are considered.

Let us consider a stochastic differential equation

$$X_t = X_0 + \int\limits_0^t (A^0_s + \int\limits_0^1 A^1_{s,r}dW_r)X_s\delta W_s, \tag{1}$$

where $X_0 = X_0(\omega)$ is a random variable with a finite Wiener chaos expansion, W_t is canonical Wiener process defined on probability space $\Omega = C_0([0,1])$, A^0_s and $A^1_{s,r}$ are commuting square integrable real matrix functions. The integral in right part of (1) is interpreted in the Skorohod sense. Using the results from [7] (received for more general case) a solution of (1) can be presented in the form

$$X_t = \exp\Big\{\int\limits_0^t \tau_{s,t}(A_s)dW_s - \frac{1}{2}\int\limits_0^t \tau_{s,t}(A_s)^2ds + \int\limits_0^t\int\limits_s^t \tau_{s,t}(D_rA_s)D_s[\tau_{s,t}(A_r)]drds\Big\}\tau_{0,t}(X_0),$$

where $\tau_{s,t}(A_s) = \gamma_{s,t}$ is a solution of equation $\gamma_{s,t} = A^0_s + \int\limits_0^1 A^1_{s,r}dW_r - \int\limits_s^t A^1_{s,r}\gamma_{r,t}dr$, D_s- the operator of functional derivative and for $I_n(f_n) = n!\int\limits_0^1\int\limits_0^{s_n}\cdots\int\limits_0^{s_2} f_n(s_1,\dots,s_n)dWs_1\cdots dWs_n$

$$\tau_{s,t}(I_n(f_n)) = \sum_{k=0}^{n}(-1)^kC^k_n \int\limits_{[s,t]^k} \gamma_{s_1,\tau}\cdots\gamma_{s_k,\tau}I_{n-k}(f_n(s_1,\dots,s_k,*))ds_1\dots ds_k,$$

$$D_t \sum_{n\geq 1} I_n(f_n) = \sum_{n\geq 1} n I_{n-1}(f_n(\cdot,t)),\ t\in[0,1].$$

We consider in our report a particle case of (1) under $d=2$, $A=\begin{pmatrix} a_0(s) & 0 \\ 0 & a_0(s)\end{pmatrix}$, $A=\begin{pmatrix} 0 & 0 \\ I_b & 0\end{pmatrix}$, $a_0(s)\in L^2[0,1]$, $I_b=\int\limits_0^1 b(\tau)dW_\tau$, $X_b=\begin{pmatrix} x_0^1 \\ x_0^2\end{pmatrix}$, $x_0^1\in R$, $x_0^2=x_0^2(\omega)=\int_0^1 g(r)dW_r$, $g_r\in L_2[0,1]$.

Theorem 1. [9] Solution to equation (1) have the form

$$X_t^1 = x_0^1 \exp\left\{\int_0^t a_0(s)dW_s - \frac{1}{2}\int_0^t a_0^2(s)ds\right\},$$

$$X_t^2 = x_0^1\left(\int_0^t\left(I_b - \int_s^t a_0(r)b(r)dr\right)dW_s - \int_0^t a_0(s)\left(I_b-\int_s^t a_0(r)b(r)dr\right)ds\right)\times$$
$$\exp\left\{\int_0^t a_0(s)dW_s - \frac{1}{2}\int_0^t a_0^2(s)ds\right\} + \left(\int_0^t g(r)dW_r - \int_0^t a_0(s)g(s)ds\right)\times$$
$$\exp\left\{\int_0^t a_0(s)dW_s - \frac{1}{2}\int_0^t a_0^2(s)ds\right\}.$$

Numerical results are considered under the conditions: $a_0(s)=s$, $b(s)=\lambda$, $x_0^1=1$, $g(s)=1$, $F(X_t^1,X_t^2)=\sin(\nu(X_t^1-X_t^2))$ $(\lambda=0.3,\nu=0.3)$. Approximating expression have the form: $J(F(X_t^1,X_t^2)) = -\frac{1}{\sqrt{2}}\sin(\sqrt{2}\nu)+2\sin(\nu)+\frac{1}{6}M_1(t,t,t)(\sin(3\nu)-3\sin(2\nu+3\sin(\nu)))$, where $M_3(t,t,t)=E[(X_t^1-X_t^2)^3]$. In Figure 1 we see comparison between the exact and the approximate values $E[F(X_t^1,X_t^2)]$.

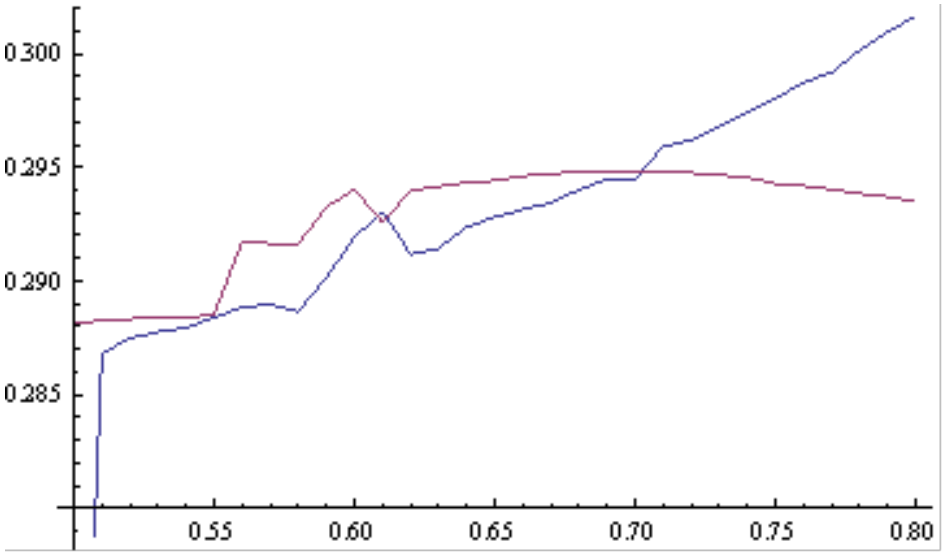


Figure 1: Comparison of the exact (blue) and the approximate (red) values of $E\{\sin(\nu(X_t^1-X_t^2))\}$

References

1. Egorov A.D., Sabelfeld K.K. (2010). Approximate formulas for expectation of functionals of solutions to stochastic differential equations. *Monte Carlo methods and applications*. Vol. **16**, No. 2, pp. 95–127.

2. Egorov A.D., Ulasik A.F. (2012). Approximate formulas of third accuracy degree for evaluation of mathematical expectation of functionals from solution to stochastic equation. *Vestsi of the National'nai akademii navuk Belarusi. Seryia fizika-matematichnych navuk = Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series*. Vol. **1**. pp. 8–12 (in Russian).

3. Egorov A.D. (2015). On approximate evaluation of mathematical expectation of functionals from solution to the Ito-Levy linear equation. *Doklady of the National Academy of Sciences of Belarus*. Vol. **59**, No. 1, pp. 13–17 (in Russian).

4. Egorov A.D. (2017). On apprpximate formulas for evaluation a class of functionals from the Poisson process. *Vestsi of the National'nai akademii navuk Belarusi. Seryia fizika-matematichnych navuk = Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series*. No. 1. pp. 7–13 (in Russian).

5. Ayryan E.A., Egorov A.D., Malyutin V.B., Sevastianov L.A. (2017). Approximate formulas for mathematical expectations of functionals of random processes defined by Ito-Levy multiple integral expansions. *Mathematical Modelling and Geometry*. Vol. **5**, No. 3, pp. 1–15.

6. Egorov A.D. (2020). An approximate formulas for calculating the expectations of functionals from random processes based on using the Wiener chaos expansion. *Monte Carlo Methods and Applications*. Vol. **26**, No. 4, pp. 285–292.

7. Buckdahn R., Nualart D. (1994) Linear stochastic differential equations and Wick products. *Probab. Th. Rel. Fields*. Vol. **99**, pp. 501–526.

8. Egorov A.D. (2021). Approximate formulas for the evaluation of the mathematical expectation of functionals from the solution to the linear Skorohod equation. *Vestsi of the National'nai akademii navuk Belarusi. Seryia fizika-matematichnych navuk = Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series*. Vol. **57**, no. 2, pp. 198-205.

9. Egorov A.D. (2023). On the calculation of functionals from the solution to linear Skorohod SDE with first-order chaos ih coefficients. *Vestsi of the National'nai akademii navuk Belarusi. Seryia fizika-matematichnych navuk = Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series*. Vol. **59**, no. 3, pp. 201-212.

A NEW STOCHASTIC ESTIMATOR FOR SYMMETRY CENTER IN MULTIDIMENSIONAL SPACE

A.A. Filatova[1], M.S. Ermakov[2]

[1,2]*St. Petersburg State University*
St. Petersburg
e-mail: [1]rii.filatova@gmail.com, [2]erm2512@mail.ru

We consider a new robust stochastic algorithm for estimating the symmetry center of multivariate distributions with convex centrally symmetric density level surfaces. We investigate its time complexity, convergence rate, and robustness. A modification of the algorithm that improves estimation accuracy is also discussed. Comparative numerical experiments with existing multivariate location parameter estimation algorithms, including Tukey, Oja, and geometric medians, are presented. Since in the considered class of distributions these algorithms estimate the symmetry center, the comparison is objective.

Keywords: robust statistics, multivariate median, stochastic approximation, symmetry center estimation

1 Introduction

It is known that there is no natural generalization of the univariate median for multivariate data [2]. This is because in multidimensional space, unlike $\mathbb{R}^1$, there is no natural ordering of points, making direct transfer of univariate concepts impossible. As a result, various approaches to defining multivariate medians have been proposed in literature, such as the geometric median [4], Oja median [5], concepts based on data depth, particularly Tukey median [1], and others [2].

Although there is no single universally accepted generalization of the univariate median to the multivariate case, there are certain properties it should possess. For example, for univariate symmetric distributions, the median coincides with the symmetry center. It is natural to assume that the same should hold in multidimensional space. In this case, as a generalization of the symmetric distribution concept, we consider distributions with convex centrally symmetric density level surfaces. Therefore, the proposed algorithm is based on the consideration that for such distributions it should estimate the symmetry center.

An important practical limitation of many existing methods is their high computational complexity [1, 5, 7], which significantly narrows their application scope. To overcome this limitation, the paper proposes a stochastic approach that potentially provides both robustness and reduced computational costs.

2 New robust algorithm for multivariate location parameter estimation

Algorithm 1 estimates the location parameter for distributions whose density level surfaces f are convex and centrally symmetric, i.e., for any $\mathbf{x} \in \mathbb{R}^d$:

- $f(\mathbf{x} + \mathbf{x}_0) = \text{const}$ is a convex set;
- $f(\mathbf{x}_0 + \mathbf{x}) = f(\mathbf{x}_0 - \mathbf{x})$, where $\mathbf{x}_0$ is the symmetry center.

In this case, the location parameter is defined as the value coinciding with the distribution's symmetry center $\mathbf{x}_0 \in \mathbb{R}^d$.

The assumptions imply that Algorithm 1 can be considered as a method for estimating the symmetry center. The quantity obtained as a result of the algorithmf will henceforth be called the **stochastic median**.

Algorithm 1 reduces the multivariate problem to a sequence of univariate subproblems. This approach allows sequential approximation to the symmetry center while using simple computations and providing an efficient solution to the location parameter estimation problem in the multivariate case.

Algorithm 1 for multivariate median estimation

1. Consider a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a distribution with density function f that satisfies assumptions of the algorithm.
2. Arbitrarily select a point $\widehat{\mathbf{m}}_1$ as the initial approximation. Set the computation accuracy ε.
3. Generate a random vector $\mathbf{u}_i$ uniformly distributed on the unit sphere. Construct the line
$$l_i = \{\widehat{\mathbf{m}}_i + \lambda \mathbf{u}_i,\ \lambda \in \mathbb{R}\}.$$
4. Project observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ onto line l_i, obtaining projection points $y_{i1}, \ldots, y_{in}$.
5. Find the median $\widehat{\mathbf{m}}_{i+1}$ of points $y_{i1}, \ldots, y_{in}$.
6. The algorithm terminates when the condition $\|\widehat{\mathbf{m}}_i - \widehat{\mathbf{m}}_{i+1}\| < \varepsilon$ is met. Otherwise, increment the step counter i by 1 and return to step 3.

To reduce the influence of randomness on the final estimate, consider the following **modification** of Algorithm 1. Instead of using the result of the last iteration, perform k additional iterations and take the average over these k iterations as the final estimate after reaching the specified accuracy ε. Let the algorithm execute N steps before reaching the specified accuracy, then

$$\widehat{\mathbf{m}}_{\text{avg}_k} = \frac{1}{k} \sum_{i=N+1}^{N+k} \widehat{\mathbf{m}}_i.$$

Averaging the last iterations reduces the estimate's variance and makes it more stable.

3 Time complexity analysis

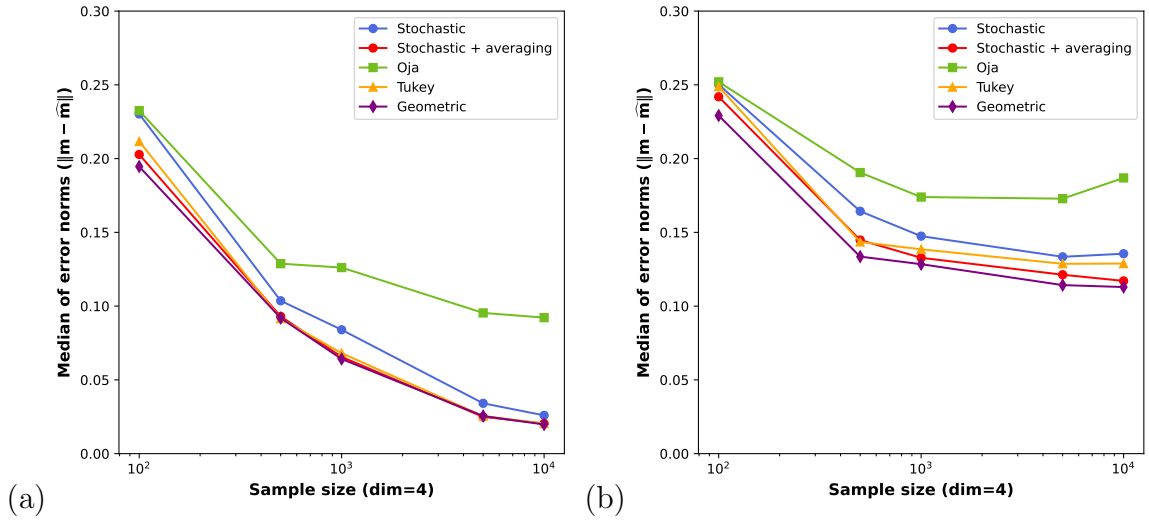
The complexity of k steps of Algorithm 1 for a sample of size n in a space of dimension d is $O(knd)$.

The time complexity of algorithms for estimating various median generalizations in a space of dimension d varies significantly across different methods. The Tukey median approximation (ABCDepth) has a time complexity of $O((d+k)n^2 + n^2 \log n)$ [3], while the exact computation for large n and d is NP-hard [7]. However, randomized optimization techniques can achieve $O(n^{d-1})$ complexity for exact Tukey median computation [6]. The Oja median has a higher complexity of $O(kdn^d \log n)$ [5], whereas the geometric and stochastic median estimation algorithms are more efficient, both with a complexity of $O(knd)$ [4]. This makes the geometric and stochastic methods preferable for large samples ($n \gg 1$) and high-dimensional spaces ($d \gg 1$).

4 Comparison with other algorithms

Let us compare the proposed algorithm with popular algorithms for multivariate location parameter estimation.

Consider median norms of error for different sample sizes in the case of 4-dimensional standard normal distribution. Simulation results are shown in Figure 1.



(a) (b)

Figure 1: Median norms of error for samples of size 100, 500, 1000, 5000 and 10000 from normal distribution (a) without and (b) with outliers. 100 simulations were made.

From Figure 1 (a) we can assume that Algorithm 1 is comparable in accuracy to the geometric median, which is its complexity competitor, as well as to the Tukey median for given sample sizes.

Now replace 5% of the sample with an outlier point $(8,\ 6,\ 5,\ 10)^{\mathbf{T}}$ and run the algorithms again. Results are shown in Figure 1 (b). We see that the proposed Algorithm 1 now estimates the location parameter slightly worse, but the accuracy is still close to the geometric median and Tukey median.

Thus, comparison results show that the new Algorithm 1 may compete with known algorithms for multivariate location parameter estimation.

5 Conclusion

A new algorithm for robust estimation of center of symmetry in multidimensional space was implemented. A comparison with common algorithms showed that the proposed algorithm and its modification provide comparable results and are not inferior in accuracy. The algorithm is efficient for large samples and high dimensions, making it a promising alternative to existing methods.

References

1. Rousseeuw P., Struyf A. Computation of robust statistics: depth, median, and related measures // Handbook of Discrete and Computational Geometry. Second ed. – Chapman & Hall/CRC. – 2004. – P. 1279-1292.

2. Small C. G. A Survey of Multidimensional Medians // International Statistical Review. – 1990. – Vol. 58, No. 3. - P. 263-277.

3. Bogievi M., Merkle M. Approximate Calculation of Tukey's Depth and Median With High-dimensional Data // Yugoslav Journal of Operations Research. – 2018. – Vol. 28, No. 4. – P. 475-499.

4. Cohen M. B., Lee Y. T., Miller G., Pachocki J., Sidford A. Geometric median in nearly linear time // Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing. – 2016. – P. 9-21.

5. Ronkainen T., Oja H., Orponen P. Computation of the multivariate Oja median // Developments in robust statistics: International Conference on Robust Statistics ICORS '01, Stift Vorau, Itvalta, heinkuu 2001. – 2002. – P. 344-359.

6. Chan T.M. An optimal randomized algorithm for maximum Tukey depth // Proc. 15th Annu. ACM-SIAM Sympos. Discrete Algorithms (SODA '04). – 2004. – P. 430-436.

7. Dyckerhoff R., Mozharovskyi P. Exact computation of the halfspace depth // Computational Statistics & Data Analysis. – 2016. – Vol. 98. – P. 19-30.

OPINION DYNAMICS CONTROL IN A SOCIAL NETWORK

A.A. Ivashko[1], V.V. Mazalov[2]

[1,2]*Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences*
Petrozavodsk, RUSSIA
[2]*Saint Petersburg State University*
Saint Petersburg, RUSSIA
e-mail: [1]aivashko@krc.karelia.ru, [2]vmazalov@krc.karelia.ru

We consider a model of opinion dynamics in a social network, in which there are N identical agents and one leader (principal). All agents have the same reputation, but the leader's reputation is much higher. The agents' reputations are determined by the matrix of agents' trust in each other. The optimal control of opinion dynamics in the social network is found to influence the final opinion of the agents. Numerical simulations for different influence matrices are carried out.

Keywords: opinion dynamics, control, social network

1 Introduction

A variety of models exist for opinion dynamics: the process through which opinions change and spread through a network. They vary in complexity, underlying assumptions, and structure of opinions generated. This paper considers a model based on the De Groot model of opinion dynamics [2]. This model is used in negotiation modeling.

Under the DeGroot opinion dynamics model, agents update their opinions at each time step to be a weighted average of their own current opinion and the opinions of everyone with whom they interact. This process is described for a network of N agents by the following equation

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t), t = 0, 1, ... \quad \mathbf{x}(0) = x_0, \tag{1}$$

where

$\mathbf{x}(t) = (x_1(t), ..., x_N(t))^T$,

$x_i(t)$ represents the opinion of agent i at time t,

$\mathbf{A} = (a_{ij})$ is a stochastic matrix of trust of agents to each other, $i, j = 1, 2, ..., N$,

a_{ij} represents the weight that agent i trusts agent j,

$x_i(0)$ is agent i's initial opinion.

Iterating (1) t times, we obtain that the agents' opinions at time t are calculated by the formula

$$\mathbf{x}(t) = \mathbf{A}^t\mathbf{x}(0), t = 0, 1, ... \tag{2}$$

Since the matrix $\mathbf{A}$ is stochastic, then if the ergodicity condition is satisfied, we obtain

$$\lim_{t\to\infty} \mathbf{A}^t \mathbf{x}(0) = \mathbf{x}(\infty) = (x, x, ..., x).$$

The vector $\mathbf{x}(\infty) = (x, x, ..., x)$ gives consensus in negotiations.

An review of opinion dynamics models is presented in [3]. The conditions for the existence of consensus for a model with two centers of influence were studied in [1]. Models with opinion dynamics controlled by an external player were considered in [4, 5, 6, 7].

In this paper, we consider opinion dynamics model in a social network with N identical agents and one leader (principal). The agents and the leader differ in the degree of trust in each other. It is assumed that the agents trust the leader more, and the leader distributes his/her trust among the agents. We study a model of the opinion dynamics optimal control in the social network in order to approximate the consensus closer to a given target value. Numerical modeling is carried out for different values of players' degrees of trust in each other, as well as initial values of agents' and leader's opinions.

2 Social Network with Principal

Consider a social network in which there are $N + 1$ agents, among which there is one principal and N identical agents.

Suppose that N agents have the same degree of trust in the leader, and the leader distributes his/her trust among all agents. Then the dynamic equation of the agents' state has the following form:

$$\begin{aligned} x_1(t+1) &= (1-p)x_1(t) + \frac{p}{N}\sum_{j=1}^{N} x_j(t), \\ x_i(t+1) &= (1-m)x_1(t) + mx_i(t),\ i = 2, ..., N+1. \end{aligned} \tag{3}$$

Therefore, the trust matrix has the form:

$$\mathbf{A} = \begin{pmatrix} 1-p & \frac{p}{N} & \frac{p}{N} & \dots & \frac{p}{N} \\ 1-m & m & 0 & \dots & 0 \\ & & & \dots & \\ 1-m & 0 & 0 & \dots & m \end{pmatrix}$$

Using formula (2), system (3) will take the form

$$
\begin{aligned}
x_1(t) &= \frac{1}{p-m+1}\Big[(1-m)x_1^0 + p\overline{x_0} + p(m-p)^t(x_1^0 - \overline{x_0})\Big],\\
x_i(t) &= \frac{1}{p-m+1}\Big[(1-m)x_1^0 + p\overline{x_0} - (1-m)(m-p)^t(x_1^0 - \overline{x_0}) + \\
&\quad + m^t(p-m+1)(x_i^0 - \overline{x_0})\Big],\ i = 2, ..., N+1,
\end{aligned} \tag{4}
$$

where $\overline{x_0} = \frac{1}{N}\sum\limits_{j=2}^{N+1} x_j^0$.

Then $\mathbf{x}(t)$ comes to the next steady state:

$$
\mathbf{x}(\infty) = \frac{1}{p-m+1}\begin{pmatrix} (1-m)x_1^0 + p\overline{x_0} \\ ... \\ (1-m)x_1^0 + p\overline{x_0} \end{pmatrix}. \tag{5}
$$

3 Opinion Dynamics Control

Suppose there is an external player who is interested in bringing the opinions of a social network to a certain value. The player can influence only the principal. Then the dynamic equation of the agents' state will take the following form:

$$
\begin{aligned}
x_1(t+1) &= (1-p+u)x_1(t) + \frac{p-u}{N}\sum_{j=2}^{N} x_j(t),\\
x_i(t+1) &= (1-m)x_1(t) + mx_i(t),\ i = 2, ..., N+1,
\end{aligned} \tag{6}
$$

where $u = \{u(t) : u(t) \in [-p, 1-p]\}$ is the player's control over the principal.

Let us consider the player's objective function:

$$
J(u) = \sum_{t=0}^{\infty}\delta^t\left[\sum_{i=1}^{N}(x_i(t)-s)^2 + \gamma u^2(t)\sum_{i=1}^{N}\sum_{j=i+1}^{N+1}\Big(x_i(t) - x_j(t)\Big)^2\right]. \tag{7}
$$

Here t denotes time, $0 < \delta \leq 1$ is a discount factor, $\gamma > 0$ represent the player's costs, and s is fixed value that the player expect all agents' opinion to reach.

The player aims to minimize his/her objective function with respect to $u(t)$.

Then from (6), taking into account (2), we obtain the dynamic equation

$$
\begin{aligned}
x_1(t) &= G\Big[M + (p-u)(m-p+u)^t K_1\Big],\\
x_i(t) &= G\Big[M - (1-m)(m-p+u)^t K_1\Big] + m^t K_i,\\
&\quad i = 2, ..., N+1.
\end{aligned} \tag{8}
$$

and steady state

$$\mathbf{x}(\infty) = \frac{1}{p-m-u+1} \begin{pmatrix} (1-m)x_1^0 + (p-u)\overline{x_0} \\ \dots \\ (1-m)x_1^0 + (p-u)\overline{x_0} \end{pmatrix}. \quad (9)$$

Here
$G = \dfrac{1}{p-m-u+1}$,
$M = (1-m)x_1^0 + (p-u)\overline{x_0}$,
$K_i = x_i^0 - \overline{x_0}$.
Note that

$$x_i(t) - x_j(t) = m^t(x_i^0 - x_j^0).$$

Assuming that the control $u(t)$ is constant, we rewrite function (7) in the following form

$$J(u) = \frac{1}{1-\delta}\sum_{i=1}^{N+1}(x_i(t)-s)^2 + \gamma\frac{u^2}{1-\delta m^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N+1}(x_i^0 - x_j^0)^2.$$

Numerical simulation results show that by controlling the principal, the player can shift the steady state of the system towards his/her target value.

References

1. Bure V. M., Parilina E. M., Sedakov A. A. (2017). Consensus in a social network with two principals. *Automation and Remote Control*, Vol. **78**, Num. **8**, pp. 14891499

2. De Groot M.H. (1974). Reaching a Consensus. *Journal of the American Statistical Association*. Vol. **69**, Num. **345**, pp. 118121.

3. Hassani H., Razavi-Far R., Saif M. (2022). Classical dynamic consensus and opinion dynamics models: A survey of recent trends and methodologies. *Information Fusion*. Vol. **88**, pp. 2240.

4. Hegselmann R., König S., Kurz S. (2014). Optimal opinion control: The campaign problem. *arXiv preprint arXiv:1410.8419*.

5. Jiang H., Mazalov V.V., Gao H. (2023). Opinion dynamics control in a social network with a communication structure. *Dynamic Games and Applications*. Vol. **13**, Num. **1**, pp. 412434.

6. Mazalov V., Parilina E. (2020). The Euler-Equation Approach in Average- Oriented Opinion Dynamics. *Mathematics*. Vol. **8**. Num. **3**, 355.

7. Wang C., Mazalov V.V., Gao H. (2021). Opinion Dynamics Control and Consensus in a Social Network. *Automation and Remote Control*. Vol. **82**, pp. 11071117.

WEAK CONVERGENCE OF HITTING TIMES FOR CRITICAL CIRCLE MAPS

A.A. Jalilov [1]

[1] *Amity University in Tashkent*
Tashkent, UZBEKISTAN
e-mail: [1]ajalilov@amity.uz

Let $Cr(\overline{\rho})$ be the set of all critical circle which maps are C^1 conjugate to $f_{cr} \in C^3$ critical circle homeomorphisms having a single $x_{\rm cr}$ critical point and rotation number $\overline{\rho} := [k, k, k, , ...]$. Let $\mu := \mu_f$ denote the unique probability invariant measure of the map $f \in Cr(\overline{\rho})$. Define a decreasing sequence $\{c_n := c_n(\theta),\ n \geq 1\}$ for some $\theta \in (0, 1)$ is such that a $\mu-$measure of the interval $(x_{cr}, c_n]$ satisfies $\mu([x_{cr}, c_n]) = \theta \cdot \mu([x_{cr}, f^{q_n}(x_{cr})])$, where q_n is the return times associated with the linear rotation $f_{\overline{\rho}} = x + \overline{\rho} \bmod 1$. We study weak convergence of normalized hitting times. Moreover, we show that limiting distribution is singular with respect to the Lebesgue measure.

Keywords: circle homeomorphism, critical point, invariant measure, rotation number, symbolic dynamics

1 Introduction

This paper aims to investigate week convergence of normilazed hitting times to shringking target intervals for critical circle maps with a single critical point. The classical Denjoy's theorem states that for ergodic circle diffeomorphisms f from class $C^2(S^1)$ is topologically conjugated to a linear rotation f_ρ (see for instance [3]).

Yoccoz in [1] extended Denjoy's classical result for the class circle maps with one or more critical points at which the derivative vanishes. Graczyk and Swiatek [2] proved that for a C^3 circle homeomorphism f with finitely many critical points of polynomial type and an irrational rotation number, the conjugacy φ is a singular function on S^1, meaning $\varphi'(x) = 0$ almost everywhere. As a result, the unique invariant probability measure μ_f is singular with respect to Lebesgue measure on the circle.

We assume that the number $\overline{\rho}$ has a continued fraction expansion

$$\overline{\rho} = [k, k, \ \ , k, ...] = \frac{1}{k + \frac{1}{k+...}},\ k \in \mathbb{N}.$$

The last relation shows that $\overline{\rho}$ is an irrational number of algebraic type. It is well known that the renormalization transformation $\mathbf{R}_{\overline{\rho}}$ has a unique fixed point f_{cr} in the space of all analytic critical maps with one cubic critical point x_{cr} with rotation number $\overline{\rho}$.

Let $Cr(\overline{\rho})$ denote the class of circle homeomorphisms on the standard circle $S^1 = \mathbb{R}/\mathbb{Z} \simeq [0, 1)$ that are C^1-conjugate to a fixed critical map f_{cr}. It is well known (see [3]) that topologically conjugate homeomorphisms share the same rotation number. Thus, every map in $Cr(\overline{\rho})$ has the rotation number $\overline{\rho}$.

Now, let f be an orientation-preserving homeomorphism of the circle $S^1 = \mathbb{R}/\mathbb{Z} \simeq [0,1)$ with an irrational rotation number $\rho = \rho_f$, and let $\mu = \mu_f$ be its unique invariant probability measure. Fix a point $z \in S^1$, and define the interval $\mathfrak{J}_\varepsilon(z) = [z, z+\varepsilon] \subset S^1$. The **first hitting time** of a point $x \in [0,1)$ to the interval $\mathfrak{J}_\varepsilon(z)$ is then defined as

$$N_\varepsilon^{(1)}(x) = inf\{i \geq 1 : f^i(x) \in \mathfrak{J}_\varepsilon(z)\}.$$

The goal is to identify conditions under which the rescaled hitting time converges in distribution as the interval $\mathfrak{J}_\varepsilon(z)$ shrinks. Since the expected hitting time is roughly $1/\mu(\mathfrak{J}_\varepsilon(z))$, it makes sense to rescale by this factor:

$$E_\varepsilon^{(1)}(x) = \mu(\mathfrak{J}_\varepsilon(z)) \cdot N_\varepsilon^{(1)}(x).$$

We study whether the distribution function

$$F_\varepsilon(t) = \mu\left(x \in S^1 : E_\varepsilon^{(1)}(x) \leq t\right)$$

converges as $\varepsilon \to 0$ at continuity points of the limit.

Ayupov and Jalilov in [7] proved the convergence of this distribution for $\rho = [1,1,1,....]$. Coelho and de Faria ([4], [5]) explored this problem for linear irrational rotations $f_\rho(x) = x + \rho \mod 1$, where Lebesgue measure ℓ is invariant. They showed that for almost every irrational ρ, the rescaled hitting times in renormalization intervals $[x_0, c_n]$ do not converge in law as $c_n \to 0$, but all possible limiting distributions along subsequences $\{c_n\}$ can be characterized.

Fix $\theta \in (0,1)$. Let q_n be the denominator of the n-th convergent of an irrational number $\overline{\rho}$. For each n, define the point $c_n(\theta)$ so that

$$\mu([x_0, c_n(\theta)]) = \theta \cdot \mu([x_0, f^{q_n}(x_0)]).$$

Define the hitting time $N_{n,\theta}^{(1)}$ to the interval $[x_0, c_n(\theta))$, and the corresponding rescaled hitting time

$$E_{n,\theta}^{(1)}(x) = \mu([x_0, c_n(\theta))) \cdot N_{n,\theta}^{(1)}(x).$$

Let $F_{\theta,n}(t)$ and $\Phi_{\theta,n}(t)$ denote the distribution functions of $E_{n,\theta}^{(1)}$ with respect to the invariant measure μ and Lebesgue measure ℓ, respectively.

Coelho ([5]) showed that for any converging subsequence $F_{\theta,n_m}(t)$, the limiting distribution $F_\theta(t)$ is piecewise linear on $[0,1]$, with $F_\theta(t) = 0$ for $t \leq 0$ and $F_\theta(t) = 1$ for $t \geq 1$. This result applies to all circle diffeomorphisms C^1-conjugate to irrational rotations, with μ replaced by Lebesque measure ℓ.

Similar limit theorems for hitting times (often exponential with parameter 1) have been found in other systems, such as Markov chains, Anosov and Axiom A diffeomorphisms, and piecewise expanding interval maps.

2 Main results

Let f be a circle homeomorphism with the rotation number $\overline{\rho} = [k, k, k, ...]$ and with a unique probability invariant measure $\mu := \mu_f$. Take an arbitrary point $x_0 \in S^1$.

Fix an arbitrary point $c_1 \in [f^{q_1}(x_0), 1)$. There exists a constant $\theta \in (0, 1]$ such that

$$\mu([x_0, c_1]) := \theta \cdot \mu([f^{q_1}(x_0), 1]) := \theta \cdot \mu(I_0^{(1)}). \tag{1}$$

For every $n \geq 1$ we define the numbers $c_{n:} = c_n(\theta) \in I_0^{(n)}(x_0)$ such that

$$\mu(I_{c_n}(x_0)) := \theta \cdot \mu(I_0^{(n)}(x_0)), \tag{2}$$

where the interval $I_{c_n}(x_0)$ has endpoints x_0 and c_n. Moreover,

$$c_n \in [f^{-q_{n+1}}(x_0),\ f^{q_n}(x_0)],\ \ n \geq 1.$$

The constraction shows that we get the embedded sequence of intervals i.e.

$$I_{c_1}(x_0)) \supset I_{c_2}(x_0)) \supset ...I_{c_n}(x_0)) \supset I_{c_{n+1}}(x_0))....$$

It is important that

$$\lim_{n\to\infty} \mu(I_{c_n}(x_0)) = 0.$$

Consider the first return time function $R_{c_n} : I_{c_n} \to \mathbb{N}$ as

$$R_{c_n}(x) := \min\{j \geq 1 : f^j(x) \in I_{c_n}\}.$$

The first return times function $R_{c_n}(x)$ takes only 3 values.

Proppsition 1. Let f be a circle homeomorphism with the rotation number $\overline{\rho} = [k, k, k, ...]$ and $x_0 \in S^1$. Assume the constant $\theta \in (0, 1)$ and $c_n \in I_0^{(n)}(x_0)$, $n > 1$ determined by (1) and (2), respectively.
(I) If $n \in \mathbb{N}$ is odd, then

$$R_{c_n}(x) = \begin{cases} q_{n+2}\ ,\ x \in [c_n, f^{q_{n+2}}(x_0)), \\ q_{n+3}\ ,\ x \in [f^{q_{n+2}}(x_0), f^{q_{n+1}}(c_n)), \\ q_{n+1}\ ,\ x \in [f^{-q_{n+1}}(c_n), x_0). \end{cases}$$

(II) If $n \in \mathbb{N}$ is even, then

$$R_{c_n}(x) = \begin{cases} q_{n+2}\ ,\ x \in [x_0, f^{-q_{n+2}}(c_n)), \\ q_{n+3}\ ,\ x \in [f^{-q_{n+2}}(c_n), f^{q_{n+1}}(x_0)), \\ q_{n+1}\ ,\ x \in [f^{q_{n+1}}(x_0), c_n). \end{cases}$$

To be definite, we consider the case when n is even. The case of odd n can be considered similarly. Introduce the following notations:

$$A_0^{(n)} := [x_0, f^{-q_{n+2}}(c_n)),\ \ n \geq 1,$$

$$C_0^{(n)} = [f^{-q_{n+2}}(c_n), f^{q_{n+1}}(x_0)), \;\; n \geq 1,$$
$$B_0^{(n)} = [f^{q_{n+1}}(x_0), c_n), \;\; n \geq 1.$$

The collection of intervals

$$\xi_n(x_0, c_n) := \{A_0^{(n)}, \; f(A_0^{(n)}), ..., f^{q_{n+2}}(A_0^{(n)})\}$$
$$\cup\{C_0^{(n)}, \; f(C_0^{(n)}), ..., f^{q_{n+3}}(C_0^{(n)})\}\cup$$
$$\{B_0^{(n)}, \; f(B_0^{(n)}), ..., f^{q_{n+1}}(B_0^{(n)})\}.$$

constitute the partition of the circle S^1. We denote it by $\xi_n(x_0, c_n)$ and call the $n-$**th generalized dynamical partition associated by the points** x_0 **and** c_n.

In this paper we study rescaled hitting times for critical circle maps $f \in Cr(\overline{\rho})$. The hitting time function $N_n^{(1)}(x)$ is normalized by its maximum value q_{n+3}, yielding the rescaled form

$$E_n^{(1)}(x) = \frac{1}{q_{n+3}} N_n^{(1)}(x).$$

We denote the distribution function of $E_n^{(1)}(x)$ with respect to Lebesgue measure on S^1 by $\Phi_{n,\theta}(t)$. Now, we formulate the main results of present work.

Theorem 2. The distribution function of the rescaled hitting time function $E_n^{(1)}(x)$ has the following form:

i) if $t < 1/q_{n+3}$, then $\Phi_{n,\theta}(t) = 0$,

ii) if $m/q_{n+3} \leq t \leq (m+1)/q_{n+3}$, $1 \leq m \leq q_{n+1}$, then

$$\Phi_{n,\theta}(t) = \sum_{i=q_{n+1}-m}^{q_{n+1}-1} |B_i^{(n)}| + \sum_{j=q_{n+2}-m}^{q_{n+2}-1} |A_j^{(n)}| + \sum_{k=q_{n+3}-m}^{q_{n+3}-1} |C_k^{(n)}|,$$

iii) if $m/q_{n+3} \leq t \leq (m+1)/q_{n+3}$, $q_{n+1} \leq m \leq q_{n+2}$, then

$$\Phi_{n,\theta}(t) = \sum_{i=0}^{q_{n+1}-1} |B_i^{(n)}| + \sum_{j=q_{n+2}-m}^{q_{n+2}-1} |A_j^{(n)}| + \sum_{k=q_{n+3}-m}^{q_{n+3}-1} |C_k^{(n)}|,$$

iv) if $m/q_{n+3} \leq t \leq (m+1)/q_{n+3}$, $q_{n+2} \leq m \leq q_{n+3}$ then

$$\Phi_{n,\theta}(t) = \sum_{i=0}^{q_{n+1}-1} |B_i^{(n)}| + \sum_{j=0}^{q_{n+2}-1} |A_j^{(n)}| + \sum_{k=q_{n+3}-m}^{q_{n+3}-1} |C_k^{(n)}|,$$

v) if $t \geq 1$, then $\Phi_{n,\theta}(t) = 1$, where $|L_i^{(n)}|$ is lebesque measure of L_i^n.

The above theorem shows that the normalized hitting times are discrete random variables and its distribution function is step function. Moreover, the values of

distribution function defined by invariant measures of intervals of dynamical partition.

Theorem 2. Let $\overline{\rho} = [k, k, k, ...]$ and let $f \in Cr(\overline{\rho})$ be critical circle map. Consider for $\theta \in (0, 1)$ the sequence of distribution functions $\{\Phi_{n,\theta}(t)\}_{n=1}^{\infty}$ with respect to Lebesgue measure on circle corresponding to the first rescaled hitting times $E_{n,\theta}^{(1)}(x)$ to interval $[x_c, c_n(\theta)]$. Then

1) for all $t \in \mathbb{R}^1$ there exists the finite limit

$$\lim_{n\to\infty} \Phi_{n,\theta}(t) = \Phi_{\theta}(t),$$

where $\Phi_{\theta}(t) = 0$, if $t \leq 0$, and $\Phi_{\theta}(t) = 1$, if $t > 1$;

2) the limit function $\Phi_{\theta}(t)$ is a strictly increasing on $[0, 1]$ and continuous distribution function on $\mathbb{R}^1$;

3) $\Phi_{\theta}(t)$ is singular on $[0, 1]$ i.e. $\frac{\Phi_{\theta}(t)}{dt} = 0$ a.e. with respect to Lebesgue measure ℓ on the circle.

The last theorem shows that the sequence of discrete rescaled hitting times weekly converges to a continues distribution function. Notice that the limit distribution function is singular on interval $[0, 1]$, i.e. it is continues , strictly increasing and its derivative is zero almost everywhere with respect to Lebesque measure ℓ on $[0, 1]$.

References

1. Yoccoz J. C. (1984) Il ny a pas de contre-exemple de Denjoy analytique. J.C.R. Acad.Sci.Paris. 298(7).pp. 141-144.

2. Graczyk J., Swiatek G. (1993) Singular measures in circle dynamics. Commun. Math. Phys. pp. 213-230. https://doi.org/10.1007/BF02099758

3. De Melo W., van Strien S. (1993). One dimensional dynamics-Berlin, New York.: Springer, pp. 3-25.

4. Coelho Z., de Faria E.. (1996) Limit laws of entrance times for homeomorphisms, Israel J. Math. 93 , pp. 93-112. https://doi.org/10.1215/00127094-2018-0017

5. Coelho Z. (2004) The Loss of Tightness of Time Distributions for Homeomorphisms of the Circle, Transactions of the American Mathematical Society, Vol. 356, No. 11, pp. 4427–4445, https://doi.org/10.1090/S0002-9947-04-03386-0

6. Kim D.H., Seo B.K. (2003) The waiting time for irrational rotations, Nonlinearity 16. pp. 1861-1868. https://doi.org/10.1088/0951-7715/16/5/318

7. Ayupov Sh.A., Jalilov A.A. (2021) Asymptotic distribution of hitting times for critical maps of the circle. Vestnik Udmurskogo Universiteta, V3, pp. 365-383.

A METHOD TO EVALUATE THE ECONOMIC GROWTH QUALITY OF HIGH-TECH INDUSTRY IN CHINA AND ITS REGIONAL LINKAGE

G. Jiacheng[1], B.A. Zhalezka[2]

[1,2]*Belarusian National Technical University*
Minsk, BELARUS
e-mail: [1]13951841399@163.com, [2]boriszh@yandex.ru

Based on China's provincial panel data, this study uses entropy weight model and vector autoregressive model to measure the economic growth quality of high-tech industries in China and test the interaction effect among regions. It is found that the quality of economic growth of high-tech industries in China has shown a steady growth trend during the study period, but there are obvious differences in the growth level among regions. The economic growth quality of high-tech industries in the eastern region is obviously stronger than that in the central region and the western region. In addition, the economic growth quality of high-tech industries in the eastern region is likely to have a siphon effect on the central region and the western region, while the central region is likely to have a radiation effect on the western region.

Keywords: high-tech industry, economic growth, growth quality, regional linkage, entropy weighting, vector autoregressive

1 Introduction

The quality of economic growth in high-tech industries and regional linkages determine the sustainability of a country's high-quality national economic development [1]. Therefore, this study takes China as an example to design a method to effectively evaluate the quality of economic growth in high-tech industries and regional linkage.

2 Methods and Methodology

Entropy weight model. In order to ensure that the evaluation results are more comprehensive and objective, referring to the existing research [2], this study adopts the entropy weight model to integrate various single indicators, and evaluates the economic growth quality of China high-tech industry by linear weighted summation. The specific model is established as follows.

$$Z_{ij} = \frac{X_{i,j} - \min(X_j)}{\max(X_j) - \min(X_j)} \times 0.999 + 0.001,$$

$$H_j = -\frac{1}{\ln(n)} \sum_{i=1}^{n} p_{i,j} \ln p_{i,j},$$

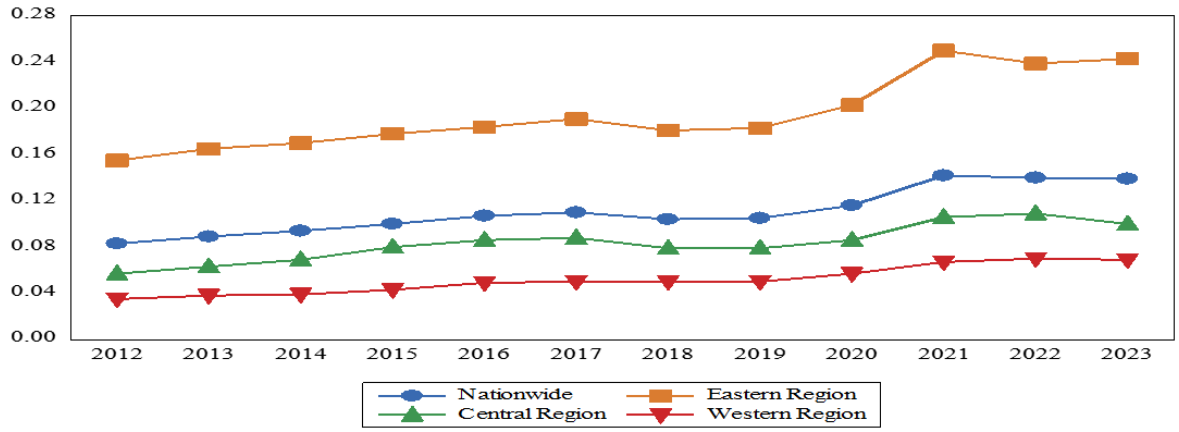


Figure 1: Measurement results and Changing trends in the quality of economic growth in China's high-tech industry

$$W_j = \frac{1 - H_j}{\sum_{j=1}^{m}(1 - H_j)} = \frac{1 - H_j}{m - \sum_{j=1}^{m} H_j},$$
$$S_i = \sum_{j=1}^{m} W_j Z_{i,j}.$$

In the above formula, S_i represents the economic growth quality of the high-tech industry of the i-th target to be evaluated. The larger the value of S_i, the higher the quality. $Z_{i,j}$ and $X_{i,j}$ respectively represent the j-th standardized and non-standardized original indicator of the target i to be evaluated. $\max(X_j)$ and $\min(X_j)$ represent the maximum and minimum values of the j-th indicator respectively. H_j represents the entropy value of the jth indicator. $p_{i,j} = \frac{Z_{i,j}}{\sum_{i=1}^{n} Z_{i,j}}$. When $p_{i,j} = 0$, let $p_{i,j} \ln p_{i,j} = 0$, and ensure that $0 \leq H_j \leq 1$. W_j represents the weight of the jth indicator. $0 \leq w_j \leq 1$, and $\sum_{j=1}^{m} w_j = 1$.

In terms of indicators selection, this study selects the number of high-tech industry enterprises, high-tech industry operating income, high-tech industry total profit, high-tech industry operating profit margin and high-tech industry operating income as the indicator system to evaluate the quality of high-tech industry economic growth. The measurement results are shown in Figure 1.

Vector autoregressive model. In order to test the interactive effect of the economic growth quality of high-tech industries in various regions of China, referring to existing research [3], this study chooses to use a vector autoregression model to estimate the response of each region to each other's shocks. The results of the Phillips-Perron unit root test show that the logarithmic series have reached a stable state and can be directly used for modeling. In addition, the AIC, SC and HQ information criteria and stability test results show that the first-order lag is the optimal lag order. The specific model is established as follows:

$$\mathcal{Q}_t = A \cdot \mathcal{Q}_{t-1} + \alpha + \mu_t,$$

where $\mathcal{Q}_t = (\ln \mathcal{E}_t, \ln \mathcal{C}_t, \ln \mathcal{W}_t)' \in \mathbb{R}^3$ is a 3-vector-column, $\mathcal{E}_t$, $\mathcal{C}_t$ and $\mathcal{W}_t$ are the

qualities of high-tech industry economic growth in the eastern region, the central region and the western region respectively, $A = (A_{i,j})_{i,j=1}^{3} \in \mathbb{R}^{3\times 3}$ is a (3×3)-matrix of the coefficients to be estimated, $\alpha = (\alpha_i)_{i=1}^{3} \in \mathbb{R}^3$ is a 3-vector-column of constant terms, $\mu_t = (\mu_{t,i})_{i=1}^{3} \in \mathbb{R}^3$ is a 3-vector-column of random perturbation terms.

Data source. In this study, the data for measuring the quality of high-tech industry economic growth in China are all from China Statistical Yearbook and China Statistical Yearbook of Science and Technology.

3 Results and analysis

Figure 1 reports the measurement results and changing trends of the quality of economic growth in high-tech industries in China and its regions. The results show that the quality of economic growth in high-tech industries in China has been growing steadily during the study period, but there are obvious regional differences. The eastern region is significantly stronger than other regions.

Figure 2 reports the interactive effect of the quality of high-tech industry economic growth among different regions in China. The results show that: the improvement of the economic growth quality of high-tech industries in the central and western regions can promote the economic growth quality of high-tech industries in the eastern region, while the improvement of the economic growth quality of high-tech industries in the eastern region not only cannot effectively drive the improvement of the economic growth quality of high-tech industries in the central and western regions, but may even have a phased negative impact on them. Considering that the economic growth quality of high-tech industries in the eastern region is far better than that in the central and western regions, the economic growth quality of high-tech industries in the eastern region is likely to have a siphon effect on the central and western regions. At the same time, the improvement of the economic growth quality of high-tech industries in the central region can promote the economic growth quality of high-tech industries in the western region, but the improvement of the economic growth quality of high-tech industries in the western region cannot effectively improve the economic growth quality of high-tech industries in the central region. Considering that the economic growth quality of high-tech industries in the central region is slightly better than that in the western region, the central region is likely to have a radiation effect on the western region.

4 Conclusion

Based on China's provincial panel data from 2012 to 2023, this study uses the entropy weighted model to measure the quality of high-tech industry economic growth in China and in each region, and on this basis, uses the vector autoregression model to test the interactive effect of the quality of high-tech industry economic growth in each region of China. The main conclusions are as follows: First, the quality of high-tech industry economic growth in China is showing a stable growth trend, but with obvious regional

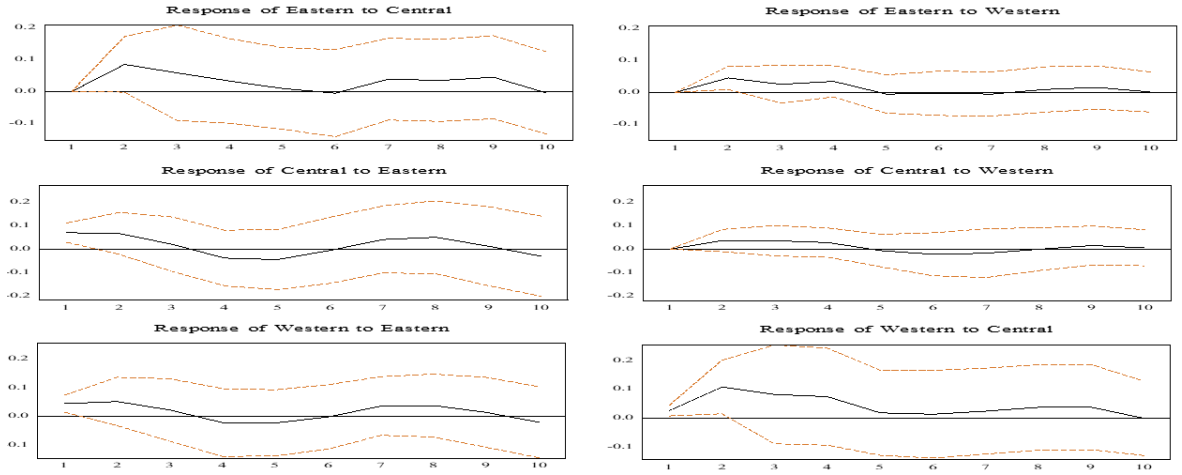


Figure 2: Interactive effects of high-tech industry economic growth quality among different regions in China

differences. The quality of high-tech industry economic growth in the eastern region is significantly stronger than that in other regions. Second, the quality of high-tech industry economic growth in the eastern region is likely to have a siphon effect on the central and western regions, and the central region is likely to have a radiation effect on the western region.

References

1. Oh D. H, Danilchanka A., Zhalezka B., Siniauskaya V. (2025).The Transition of Economy from Analogue to Digital in the XXI Century by the case of the Republic of Korea. *Eastern European Journal for Regional Studies*. Vol. **7**, Num. **1**, pp. 107-131.

2. Gao J. C. (2025). Internet Development Empowering Innovation Activities to Achieve Efficient and Balanced Development: Evidence From China. *International Journal of Knowledge Management*. Vol. **21**, Num. **1**, pp. 1-20.

3. Rajab K., Kamalov F., Cherukuri A. K. (2022). Forecasting COVID-19: vector autoregression-based model. *Arabian journal for science and engineering*. Vol. **47**, Num. **1**, pp. 68516860.

SEQUENTIAL ANALYSIS OF DATA UNDER DISTORTION: PERFORMANCE, ROBUSTNESS AND IMPLEMENTATION

A.Yu. Kharin[1]
[1]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]KharinAY@bsu.by

The problem of sequential testing of hypotheses on parameters of a stochastic data flow is considered under distortions (deviations from the hypothetical model assumptions). Simple and composite hypotheses setting are investigated for different hypothetical models of data. The interest is focused on three areas: performance characteristics (error probabilities and mathematical expectation of the random number of observations) calculation; robustness analysis of sequential tests under distortions; computer implementation of the considered sequential tests. The research is partially supported by the National Science Foundation, Grant No. F023Uzb-080.

Keywords: sequential test, stochastic data, distortion, error probability, expected number of observations, robustness

1 Introduction

In computer analysis of stochastic data, tasks of hypotheses testing appear quite frequently. To solve these problems mathematically, probability models and the approach based on sequential analysis [11] are intensively used [9], where the number of observations is considered to be not fixed a priori, supposed to be a random variable that depends on stochastic observations themselves. The sophisticated scheme of statistical inference in sequential analysis results in optimality of the decision process (the expected number of observations is minimized via that scheme provided error probabilities are restricted below predefined small levels) [10], with the price that the performance characteristics of sequential tests (error probabilities, expected number of observations) are problematic to be calculated with a given accuracy even for basic hypothetical probability models of data flows [8].

Sequential statistical tests are constructed mathematically to be optimal [1] under the hypothetical model of stochastic data flow, but in practice they are applied to real data sets that do not follow that hypothetical model exactly, the hypothetical model is distorted [2, 4, 7].

In the talk we present the results on performance and robustness analysis [6] of sequential statistical tests under simple and composite hypotheses setting for different models of stochastic data under distortions, and on robust sequential test construction.

2 Independent homogeneous observations

2.1 Case of simple hypotheses

Let on a probability space $(\Omega, \mathcal{F}, P)$ random variables x_1, $x_2, \dots$ be defined, $\forall t \in \mathbb{N}$, $x_t \in U = \{u_1, u_2, \dots, u_M\}$, $M < \infty$, $u_1 < u_2 < \dots < u_M$. Let these random variables be independent identically distributed, from a discrete probability distribution with a parameter $\theta \in \Theta = \{\theta_0, \theta_1\}$:

$$P(u;\theta) = P_\theta\{x_t = u\} = a^{-J(u;\theta)}, \quad t \in \mathbb{N}, \quad u \in U, \tag{1}$$

$a \in \mathbb{N} \setminus \{1\}$; $J(u;\theta)$: $U \times \Theta \longrightarrow \mathbb{N}_0$ is a function satisfying $\sum_{u \in U} a^{-J(u;\theta)} = 1$.

Consider two simple hypotheses w.r.t. the parameter θ:

$$H_0 : \ \theta = \theta_0, \ H_1 : \ \theta = \theta_1. \tag{2}$$

Introduce the notation:

$$\Lambda_n = \Lambda_n(x_1, \dots, x_n) = \sum\nolimits_{t=1}^{n} \lambda_t; \ \lambda_t = \log_a \left(P(x_t;\theta_1)/P(x_t;\theta_0)\right) \in \mathbb{Z}.$$

To test hypotheses (2) by n ($n = 1, 2 \dots$) observations consider the sequential probability ratio test (SPRT) [11]:

$$d_n = \mathbb{1}_{[C_+, +\infty)}(\Lambda_n) + 2 \cdot \mathbb{1}_{(C_-, C_+)}(\Lambda_n), \tag{3}$$

where $\mathbb{1}_D(\cdot)$ is the indicator function of the set D. The decisions $d_n = 0$ and $d_n = 1$ mean stopping of the observation process and the acceptance of the appropriate hypothesis. The decision $d_n = 2$ means that it is necessary to make the $(n+1)$-th observation. In (3) the thresholds $C_-, C_+ \in \mathbb{R}$, $C_- < C_+$ are the given values (parameters of the test). According to [11], we use

$$C_+ = \left[\log_a \left((1-\beta_0)/\alpha_0\right)\right], \ C_- = \left[\log_a \left(\beta_0/(1-\alpha_0)\right)\right], \tag{4}$$

where α_0, β_0 are given maximal possible values of the probabilities of type I and type II errors respectively. In fact, the true values α, β for the probabilities of type I and type II errors differ from α_0, β_0 [3].

For this model of data, the performance characteristics (error probabilities α, β and the mathematical expectations of the sample size t_0, t_1 under the correspondent hypothesis being true) are calculated in the explicit form in [4]. In the situation, where the assumption 1 does not hold, the approach to calculate the performance characteristics is given in [3].

For the case of distorted observations, the correspondent asymptotic expansions (w.r.t. one more extra parameter – the distortion level) are derived. The robust sequential test is constructed.

2.2 Binary random vectors case

The case where $U = \{u^1, \ldots, u^{2^K}\} = \left\{ \begin{array}{c} u_1 \\ \vdots \\ u_K \end{array} , \; u_i \in \{0,1\}, \; i = 1, \ldots, K \right\}$ is considered specifically. For this case, the test statistic is

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = n_0 \left(J(0; p^0) - J(0; p^1) \right) + (n - n_0) \left(J(1, p^0) - J(1, p^1) \right),$$

where n_0 is used for the number of observations equal to 0; $J(\cdot, \cdot)$ is a function described in (1); p^0 and p^1 are parameter vectors of the probabilities.

For the case of distorted observations, the correspondent asymptotic expansions are justified, and the minimax robust sequential test is constructed.

2.3 Multivariate data with block structure

In modern applied problems of econometrics, medicine, insurance and some others, data are often characterized by a high dimension. Also observations are characterized by a block structure: they can be split into blocks that may be considered as stochastically independent:

$$x_i = \left(x_i^1 \vdots \ldots \vdots x_i^K \right), \; i = 1, 2, \ldots .$$

This allows to use sequential scheme also within observations themselves, taking one block after another. Two positive aspects are brought with this scheme: if some components are missed, nevertheless the sequential test still can be used; mathematical expectation of the sample size becomes even less if compare to the case where entire observations only can be used. Both aspects are especially important when the number of observations is highly critical.

For this model of data, robustness analysis is performed and robust sequential tests are constructed.

2.4 Case of composite hypotheses

Suppose a sequence $x_1, x_2, \ldots$ of i.i.d. random variables is observed from a continuous distribution with the p.d.f. $p(x \mid \theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$ is an unknown value of random parameter. Consider two composite hypotheses [5]

$$H_0 : \; \theta \in \Theta_0, \; H_1 : \; \theta \in \Theta_1; \tag{5}$$

$\Theta_0, \Theta_1 \in \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$. Assume that the prior p.d.f. $p(\theta)$ is known.

One of the possible techniques to test the hypotheses (5) is using of weight functions proposed by Wald [11]. Introduce the notation:

$$W_i = \int_{\Theta_i} p(\theta) d\theta, \; w_i(\theta) = \frac{1}{W_i} \cdot p(\theta) \cdot \mathbb{1}_{\Theta_i}(\theta), \; \theta \in \Theta, \; i = 0, 1; \tag{6}$$

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \ln \frac{\int_\Theta w_1(\theta) \prod_{i=1}^n p(x_i \mid \theta) d\theta}{\int_\Theta w_0(\theta) \prod_{i=1}^n p(x_i \mid \theta) d\theta}. \quad (7)$$

For testing hypotheses (5), under the notation (6), (7) the following parametric family of tests is used:

$$N = \min\{n \in \mathbb{N} : \ \Lambda_n \notin (C_-, C_+)\}, \quad (8)$$

$$d = \mathbb{1}_{[C_+, +\infty)}(\Lambda_N), \quad (9)$$

where (8) gives the stopping rule, N is the random number of the observation, at which the decision d is made according to (9); $d = i$ means that the hypothesis H_i, $i = 0, 1$, is accepted; $C_- < 0$, $C_+ > 0$ are parameters of the test, which are usually choosen in practice according to (4).

Expressions in the explicite form are derived for the special case of the data distribution, and asymptotic expansions are constructed in the general case for the performance characteristics, also under distortions. The robust sequential test is constructed by the minimax risk criterion.

3 Heterogeneous observations

Let $x_1, x_2, ...$ be observations of time series with a trend:

$$x_t = \theta^T \psi(t) + \xi_t, \ t = 1, 2, 3, ...,$$

where $\psi(t) = (\psi_1(t), \psi_2(t), ..., \psi_m(t))^T$, $t \geq 1$, are the vectors of basic functions of trend, $\theta = (\theta_1, \theta_2, ..., \theta_m)^T \in \mathbb{R}^m$ is an unknown vector of coefficients, and $\{\xi_t, t \geq 1\}$ is the sequence of independent identically distributed random variables, $\xi_t \sim \mathcal{N}(0, \sigma^2)$.

Consider two simple hypotheses (2).

Denote the accumulated log-likelihood ratio statistic: $\Lambda_n = \Lambda_n(x_1, x_2, ..., x_n) = \sum_{t=1}^n \lambda_t$, where $\lambda_t = ln\left(\frac{p_t(x_t, \theta^1)}{p_t(x_t, \theta^0)}\right)$ is the log-likelihood ratio calculated on the observation x_t, and $p_t(x, \theta)$ is the probability density function of x_t provided the parameter value is θ.

To test these hypotheses, decision rule (3) is used.

For this model of data, the approach to calculate the performance characteristics of the sequential test is developed.

The situation where certain observations can be missed, is considered. The sequential test for this situation is constructed and its performance characteristics are evaluated.

The case of $M > 2$ simple hypotheses is also considered w.r.t. the vector θ. The following two sequential test were analyzed.

M-ary sequential probability ratio test. It uses the posterior probabilities of the hypotheses. The stopping time N_a and the final decision d_a for this test are defined by the equations:

$$N_a = \inf\left\{n \geq 1 : \ \exists m \in \{1, \ldots, M\}, \ P\{\mathcal{H}_m \mid x_1, \ldots, x_n\} > \frac{1}{1 + A_m}\right\},$$

$$d_a = \arg \max_{1 \le m \le M} P\{\mathcal{H}_m \mid x_1, \ldots, x_{N_a}\},$$

where $A_m \in (0, 1]$ are some specified constants, $m \in \{1, \ldots, M\}$, $d_a = m$ means that the decision in favor of the hypothesis H_m is made.

Matrix sequential probability ratio test. Denote

$$\Lambda_n(i, j) = \ln \left(\prod_{t=1}^{n} \frac{n_1(x_t; (\theta_i)^T \psi(t), \sigma^2)}{n_1(x_t; (\theta_j)^T \psi(t), \sigma^2)} \right);$$

$$\tau_i = \inf\{n \in \mathbb{N} : \ \Lambda_n(i, j) > b_{ij}, \ \forall j \in \{1, \ldots, M\} \setminus \{j\}\}, \ i \in 1, \ldots, M,$$

where $B = (b_{ij})$, $i, j \in \{1, \ldots, M\}$, is the matrix of the test thresholds (using them, the error probabilities of the test are controlled by the user f the decision rule). For this test the stopping time N_b and the final decision d_b are defined as follows:

$$N_b = \min\{\tau_i : \ i \in \{1, \ldots, M\}\}, \ d_b = \arg \min_{i \in \{1, \ldots, M\}} \tau_i.$$

For the two sequential tests defined above, the termination with probability 1 property and the finiteness of all moments of the random stopping time are proved. For the M-ary sequential probability ratio test, upper bounds for the error probabilities are derived. A robustified version of the matrix sequential probability ratio test is constructed and its properties are analyzed via numerical experiments.

4 Observations with Markov dependencies

Let the data flow be dependent observations forming a homogeneous Markov chain $x_1, x_2, \ldots$, with possible values in the set $V = \{0, 1, \ldots, M-1\}$. Denote the vector of initial states probabilities by $\pi = (\pi_i)$, $i \in V$, and the one-step transition probabilities matrix by $P = (p_{ij})$, $i, j \in V$, that are: $\mathrm{P}\{x_1 = i\} = \pi_i$, $\mathrm{P}\{x_n = j \mid x_{n-1} = i\} = p_{ij}$, $i, j \in V$, $n > 1$.

There are two hypotheses concerning the Markov chain parameters introduced above: $\mathcal{H}_0$: $\pi = \pi^{(0)}$, $P = P^{(0)}$ with the alternative $\mathcal{H}_1$: $\pi = \pi^{(1)}$, $P = P^{(1)}$, where $\pi^{(0)}$, $\pi^{(1)}$ are the given values of the initial states probabilities vector, $P^{(0)} \neq P^{(1)}$ are the one-step transition probabilities matrices for correspondent hypotheses. Denote also:

$$\lambda_1 = \ln \frac{P_1\{x_1\}}{P_0\{x_1\}}, \ \lambda_k = \ln \frac{P_1\{x_k \mid x_{k-1}\}}{P_0\{x_k \mid x_{k-1}\}}, \ k > 1, \ \Lambda_n = \sum_{k=1}^{n} \lambda_k, \ n \in \mathbb{N},$$

where $P_s\{x_1\}$ is the probability to observe the value x_1, $P_s\{x_k \mid x_{k-1}\}$ is the probability to observe x_k at the moment k provided at the moment $k-1$ the value x_{k-1} was observed, if hypothesis $\mathcal{H}_s$, is true $s \in \{0, 1\}$.

Construct the sequential decision rule to decide in favor of $\mathcal{H}_0$ or $\mathcal{H}_1$. According to this decision rule, with given thresholds values $C_-, C_+ \in \mathbb{R}$, $C_- < 0$, $C_+ > 0$, hypothesis $\mathcal{H}_0$ is accepted on the basis of n observations, if $\Lambda_n \le C_-$. Hypothesis $\mathcal{H}_1$ is accepted, if $\Lambda_n \ge C_+$, otherwise the observation process is not stopped, and $(n+1)$-th observation is requested.

Correspondent families of modified sequential decision rules are developed. Within the developed families, the robust sequential decision rules are constructed with the minimax risk criterion. Results are generalized to the situation where data form a high order Markov chain.

5 Implementation and conclusion

The sequential tests considered in the talk are implemented in the form of the computer procedures and are forming the library within R statistical software. Each procedure includes the likelihood ratio sequential test implementation, calculation of its performance characteristics, robustness analysis, and a robustified version.

The future research includes sequential test construction and analysis under missing values and partially available observations.

References

1. Aivazian S. A. (1959) Comparison of Optimal Properties of the Tests of Neyman-Pearson and Wald. *Teoriya veroyatnostei i ee primeneniya*. Vol. **4** (1), pp. 86-93.

2. Huber P. J. (2004). *Robust Statistics: Theory and Methods*. Wiley, New York.

3. Kharin A. (2013). *Robustness of Bayesian and Sequential Statistical Decision Rules*. BSU, Minsk.

4. Kharin A., Galinskij V. (1999). On minimax robustness of Bayesian statistical prediction. *Probability Theory and Mathematical Statistics*. TEV, pp. 259-266.

5. Kharin A. (2017). An Approach to Asymptotic Robustness Analysis of Sequential Tests for Composite Parametric Hypotheses. *Journal of Mathematical Sciences*. Vol. **227** (2), pp. 196-203.

6. Kharin A.Y., Kishylau D.V. (2015) Robust sequential test for hypotheses about discrete distributions in the presence of "outliers". *Journal of Mathematical Sciences*. Vol. **205** (1), pp. 68-74.

7. Kharin A. (2005). Robust Bayesian prediction under distortion of prior and conditional distributions. *Journal of Mathematical Sciences*. Vol. **126** (1), pp. 992-997.

8. Lai T. L. (2001). Sequential Analysis: Some Classical Problems and New Challenges. *Statistica Sinica*. Vol. **11**, pp. 303-408.

9. Mukhopadhyay N., Datta S., Chattopadhyay S. (2004). *Applied Sequential Methodologies*. Marcel Dekker, New York.

10. Sirjaev A.N. (1973) *Statistical Sequential Analysis: Optimal Stopping Rules*. AMS, New York.

11. Wald A. (1947). *Sequential Analysis*. Wiley, New York.

STATISTICAL ANALYSIS OF HIGH-ORDER MARKOV CHAINS

Yu.S. Kharin[1]

[1]*Research Institute for Applied Problems of Mathematics and Informatics*
[1]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]kharin@bsu.by

Results on development of the theory of probabilistic and statistical analysis of high order Markov chains are presented.

Keywords: high order Markov chain, statistical analysis, parsimonious model

1 Introduction

Digitalization of any society gives a significant rise to a lot of discrete-valued data. If a discrete-valued data are considered and analyzed in dynamics (in dependence of discrete time $t \in \mathbf{Z}$ we get a discrete-valued time series (DTS) $x_t \in A$, where A is some discrete set:

$$A = \{0, 1, \ldots, N-1\},\ N = |A|,\ 2 \le N < +\infty.$$

The existed theory of probabilistic and statistical analysis of time series is deep developed for the so-called "continious" time series when A is a nonzero Lebesque measure subset in R^m for some $m \in \mathbf{N}$. But for the DTS this theory is on the beginning stage only [1]. The main problem in this stage appears in modelling of high depth stochastic dependencies in DTS $\{x_t\}$.

Indicate main possible applications of this theory in practice: genetics (computer recognition and analysis of genetic sequences, $N = 4$); economics and finance (prediction of financial time series); sociology (modeling of social behavior); medicine (computer diagnostics, monitoring in personalized medicine; cybersecurity (evaluation of the safety for computer information systems, $N = 2$).

A short history of the development of the theory of statistical analysis for DTS can be found in [2, 3, 4]. A fresh review [5] indicates following topical research directions: 1) methods based on the generalized linear model GLM; 2) methods based on the integer autoregression; 3) models governed by dynamical parameters; 4) parsimonious models based on high order Markov chains. This paper aims to contribute to the appearing theory in the fourth direction.

2 Parsimonious models for high order Markov chains and approaches to construction of these models

An universal model for description of high depth stochastic dependencies is proposed by J. Doob: the homogeneous Markov chain (MC(s)) of sufficiently large order $s \in \mathbf{N}$ on some probabilistic space $(\Omega, \mathcal{F}, \mathbf{P})$. It is determined by the conditional transition probabilities:

$$\mathbf{P}\{x_t = j_t | \mathcal{F}_{t-1}\} = \mathbf{P}\{x_t = j_t | X_{t-s}^{t-1} = J_{t-s}^{t-1}\} = p_{J_{t-s}^{t-1}, j_t},\ t \in \mathbf{Z}, \tag{1}$$

where $\mathcal{F}_{t-1} = \sigma\left(\{x_\tau : \tau \le t-1\}\right)$, $X_{t-s}^{t-1} = (x_{t-s}, \ldots, x_{t-1})' \in A^s$, $J_1^s = (j_1, \ldots, j_s)' \in A^s$, $P = \left(p_{J_1^{s+1}}\right)$, $J_1^{s+1} \in A^{s+1}$ is the $(s+1)$-dimensional matrix of one-step transition probabilities.

Under the well known ergodicity conditions [1] there exists the single stationary s-dimensional distribution $\pi_{J_1^s}$, $J_1^s \in A^s$, satisfying the system of linear algebraic equations:

$$\begin{array}{l} \sum\limits_{j_1 \in A} \pi_{J_1^s} p_{J_1^{s+1}} = \pi_{J_2^{s+1}},\ J_2^{s+1} \in A^s; \\ \sum\limits_{J_1^s \in A^s} \pi_{J_1^s} = 1. \end{array} \tag{2}$$

Theorem 1. *The s-dimensional stationary distribution (2) determines all finite-dimensional joint probability distributions for the MC(s), $T \in \mathbf{N}$:*

$$\mathbf{P}\{x_1 = j_1, x_2 = j_2, \ldots, x_T = j_T\} = \begin{cases} \sum\limits_{j_{T+1}, \ldots, j_s \in A} \pi_{J_1^s},\ \textit{if } T < s, \\ \pi_{J_1^s},\ \textit{if } T = s, \\ \pi_{J_1^s} \prod\limits_{t=s+1}^{T} p_{J_{t-s}^{t-1}, j_t},\ \textit{if } T > s. \end{cases}$$

For the universality of the MC(s)-model we need to pay by its complexity: the number of independent parameters $P = \left(p_{J_1^{s+1}}\right)$ of this model $D_{\mathrm{MC}(s)} = N^s(N-1) = O\left(N^{s+1}\right)$ increases exponentially w.r.t. the order s. To avoid this "curse of dimensionality" we propose to use parsimonious Markov chains of order s (PMC(s)) that have parsimonious parametric representation of the one-step transition probabilities matrix:

$$P = \left(p_{J_1^{s+1}}\right),\ p_{J_1^{s+1}} =:: p_\alpha\left(J_1^{s+1}\right), \quad \alpha = (\alpha_1, \ldots, \alpha_d)' \in R^d, \tag{3}$$

where α is the vector parameter with a small dimensionality $d = D_{\mathrm{PMC}(s)} \ll D_{\mathrm{MC}(s)}$; the compression coefficient $\varkappa = D_{\mathrm{PMC}(s)} / D_{\mathrm{MC}(s)} \ll 1$.

We develop four approaches to construction of PMC(s): 1) reduction of the set of possible values for the elements of matrix P; 2) using of standard parametric families of discrete probability distributions for transition probabilities; 3) PMC(s) based on artificial neural networks; 4) PMC(s) based on exponential families and sufficient statistics.

3 PMC(s) based on reduction of the set of transition probabilities

Let $Q = \left(q_{J_1^{r+1}}\right)$ be some stochastic $(r+1)$-dimensional matrix $(1 \le r < s)$: $0 \le q_{J_1^{r+1}} \le 1$, $J_1^{r+1} \in A^{r+1}$, $\sum\limits_{j_{r+1} \in A} q_{J_1^{r+1}} \equiv 1$; $B = B(J_1^s; \alpha) : A^s \times R^m \to A^r$, be some discrete function with a parameter $\alpha = (\alpha_i) \in R^m$. Then the $(s+1)$-dimensional matrix of transition probabilities P is reduced into the $(r+1)$-dimensional matrix Q by the discrete transformation:

$$p_{j_1,\ldots,j_s,j_{s+1}} = q_{B(j_1,\ldots,j_s;\alpha),j_{s+1}}, \quad J_1^{s+1} \in A^{s+1}. \tag{4}$$

The total number of parameters for the model (4) is $d = N^r(N-1) + m$.

Examples for the model (4) are: Markov chain MC(s,r) of order s with r partial connections [6, 7]; Markov chain of conditional order MCCO(s,L) [8, 9]; variable length Markov chain [4].

Illustrate now our results on the family (4) for the MC(s,r)-model [6, 7]:

$$p_{j_1,\ldots,j_s,j_{s+1}} = q_{j_{m_1^0},\ldots,j_{m_r^0},j_{s+1}}, \quad J_1^{s+1} \in A^{s+1}, \tag{5}$$

where $r \in \{1, 2, \ldots, s\}$ is the number of connections; $M_r^0 = (m_1^0, m_2^0, \ldots, m_r^0)$ is the connection template; $Q = \left(q_{J_1^{r+1}}\right)$, $J_1^{r+1} \in A^{r+1}$, is the $(r+1)$-dimensional stochastic matrix. If $r = s$, then MC$(s,s) \equiv$ MC(s) is the Markov chain with full connections.

Introduce the notation: $X_1^n = (x_1, x_2, \ldots, x_n)$ is an observed realization of length $n \in \mathbf{N}$; $F\left(X_t^{t+s}; M_r\right) = (x_{t+m_1-1}, \ldots, x_{t+m_r-1}, x_{t+s})$ is the selector function of the $(r+1)$-th order; $\mathbf{1}\{B\}$ is the indicator function of the event B; the dot "$\bullet$" used instead of any index means summation on all its values;

$$\begin{gathered} \nu_{J_1^{r+1}}(M_r) = \sum_{t=1}^{n-s} \mathbf{1}\{F\left(X_t^{t+s}; M_r\right) = J_1^{r+1}\},\ J_1^{r+1} \in A^{r+1}; \\ \hat{\mu}_{J_1^{r+1}}(M_r) = \nu_{J_1^{r+1}}(M_r)/(n-s); \end{gathered} \tag{6}$$

$$\hat{I}_{r+1}(M_r) = \sum_{J_1^{r+1} \in A^{r+1}} \hat{\mu}_{J_1^{r+1}}(M_r) \ln \frac{\hat{\mu}_{J_1^{r+1}}(M_r)}{\hat{\mu}_{J_1^r \bullet}(M_r)\hat{\mu}_{\bullet j_{r+1}}} \ge 0$$

is the "plug-in" estimator of the Shannon information.

Theorem 2. *MLEs* $\hat{M}_r$, $\hat{Q} = \left(\hat{q}_{J_{r+1}}\right)$, $J_{r+1} \in A^{r+1}$, *for the parameters* M_r^0, Q^0 *are determined by the following expressions:*

$$\hat{M}_r = \arg \max_{M_r \in M} \hat{I}_{r+1}(M_r), \tag{7}$$

$$\hat{q}_{J_1^{r+1}} = \begin{cases} \hat{\mu}_{J_1^{r+1}}\left(\hat{M}_r\right) / \hat{\mu}_{J_1^r \bullet}\left(\hat{M}_r\right), \ \textit{if} \ \hat{\mu}_{J_1^r \bullet}\left(\hat{M}_r\right) > 0, \\ 1/N, \ \textit{if} \ \hat{\mu}_{J_1^r \bullet}\left(\hat{M}_r\right) = 0. \end{cases} \tag{8}$$

Theorem 3. *If MC*(s, r) *determined by (5) is stationary and the connection template* $M_r^0 \in \mathrm{M}$ *satisfies the identification condition, then at* $n \to \infty$ *MLEs* $\hat{M}_r$, $\hat{Q}$ *determined by (7) are consistent:*

$$\hat{M}_r \xrightarrow{\mathbf{P}} M_r^0, \quad \hat{Q} \xrightarrow{L_2} Q^0, \tag{9}$$

$$\Delta_n^2 = \mathbf{E}\left\{\left\|\hat{Q} - Q^0\right\|^2\right\} = \frac{1}{n-s} \cdot \sum_{J_1^{r+1} \in A^{r+1}} \frac{\left(1 - q^0_{J_1^{r+1}}\right) q^0_{J_1^{r+1}}}{\mu_{J_1^r \bullet}(M_r^0)} + o\left(\frac{1}{n}\right). \tag{10}$$

4 PMC(s) based on parametric families of standard discrete distributions for transition probabilities

Let $\{q_j(\theta) : j \in A\}$ be some standard discrete probability distribution on A with some parameter $\theta = (\theta_j) \in \Theta \subseteq R^L$; $\theta = \theta(j_1, \ldots, j_s; \alpha) : A^s \times R^m \to R^L$ be some parametric function with a parameter $\alpha \in R^m$ that determines dependence on the prehistory. Generating equation for transition probabilities is:

$$p_{J_1^{s+1}} = q_{j_{s+1}}\left(\theta(J_1^s; \alpha)\right), J_1^{s+1} \in A^{s+1}. \tag{11}$$

Examples for the PMC(s)-model (11) are: Jacobs-Lewis model [2]; Raftery model [3]; Binomial conditionally nonlinear autoregressive model BiCNAR(s) [10, 11]; Semibinomial conditionally nonlinear autoregressive model SBiCNAR(s) [10]; Binary conditionally nonlinear autoregressive model BCNAR(s) [12]; Poisson conditionally nonlinear autoregressive model $\Pi CNAR(s)$ [13].

Illustrate here our results on the BiCNAR(s) model:

$$\begin{gathered} \mathbf{P}\left\{x_t{=}j_t | X_{-\infty}^{t-1}{=}J_{-\infty}^{t-1}\right\} \equiv \mathbf{P}\left\{x_t{=}j_t | X_{t-s}^{t-1}{=}J_{t-s}^{t-1}\right\} = C_{N-1}^{j_t} p_t^{j_t} (1-p_t)^{N-1-j_t}, \\ j_k \in A, \ k \in \mathbf{Z}; \\ p_t = p\left(J_{t-s}^{t-1}\right) = F_0\left(J_{t-s}^{t-1}\right), \ t \in \mathbf{Z}, \end{gathered} \tag{12}$$

where $p_t \in [0, 1]$ is the parameter of the Binomial distribution. Consider the case when the function $F_0(\cdot)$ in (12) is approximated by the given system of m linearly independent on A^s base functions $\psi(J_1^s) = (\psi_i(J_1^s)) \in R^m$:

$$F_0\left(J_{t-s}^{t-1}\right) ::= F\left(\sum_{i=1}^m a_i \psi_i\left(J_{t-s}^{t-1}\right)\right), \ J_{t-s}^{t-1} \in A^s, \tag{13}$$

where $F(\cdot): R^1 \to [0,1]$ is some absolutely continuous distribution function; $B=(b_i)$ are unknown parameters of the model.

To estimate unknown parameters $\{a_i\}$ by an observed realization $X_1^T{=}(x_1,...,x_T){\in}A^T$ of length $T \in \mathbf{N}$ we use the FBE-method [12]. Introduce the notation:

$$\nu_s^T(J_1^s)=\sum_{t=s}^{T}\mathbf{1}\left\{X_{t-s+1}^T=J_1^s\right\},\ J_1^s\in A^s;\ J^{(T)}=\left\{J_1^s\in V^s:\nu_s^T(J_1^s)>0\right\}\subseteq A^s;$$

$$\hat{\pi}_{J_1^s}{=}\frac{\nu_s^T(J_1^s)}{(T-s+1)},\ \hat{p}_J{=}\begin{cases}\dfrac{\sum\limits_{t=s+1}^{T}x_t\mathbf{1}\left\{X_{t-s}^{t-1}=J\right\}}{(N-1)\hat{\pi}_{J_1^s}},\ J{\in}J^{(T)};\ \hat{u}(J_1^s){=}F^{-1}\left(\hat{p}_{J_1^s}\right),\ J_1^s{\in}J(T);\\ 1/N,\quad J\notin J^{(T)},\end{cases}$$

$$D=(d_{kl})=\sum_{J_1^s\in J^{(T)}}\psi(J_1^s)\psi'(J_1^s)\in\mathbb{R}^{m\times m},\ E=(e_k)=\sum_{J_1^s\in J^{(T)}}\hat{u}(J_1^s)\psi(J_1^s)\in\mathbb{R}^m. \quad (14)$$

Theorem 4. *If the PMC(s) $x_t \in A$ determined by (12)–(14) is ergodic and $|D| \neq 0$, then for the increasing length of the time series $T \to +\infty$ there exists the FBE-estimator $\hat{B}{=}\left(\hat{b}_i\right){=}D^{-1}E$ that is consistent: $\hat{B} \xrightarrow{\mathbf{P}} B^0$, and asymptotically normal: $\sqrt{T}(\hat{B}-B^0) \xrightarrow{D} \mathcal{N}_m(O_m,\Sigma)$, $\Sigma=\left(\Psi^T\Psi\right)^{-1}\Psi^T\Lambda\Psi\left(\Psi^T\Psi\right)^1$, where Λ is the known $(N^s{\times}N^s)$-matrix [14] and the $(m\times N^s)$-matrix Ψ consists of N^s m-columns $\psi(J)$ for all lexicographically ordered values $J \in A^s$.*

5 PMC(s) based on Artificial Neural Networks (ANNs)

Illustrate an application of ANNs for modelling of high depth stochastic dependencies for the example of BiCNAR(s) model (12). Instead of the basic approximation (13) let us use the neural network approximation [15]:

$$F_0\left(J_1^s\right)=F\left(\sum_{i=1}^{m}b_iF_i\left(\sum_{k=1}^{s}a_{ik}j_k\right)\right),\ J_1^s\in A^s, \quad (15)$$

where $F(\cdot),F_1(\cdot),\ldots,F_m(\cdot): R^1 \to [0,1]$ are some given absolutely continious distribution functions; $2 \le m < +\infty$; $B=(b_1,\ldots,b_m)' \in R^m$, $A_i=(a_{i1},\ldots,a_{is})' \in R^s$, $i \in \{1,\ldots,m\}$ are unknown parameters. The nonlinear dependence (15) is represented by the ANN (see Figure 1) with s inputs, one output, m neurons in the first layer and one neuron in the second layer. Number of parameters depends linearly on the order $s: d=m(s+1)$.

Let us note, that the problem of statistical estimation of parameters $B, A_1, \ldots, A_m$ by the observed time series X_1^T is invariant to the substitution:

$$\tilde{B} = \left(\tilde{b}_1, \ldots, \tilde{b}_m\right), \; \tilde{A}_i = \left(\tilde{a}_{i1}, \ldots, \tilde{a}_{im}\right),$$

$$\tilde{b}_i = b_{\pi_i}, \; \tilde{a}_{ij} = a_{\pi_i j}, \; j = 1, \ldots, s; \; i = 1, \ldots, m,$$

where $\pi = (\pi_1, \ldots, \pi_m)$ is an arbitrary substitution on $\{1, 2, \ldots, m\}$. In [15] we propose an algorithm to construct consistent estimators $\hat{B}, \{\hat{A}_i\}$.

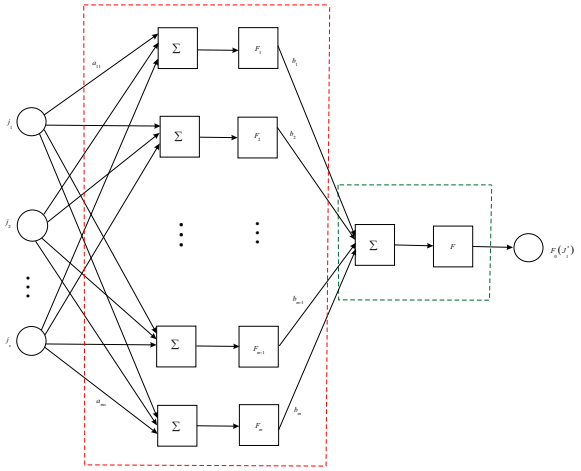


Figure 1: ANN representing the approximation (15)

6 PMC(s) based on exponential families and sufficient statistics

This class of PMC(s) models is determined by the following exponential family of conditional transition probabilities (1):

$$\begin{gathered}\mathbf{P}\left\{x_t{=}j_s|X_{t-s}^{s-1}{=}J_0^{s-1}\right\} = \exp\left\{h_0\left(j_s; J_0^{s-1}\right) + \sum_{i=1}^{m} \eta_i h_i\left(j_s; J_0^{s-1}\right) - \varphi\left(\eta; J_0^{s-1}\right)\right\}, \\ \varphi\left(\eta; J_0^{s-1}\right) = \ln \sum_{j_s \in A} \exp\left\{h_0\left(j_s; J_0^{s-1}\right) + \sum_{i=1}^{m} \eta_i h_i\left(j_s; J_0^{s-1}\right)\right\}, \; J_0^s \in A^{s+1},\end{gathered} \tag{16}$$

where $\{h_1(\cdot), \ldots, h_m(\cdot)\}$ are known base functions called sufficient statistics. For abbreviation we call this model MCSS(s): Markov chain of order s with sufficient statistics. The problem of statistical estimation of the unknown parameters $\eta = (\eta_i) \in R^m$ is solved in [16].

7 Application of the developed theory in computer analysis of real data sets

All algorithms based on the developed theory were tested on simulated data and illustrated good accordance with the proved theoretical properties. Here we give some results of computer experiments with the developed algorithms on real data.

7.1 Modelling of the wind speed data

The discrete-valued time series of the daily average wind speed at Malin Head (North of Ireland during the period 1961 – 1978) $x_t \in \{0, 1, 2\}$, $N = 3$, of the length $T = 6574$ was fitted by the MC(s, r)-model for $s = \{1, 2, \ldots, 7\}$, $r = \{1, 2, \ldots, 7\}$. Table 1 presents the values of the BIC for different pairs (s, r).

Table 1: Bayesian Information Criterion for models of the wind speed data

Model	BIC	Model	BIC	Model	BIC	Model	BIC
MC(1,1)	8127.52	MC(4,2)	8139.12	MC(5,5)	8621.97	MC(7,1)	9041.43
MC(2,1)	8777.63	MC(4,3)	8164.79	MC(6,1)	9016.23	MC(7,2)	8163.07
MC(2,2)	8096.08	MC(4,4)	8332.77	MC(6,2)	8148.48	MC(7,3)	8197.91
MC(3,1)	8849.90	MC(5,1)	8984.10	MC(6,3)	8190.78	MC(7,4)	8323.19
MC(3,2)	**8079.81**	MC(5,2)	8129.83	MC(6,4)	8350.82	MC(7,5)	8599.09
MC(3,3)	8143.13	MC(5,3)	8177.92	MC(6,5)	8576.92	MC(7,6)	8973.15
MC(4,1)	8956.11	MC(5,4)	8349.62	MC(6,6)	8969.54	MC(7,7)	9575.64

The best fitted model is the MC(3,2) with $\hat{M}_r = (1, 3)$ and the transposed matrix

$$\hat{Q}' = \begin{pmatrix} 0.27 & 0.08 & \mathbf{0} & 0.22 & 0.04 & \mathbf{0} & 0.21 & 0.02 & 0 \\ 0.73 & 0.86 & 0.63 & 0.78 & 0.82 & 0.52 & 0.79 & 0.72 & 0.43 \\ \mathbf{0} & 0.06 & 0.37 & \mathbf{0} & 0.14 & 0.48 & \mathbf{0} & 0.26 & 0.57 \end{pmatrix}.$$

The fitted model MC$(3, 2)$ detects significant dependencies in this data.

7.2 Genomic sequencing

We used the drosophila genome sequence (www.fruitfly.org): N=4, T=$5 \cdot 10^5$, s_-=1, $s_+ = 8$, $r_- = 1$, $r_+ = 8$. The best fitted model is the MC$(6, 3)$ with the template $\hat{M}_r$=$(1, 5, 6)$ and the matrix $\hat{Q}$ visualized in Figure 2. Here on "x-axis" the values of $\hat{M}_r$-prehistory are indicated, "y-axis" gives the values of one-step transition probabilities to four states indicated by different levels of color.

7.3 Analysis of CG-patterns in genome

We took the complete Panthera tigris mitochondrion genome of the length T=16990 (available from NCBI Nucleotide data base, ID EF551003.1) and extracted the binary

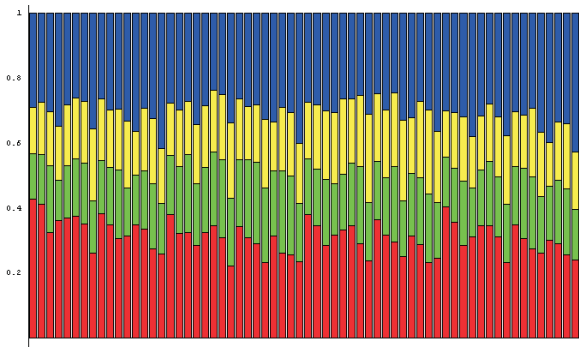


Figure 2: Visualization of the matrix $\hat{Q}$ for the genomic sequencing

sequence x_t of its CG-indicators: $x_t = 1$ iff the t'th nucleotide is Guanine or Cytosine, $t = 1, \ldots, T$. Portion of "1" in X_1^T is known as CG-content and plays important role in bioinformatics.

In order to evaluate individual and pairwise impact of the lagged variables X_{t-s}^{t-1} on x_t we fitted the BCNAR(s)-model (up to $s = 15$) for $N = 2$ with the bilinear bases $\{\psi_i(\cdot)\}$ and the Gaussian c.d.f. $F(\cdot) = \Phi(\cdot)$. Two fitted BIC – adequate BCNAR-models for $s = 10; 15$ respectively, are ($\zeta_t = (-1)^{x_t}$):

$$\mathbf{P}\{x_t|x_{t-1}, \ldots\} = \Phi(-0.3962 + 0.0313\zeta_{t-1} + 0.0241\zeta_{t-3} + 0.033\zeta_{t-10} +$$
$$+0.045\zeta_{t-3}\zeta_{t-6} - 0.0576\zeta_{t-3}\zeta_{t-10}),$$
$$\mathbf{P}\{x_t|x_{t-1}, \ldots\} = \Phi(-0.1319 + 0.022\zeta_{t-1} + 0.0269\zeta_{t-6} + 0.0248\zeta_{t-15} -$$
$$-0.0434\zeta_{t-6}\zeta_{t-15}),$$

8 Conclusion

This paper summarizes research results in the Research Institute for Applied Problems of Mathematics and Informatics on the development of the probabilistic and statistical analysis of high order Markov chains. To avoid the hard "curse of dimensionality" we propose some approaches to construct parsimonious models for high order Markov chains. These approaches are illustrated on parsimonious models constructed by new methods and algorithms for statistical estimation of model parameters. Theoretical results are accompanied by computer analysis of real data sets. The developed theory should be extended for statistical analysis of multivariate and space-temporal discrete-valued sequences.

References

1. Kharin Yu. (2013). *Robustness in Statistical Forecasting*. Springer, Heidelberg / New York / Dordrecht / London. – 356 p.

2. Jacobs P.A., Lewis P.A.W. (1978). Discrete time series generated by mixtures I: correlational and runs properties. *Journal of the Royal Statistical Society. Ser. B.* Vol. **40**, No. 1, pp. 94-105.

3. Raftery A. (1985). A model for high-order Markov chains. *Journal of the Royal-Statistical Society. Ser. B.* Vol. **47**, No. 3, pp. 528-539.

4. Buhlmann P., Wyner A.J. (1999). Variable length Markov chains. *The Annals of Statistics*. Vol. **27**, No. 2, pp. 480-513.

5. Kharin Yu., Fokianos K., Fried R., Voloshko V. (2022). Statistical analysis of multivariate discrete-valued time series. *Journal of Multivariate Analysis*. Vol. **188**: 104805.

6. Kharin Yu.S. (2004). Markov chains with r-partial connections and their statistical estimators. *Trans. Nat. Acad. Sci. Belarus*. Vol. **48**, No. 1, pp. 40-44.

7. Kharin Yu.S., Piatlitski A.I. (2007). A Markov chain of order s with r partial connections and statistical inference on its parameters. *Discrete Mathematics and Applications*. Vol. **17**, No. 3, pp. 295-317.

8. Kharin Yu., Maltsew M. (2014). Markov chain of conditional order: Properties and Statistical Analysis. *Austrian Jornal of Statistics*. Vol. **43**, No. 3-4, pp. 205-217.

9. Kharin Yu., Maltsew M. (2017). Statistical analysis of high-order dependencies. *Acta et Commentationes Universitetas Tartuensis de Mathematica*. Vol. **21**, No. 1, pp. 37-45.

10. Kharin Yu. (2020). Statistical analysis of discrete-valued time series by parsimonious high-order Markov chains. *Austrian Journal of Statistics*. Vol. **49**, No. 4, pp. 76-88.

11. Kharin Yu.S., Voloshko V.A. (2019). Binomial conditionally nonlinear autoregressive model of discrete-valued time series and its probabilistic and statistical properties. *Transactions of the Institute of Mathematics of NAS Belarus*. Vol. **26**, No. 1, pp. 95-105.

12. Kharin Yu.S., Voloshko V.A., Medved E.A. (2019). Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series. *Mathematical Methods of Statistics*. Vol. **26**, No. 2, pp. 103-118.

13. Kharin Y., Kislach M. (2019). Statistical analysis of Poisson conditionally nonlinear autoregressive time series by frequencies-based estimators. *Proceedings of the 14-th International Conference. Minsk, Belarus "Pattern Recognition and Information Processing"*, pp. 233-236.

14. Kharin Yu., Voloshko V. (2021). Robust estimation for binomial conditionally nonlinear autoregressive time series based on multivariate conditional frequencies. *Journal of Multivariate Analysis*. Vol. **185(2)**: 104777.

15. Kharin Yu. (2021). Neural net models of binomial time series in data analysis. *Reports of National Academy od Sciences of Belarus*. Vol. **65**, No. 6, pp. 654-660.

16. Kharin Yu., Voloshko V. (2025). Statistical Analysis of Parsimonious High-Order Multivariate Finite Markov Chains Based on Sufficient Statistics. *Journal of Multivariate Analysis*. Vol. **208**: 105422.

ON NEW PARSIMONIOUS MODEL FOR HIGH-ORDER MARKOV CHAINS BASED ON SUFFICIENT STATISTICS

Yu.S. Kharin[1], V.A. Voloshko[2]

[1,2]*Research Institute for Applied Problems of Mathematics and Informatics*
[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]kharin@bsu.by, [2]valoshka@bsu.by

We construct a new parsimonious model MCSS(s) (Markov Chain of order s based on Sufficient Statistics) for discrete-valued time series. This MCSS(s) model has sufficient statistics of a simple form and strongly concave loglikelihood function under mild regularity conditions, which provides the uniqueness of the maximum likelihood estimator and its good computational properties. The research is partially supported by the National Science Foundation, Grant No. F023Uzb-080.

Keywords: efficient estimator, exponential family, parsimonious high-order Markov chain, sufficient statistics

1 MCSS(s) model for discrete-valued time series

Let us introduce the notation: $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{C}$ are respectively the sets of positive integers, nonnegative integers, all integers, real and complex numbers; $(\Omega, \mathcal{F}, \mathsf{P})$ is a general probabilistic space; $\mathbb{1}\{\cdot\}$, $\mathsf{P}\{\cdot\}$, $\mathsf{E}\{\cdot\}$, $\mathcal{L}\{\cdot\}$, $\mathcal{I}(\cdot)$ are respectively the indicator function of an event, functionals of the probability of an event, the expectation of a random variable, the probability distribution (law) of a random variable, a Fisher information matrix w.r.t. the model parameter; $0_n \in \mathbb{R}^n$, $\mathrm{Id}_n \in \mathbb{R}^{n\times n}$ are respectively a zero n-vector and the identity square matrix of order n; $u_a^b = (u_a, u_{a+1}, \ldots, u_b)$ is a subvector in some sequence $\{u_i\}$ for $a \le b$; $\langle u, v\rangle = \sum_i u_i v_i$ is the standard scalar product of real vectors $u = (u_i), v = (v_i)$.

The observed discrete-valued time series $\mathbf{x}_t \in \mathrm{A}$ on the probabilistic space $(\Omega, \mathcal{F}, \mathsf{P})$ is determined for discrete time $t \in \mathbb{Z}$ and ranging over some N-state space A, $N = |\mathrm{A}| < +\infty$. To avoid the well-known "curse of dimensionality" of fully connected Markov chain let us introduce a parsimonious model for a d-variate Markov chain of order s based on sufficient statistics developed in [1] (we call it an MCSS(s) model):

$$\mathsf{P}\left\{\mathbf{x}_t = x | \mathbf{x}_{t-s}^{t-1} = q\right\} = \mathcal{E}^{(q)}(x;\eta) ::= \exp\left(\mathfrak{h}_0^{(q)}(x) + \left\langle \eta, \mathfrak{h}^{(q)}(x)\right\rangle - \phi^{(q)}(\eta)\right), \quad (1)$$

$$\phi^{(q)}(\eta) = \ln \sum_{x\in\mathrm{A}} \exp\left(\mathfrak{h}_0^{(q)}(x) + \left\langle \eta, \mathfrak{h}^{(q)}(x)\right\rangle\right), \quad (2)$$

$$\mathfrak{h}^{(q)}(x) = \left(\mathfrak{h}_i^{(q)}(x)\right)_{i=1}^{m} \in \mathbb{R}^m,\ \eta = (\eta_i)_{i=1}^m \in \mathbb{R}^m,\ q \in \mathrm{A}^s,\ m \le N^s(N-1). \quad (3)$$

Relations (1)–(3) mean that each prehistory $q \in \mathrm{A}^s$ has its own exponential family $\mathcal{E}^{(q)}$ of conditional probability distributions $\mathcal{L}\left\{\mathbf{x}_t | \mathbf{x}_{t-s}^{t-1} = q\right\}$ on the support A, and all these

conditional probability distributions for all prehistories are determined by the common canonic parameter $\eta \in \mathbb{R}^m$, that is the parameter of model (1). The exponential family $\mathcal{E}^{(q)}$, $q \in \mathrm{A}^s$, is determined by $m+1$ functions $\{\mathfrak{h}_i^{(q)}(\cdot)\}_{i=0}^m$ on A; let us call $\mathfrak{h}_0^{(q)}(\cdot)$ the supporting function, and $\mathfrak{h}_i^{(q)}(\cdot)$, $i \in \{1,\dots,m\}$, the base functions. The notation $\mathcal{E}^{(q)}(x;\eta)$ in (1) means the value of the probability function at point $x \in \mathrm{A}$ for the probability distribution from the exponential family $\mathcal{E}^{(q)}$ with the canonic parameter $\eta \in \mathbb{R}^m$.

Lemma 1 ([1]). *Model* (1)–(3) *determines an ergodic Markov chain with strictly positive transition probabilities* (1) *and the unique nonsingular* s*-dimensional stationary probability distribution of* s*-tuples:*

$$\pi^{(s)} = (\pi^{(s)}_{q_1^s})_{q_1^s \in \mathrm{A}^s}, \quad \pi^{(s)}_{q_1^s} ::= \mathbf{P}\left\{\mathbf{x}_{t-s}^{t-1} = q_1^s\right\} > 0, \quad q_1^s \in \mathrm{A}^s, \quad t \in \mathbb{Z}. \tag{4}$$

The Fisher information matrix of MCSS(s) model (1)–(3) w.r.t. the parameter $\eta \in \mathbb{R}^m$ has the form:

$$\mathcal{I}(\eta) ::= \sum_{q \in \mathrm{A}^s} \pi_q^{(s)} \frac{d^2\phi^{(q)}(\eta)}{d\eta^2} \in \mathbb{R}^{m \times m}. \tag{5}$$

Note that MCSS(s) model (1)–(3) holds also for the case of a countable set A of possible values $\mathbf{x}_t$, i.e. for $N = |\mathrm{A}| = +\infty$, but the ergodicity conditions are more complex [2].

Let $\mathrm{LIND}_U\{\dots\}$ be the condition of linear independence of functions in the braces on their common support U. We will use the following regularity conditions.

- Regularity condition `R.0`:

$$\mathrm{LIND}_{\mathrm{A}^{s+1}}\{\mathfrak{h}_1,\dots,\mathfrak{h}_m, 1_{q'}, q' \in \mathrm{A}^s\}, \tag{6}$$

  where $1_{q'}^{(q)}(x) ::= \mathbb{1}\{q = q'\}$, $q, q' \in \mathrm{A}^s$, $x \in \mathrm{A}$, is the function on $\mathrm{A}^{s+1} = \{(q,x) : q \in \mathrm{A}^s, x \in \mathrm{A}\}$ indicating that the prehistory q equals to some fixed prehistory q'.

- Regularity condition `R.1`:

$$\exists q \in \mathrm{A}^s : \quad \mathrm{LIND}_{\mathrm{A}}\{1, \mathfrak{h}_1^{(q)}, \dots, \mathfrak{h}_m^{(q)}\},$$

  where $1(x) \equiv 1$, $x \in \mathrm{A}$.

The regularity condition `R.1` is stronger than `R.0`. Note that upper bound (3) for the number of parameters m is a necessary condition for `R.0`. Another necessary condition for `R.0` is that none of the base functions $\mathfrak{h}_i^{(q)}(x)$ are constant w.r.t. $x \in \mathrm{A}$.

Lemma 2 ([1]). *Under the regularity condition* `R.0` *model* (1) *is identifiable.*

According to Lemma 2, under the regularity condition R.0 Fisher information matrix (5) is strictly positively definite: $\mathcal{I}(\eta) \succ 0$.

In the general case the condition $\mathrm{LIND}_U\{g_1, \ldots, g_k\}$ for any $k \in \mathbb{N}$ functions $\{g_i : U \to \mathbb{R}\}_{i=1}^k$ on some finite set U, $k \leq |U|$, can be checked by checking the full rank property of the matrix of their values: $\mathrm{rank}(g_i(u))_{(i,u)\in\{1,\ldots,k\}\times U} = k$, which is a rather difficult problem for large $k = m + N^s$ and $|U| = |\mathrm{A}^{s+1}| = N^{s+1}$ in (6). However, a wide class of base functions $\{\mathfrak{h}_i\}$ guaranteed to satisfy R.0 can be constructed in a simple form with some additional features (see Section 3).

Theorem 1 ([1]). *Let $\mathbf{x}_1^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \in \mathrm{A}^T$ be an observed time series (1)–(3) of length $T \in \mathbb{N}$. The loglikelihood function for the* MCSS(s) *model has the form:*

$$L(\eta) = \left\langle \eta, \sum_{t=s+1}^{T} \mathfrak{h}^{(\mathbf{x}_{t-s}^{t-1})}(\mathbf{x}_t) \right\rangle - (T-s) \sum_{q\in\mathrm{A}^s} \hat{\pi}_q^{(s)} \phi^{(q)}(\eta) + \sum_{t=s+1}^{T} \mathfrak{h}_0^{(\mathbf{x}_{t-s}^{t-1})}(\mathbf{x}_t), \qquad (7)$$

$$\hat{\pi}_q^{(s)} = \frac{1}{T-s} \sum_{t=s+1}^{T} \mathbb{1}\left\{\mathbf{x}_{t-s}^{t-1} = q\right\}, \quad q \in \mathrm{A}^s.$$

Sufficient statistic for the MCSS(s) *model consists of two vectors* H *and* $\hat{\pi}^{(s)}$*:*

$$\mathrm{H} = (\mathrm{H}_i)_{i=1}^m ::= \sum_{t=s+1}^{T} \mathfrak{h}^{(\mathbf{x}_{t-s}^{t-1})}(\mathbf{x}_t) \in \mathbb{R}^m, \quad \hat{\pi}^{(s)} = (\hat{\pi}_q^{(s)})_{q\in\mathrm{A}^s} \in \mathbb{R}^{N^s}. \qquad (8)$$

Note that the first term of the loglikelihood (7) is linear w.r.t. the model parameter η, the second term (linear combination of $\phi^{(q)}(\eta)$) is nonlinear, and the third term does not depend on η. According to Theorem 1 all information on $(s+1)$-tuples of the observed time series $\{\mathbf{x}_t\}$ is gathered in the m-vector H of sufficient statistics (8), while the estimator $\hat{\pi}^{(s)}$ contains information on s-tuples only. Theorem 1 holds also for the case of countable Markov chain (1)–(3) with $N = |\mathrm{A}| = +\infty$.

2 Statistical parameter estimation for MCSS(s)

Let us construct the Maximum Likelihood Estimator (MLE) by the loglikelihood (7) of the observed time series $\mathbf{x}_1^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \in \mathrm{A}^T$ of length $T \in \mathbb{N}$:

$$\hat{\eta} = \arg\max_{\eta\in\mathbb{R}^m} L(\eta). \qquad (9)$$

Theorem 2 ([1]). *Under the regularity condition R.1 the loglikelihood (7) is asymptotically (almost surely for large enough T) strongly concave and has a unique global maximum MLE (9), that is asymptotically consistent, asymptotically normal and asymptotically efficient as $T \to +\infty$:*

$$\sqrt{T}\,(\hat{\eta} - \eta) \xrightarrow[T\to+\infty]{D} \mathcal{N}_m(0_m, \mathcal{I}^{-1}(\eta)), \qquad (10)$$

where the Fisher information matrix $\mathcal{I}(\eta)$ for the MCSS(s) *model is determined by (5).*

Due to asymptotical strong concavity of the loglikelihood (7), the MLE (9) can be computed by the gradient ascent algorithm:

$$\eta^{(k+1)} = \eta^{(k)} + \alpha_k \left.\frac{d}{d\eta} L(\eta)\right|_{\eta=\eta^{(k)}}, \quad k \in \mathbb{N}_0, \tag{11}$$

where $\alpha_k > 0$ is some step value for the k-th iteration. Due to the uniqueness property of the global maximum in (9), there is an arbitrariness in the choice of initial approximation $\eta^{(0)} \in \mathbb{R}^m$ in (11), so we can use zero initial vector $\eta^{(0)} ::= 0_m$ or some another initial statistical estimator $\eta^{(0)} ::= \tilde{\eta}$. The computational complexity of algorithm (11) is $\mathcal{O}\left(N \cdot \min\{T, N^{s+1}\}\right)$ [1]. Besides, this computational complexity depends on the computational accuracy, that is determined by the parameter $\varepsilon \ll 1$ of the stop condition for the gradient ascent algorithm (11): $L(\eta^{(k+1)}) - L(\eta^{(k)}) < \varepsilon$. After fulfilment of the stop condition algorithm (11) stops and returns the last iteration value $\hat{\eta}^{(\varepsilon)} ::= \eta^{(k+1)}$ with the asymptotics $\left\|\hat{\eta}^{(\varepsilon)} - \hat{\eta}\right\| \underset{\varepsilon\to 0}{=} \mathcal{O}\left(\sqrt{\varepsilon}\right)$ for the error of optimization process (9), (11).

3 Construction of base functions for MCSS(s)

Let us assume that A forms an abelian group w.r.t. some operation "+". We denote by $\mathbf{0} \in \mathrm{A}$ the neutral (zero) element of the group $(\mathrm{A}, +)$, and by $-a \in \mathrm{A}$ the inverse element for $a \in \mathrm{A}$. Introduce some auxiliary notation:

$$\mathfrak{J} ::= \{\mathrm{j} = (\mathrm{j}_i)_{i\in\mathbb{N}_0} \in \mathrm{A}^\infty, \mathrm{j}_0 \neq \mathbf{0}, \exists i_* : \mathrm{j}_i \equiv \mathbf{0}, i > i_*\} \subset \mathrm{A}^\infty \tag{12}$$

is the subset of infinite A-valued sequences with a zero tail and a nonzero first element (the set of indices for the base functions);

$$\mathfrak{J}(s) ::= \{\mathrm{j} = (\mathrm{j}_i)_{i\in\mathbb{N}_0} \in \mathfrak{J}, \mathrm{j}_i \equiv \mathbf{0}, i > s\}, \quad s \in \mathbb{N}_0; \tag{13}$$

is the subset of indices that we call s-indices ($\mathfrak{J}(s) \subset \mathfrak{J}(s')$, $s < s'$, $\cup_{s\in\mathbb{N}_0}\mathfrak{J}(s) = \mathfrak{J}$); $\Gamma ::= \{z \in \mathbb{C} : |z| = 1\}$ is the multiplicative group of all complex numbers $z \in \mathbb{C}$ with the absolute value $|z| = 1$ (the circle group); $\chi(\cdot,\cdot) : \mathrm{A} \times \mathrm{A} \to \Gamma$ is the function with the following properties (its existence follows from the character theory [3]):

- it is a group homomorphism $\mathrm{A} \to \Gamma$ called the character [3] under any one of two arguments fixed:

$$\chi(a, b + b') = \chi(a, b)\chi(a, b'), \quad \chi(b + b', a) = \chi(b, a)\chi(b', a), \quad \forall a, b, b' \in \mathrm{A};$$

- all the characters $\chi(a, \cdot)$ are distinct for $a \in \mathrm{A}$, i.e., $\forall a \neq a'$, $a, a' \in \mathrm{A}$, $\exists b \in \mathrm{A}$, $\chi(a, b) \neq \chi(a', b)$; similarly, all the characters $\chi(\cdot, a)$, $a \in \mathrm{A}$, are distinct;

$\mathrm{A}_{\mathrm{INV}} ::= \{a \in \mathrm{A}, a + a = \mathbf{0}\} \subset \mathrm{A}$ is the subgroup of involutive elements; $\sigma(\cdot) : \mathrm{A}\backslash\mathrm{A}_{\mathrm{INV}} \to \{\pm 1\}$ is any arbitrary odd function on the subset of non-involutive elements:

$\sigma(-a) = -\sigma(a)$, $a \in \mathrm{A} \setminus \mathrm{A}_{\mathrm{INV}}$; $\chi_*(\cdot,\cdot) : \mathrm{A} \times \mathrm{A} \to \mathbb{R}$ is the following real-valued modification of the complex-valued function $\chi(\cdot,\cdot)$:

$$\chi_*(a,b) ::= \begin{cases} \chi(a,b), & a \in \mathrm{A}_{\mathrm{INV}}, \\ \dfrac{\chi(a,b)+\chi(-a,b)}{\sqrt{2}}, & a \notin \mathrm{A}_{\mathrm{INV}}, \sigma(a) = +1, \\ \dfrac{\chi(a,b)-\chi(-a,b)}{\sqrt{2}}\sqrt{-1}, & a \notin \mathrm{A}_{\mathrm{INV}}, \sigma(a) = -1, \end{cases} \quad a,b \in \mathrm{A}. \tag{14}$$

Define the following set of harmonic base functions (harmonic set):

$$\mathfrak{h}_{\mathfrak{j}}^{(q_1^\infty)}(q_0) ::= \prod_{i=0}^{\infty} \chi_*(\mathfrak{j}_i, q_i), \quad \mathfrak{j} \in \mathfrak{J}, \quad q \in \mathrm{A}^\infty, \tag{15}$$

where the infinite product converges because of tail of unit factors by definition of the set $\mathfrak{J}$: $\exists i_*$, $\mathfrak{j}_i \equiv \mathbf{0}$, $i > i_*$, $\chi_*(\mathfrak{j}_i, q_i) \equiv \chi_*(\mathbf{0}, q_i) \equiv \chi(\mathbf{0}, q_i) \equiv 1$, $\chi(\mathbf{0},\cdot) \equiv 1$ is the principal character. For any $s \in \mathbb{N}_0$ and $\mathfrak{j} \in \mathfrak{J}(s)$ we can correctly use the function $\mathfrak{h}_{\mathfrak{j}}^{(q_1^\infty)}(q_0) : \mathrm{A}^\infty \to \mathbb{R}$ as the function of the form $\mathrm{A}^{s+1} \to \mathbb{R}$: $\mathfrak{h}_{\mathfrak{j}}^{(q_1^s)}(q_0) ::= \mathfrak{h}_{\mathfrak{j}}^{(q_1^\infty)}(q_0)$, $q_0^s \in \mathrm{A}^{s+1}$.

Let $G \subset \mathfrak{J}$, $|G| < +\infty$, be any finite subset of indices, $s_*(G) ::= \min\{s \in \mathbb{N}_0 : G \subset \mathfrak{J}(s)\}$ be its order, $\mathrm{MCSS}_*(G)$ be the model (1)–(3) of order $s = s_*(G)$ with zero supporting function $\mathfrak{h}_0^{(q)}(\cdot) \equiv 0$ and $m = |G|$ base functions $\mathfrak{h}_{\mathfrak{j}}$, $\mathfrak{j} \in G$.

Theorem 3 ([1]). *For any finite subset of indices $G \subset \mathfrak{J}$ the model $\mathrm{MCSS}_*(G)$ satisfies the regularity condition* `R.0` *and has identity Fisher information matrix* (5) *at the zero vector parameter $\eta = 0_m$: $\mathcal{I}(0_m) = \mathrm{Id}_m$. For any $s \in \mathbb{N}_0$ the model $\mathrm{MCSS}_*(\mathfrak{J}(s))$ is equivalent to the fully connected Markov chain MC(s) of order s.*

Harmonic base functions (15) can be used as an "elementary bricks" for the construction of parsimonious Markov models. These bricks may be taken sequentially from some previously fixed bigger set $G_* \subset \mathfrak{J}$ in a data-adaptive way by minimizing Bayesian information criterion (BIC) value $BIC(G) = |G| \cdot \ln T - 2L$ on each step, where L is the maximum loglikelihood value of the MCSS(s) model for the time series $\mathbf{x}_1^T$ ($L = L(\hat{\eta})$ in terms of (9)). The computational complexity of this data-adaptive model construction algorithm is $\mathcal{O}\left(|G_*|^2\right)$ operations. In particular, for $G_* = \mathfrak{J}(s)$ (construction of parsimonious Markov model with bounded order) $\mathcal{O}\left(|\mathfrak{J}(s)|^2\right) = \mathcal{O}\left(N^{2(s+1)}\right)$.

References

1. Kharin, Yu., Voloshko, V. (2025). Statistical Analysis of Parsimonious High-Order Multivariate Finite Markov Chains Based on Sufficient Statistics. *J. Multivariate Analysis*. Vol. **208**. Art. 105422.

2. Kemeny, J.G., Snell, J.L., Knapp, A.W. (1966). *Denumerable Markov Chains*. Van Nostrand: New York.

3. Luong, B. (2009). *Fourier Analysis on Finite Abelian Groups*. Birkhauser: Boston.

STATISTICAL CLASSIFICATION OF DISCRETE DATA BY THE SCDD PYTHON LIBRARY

Yu.S. Kharin[1], V.A. Voloshko[2], N.A. Prokhorchik[3]

[1,2,3]*Research Institute for Applied Problems of Mathematics and Informatics*
[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]kharin@bsu.by, [2]valoshka@bsu.by

We present a new Python package SCDD (for "Statistical Classification of Discrete Data") being developed in Research Institute for Applied Problems of Mathematics and Informatics of Belarusian State University. This package supports three types of discrete data: multivariate data, time series and random fields. For these types of discrete data the new developed by authors and known in the literature Markov models are implemented for statistical parameter estimation and data classification.

Keywords: statistical classification, discrete data, Python, Markov model

1 Structure of SCDD Package

Below we give a list of modules of a Python package SCDD (for "Statistical Classification of Discrete Data") and the implemented probabilistic models. Each module works with some type of discrete data $\mathbf{x}$ and classify it based on some m-parametric probabilistic model $M(\theta)$ (one of a models pool of this module) with parameter $\theta \in \mathbb{R}^m$. At the training stage an input training data is used for parameter estimation:

$$(\mathbf{x}^{(\mathrm{Train})}, M) \mapsto \hat{\theta}.$$

At the classification stage an input data is being classified based on model fitted on the training stage:

$$(\mathbf{x}^{(\mathrm{Classify})}, M(\hat{\theta})) \mapsto \mathrm{Class}.$$

Without loss of generality we assume that discrete data $\mathbf{x}$ of each type is represented by some indexed vectors (matrices, fields, etc.) $\mathbf{x} = (\mathbf{x}_i)_{i \in I}$ with some finite index set I and nonnegative integer entries $\mathbf{x}_i \in \mathbb{N}_0$.

1.1 Module "Visualization and preprocessing"

This module provides standard sample statistical data characteristics and their visualization to help user chose a suitable kind of probabilistic model for statistical classification:

- mutual covariances of variables;
- spectral characteristics;
- entropic functionals.

1.2 Module "Multivariate discrete data"

This module works with samples of i.i.d. multivariate discrete vectors $\mathbf{x}_t = (\mathbf{x}_{t,i})_{i=1}^d \in \mathbb{N}_0^d \sim M(\theta)$, $t \in \{1, \ldots, T\}$, where $M(\theta)$ is some multivariate discrete distribution family. The following standard families are implemented in the module:

- Multinomial;
- Negative Multinomial;
- Poisson;
- Hypergeometric;
- Logarithmic;
- Ewens;
- Based on power series.

1.3 Module "Univariate discrete time series"

This module works with univariate discrete time series $\mathbf{x}_t \in \mathbb{N}_0$, $t \in \{1, \ldots, T\}$, distributed according to some parsimonious high-order Markov model $M(\theta)$. The following new developed by authors and known in the literature Markov models are implemented in the module:

- Fully connected Markov chain MC(s);
- Markov chain of order s with r partial connections MC(s, r) [1];
- Raftery MTD model [2];
- Binary (MCSS [3], CNAR [4], MC);
- Conditionally binomial CNAR [5] (including outliers [6]);
- Conditionally semibinomial CNAR [7];
- Conditionally Poisson.

1.4 Module "Multivariate discrete time series"

This module works with multivariate discrete time series $\mathbf{x}_t \in \mathbb{N}_0^d$, $t \in \{1, \ldots, T\}$. The following models are implemented in the module:

- Conditionally Multinomial (CNAR [8]);
- Conditionally Negative Multinomial (CNAR [8]);
- Multinomial with conditionally independent Binomial components (CNAR [8]);

- Multinomial with conditionally independent Poisson components (CNAR);
- Multinomial binary with the standard base functions (MCSS [3]);
- Multinomial discrete with the standard base functions (MCSS [3]);
- Multinomial binary based on artificial neural networks (two-layer CNAR [9, 10]).

1.5 Module "Discrete random fields"

This module works with discrete random fields $\mathbf{x}_t \in \mathbb{N}_0$, $t \in V$, defined on some graph $G = (V, E)$ with vertices set V and edge set E. The following Markov Random Field (MRF) models are implemented in the module:

- Ising model on standard lattices (square, hexagonal, triangle);
- based on MRF on spanning tree of a graph (standard lattices and spanning trees);
- based on MRF on spanning tree of a graph (arbitrary user defined graph and spanning tree).

2 Implemented probabilistic models and their features

Here we briefly describe the used probabilistic models and their features.

The new recently developed by authors MCSS models [3] and CNAR models [4, 5, 6, 7, 8] for discrete time series provide flexible tools for data adaptive model extension with fast parameter statistical re-estimation algorithms.

The models of MRFs on graphs based on MRFs on graph's spanning trees allow to build a wide class of Markov models with explicit analytic expressions for statistical parameter estimation and to avoid a problem of exact solutions [11] for exponential families of probability measures. The standard Ising models with known exact analytic solutions are also implemented.

References

1. Kharin, Yu.S., Petlitskii, A.I. (2007). A Markov chain of order s with r partial connections and statistical inference on its parameters. *Discrete Mathematics and Applications*. Vol. **17**, No. **3**. P. 295–317.
2. Raftery, A., Tavare, S. (1994). Estimation and modelling repeated patterns in high order Markov chains with the Mixture Transition Distribution model. *J. Applied Statistics*. Vol. **43**, No. **1**. P. 179–199.

3. Kharin, Yu., Voloshko, V. (2025). Statistical analysis of parsimonious high-order multivariate finite Markov chains based on sufficient statistics. *J. Multivariate Analysis*. Vol. **208**. Art. 105422.

4. Kharin, Yu.S., Voloshko, V.A., Medved, E.A. (2018). Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series. *Mathematical Methods of Statistics*. Vol. **27**, No. **2**. P. 103–118.

5. Kharin, Yu.S., Voloshko, V.A. (2020). Statistical analysis of conditionally binomial nonlinear regression time series with discrete regressors. *Theory of Probability and Mathematical Statistics*. Vol. **100**. P. 181–190.

6. Kharin, Yu., Voloshko, V. (2021). Robust estimation for binomial conditionally nonlinear autoregressive time series based on multivariate conditional frequencies. *J. Multivariate Analysis*. Vol. **185**. Art. 104777.

7. Kharin, Yu.S., Voloshko, V.A. (2024). Semibinomial conditionally nonlinear autoregressive models of discrete random sequences: probabilistic properties and statistical parameter estimation. *Discrete Mathematics and Applications*. Vol. **30**, No. **6**. P. 417–437.

8. Voloshko, V.A., Kharin, Yu.S. (2022). Discrete-valued time series based on the exponential family with the multidimensional parameter and their probabilistic and statistical analysis. *Proc. National Academy of Sciences of Belarus. Physics and Mathematics series*. Vol. **58**, No. **3**. P. 280–291. (In Russian)

9. Fokianos, K. [et. al.] (2022). Statistical analysis of multivariate discrete-valued time series. *J. Multivariate Analysis*. Vol. **188**. Art. 104805.

10. Kharin, Yu.S., Voloshko, V.A. (2024). On the approximation of high-order binary Markov chains by parsimonious models. *Discrete Mathematics and Applications*. Vol. **34**, No. **2**. P. 71–87.

11. Baxter, R.J. (1982). *Exactly solved models in statistical mechanics*. Academic Press.

CONDITIONAL OPTIMIZATION IN UPLIFT MODELING

V.V. Kharlamov[1]

[1] *T-Bank, Applied Statistics Laboratory*
Moscow, RUSSIA
e-mail: [1]vi.v.kharlamov@gmail.com

The article is focused on a conditional optimization problem for uplift models with two given target metrics. This problem arises if we want to simultaneously maximize two metrics, for example, the customer happiness and the net profit. We present a method which maximizes the average value of one metric while the average value of another metric is fixed.
The difficulty of conditional optimization is that we need to estimate the average metric value for a policy proposed by the uplift model. We cannot use the predictions of the uplift model for this estimation. We present an effective algorithm that estimates the average metric value for an arbitrary policy based on the uplift model.

Keywords: uplift modeling, condilional optimization, model evaluation, randomized experiment

1 Introduction

Randomized experiments help people select an optimal treatment for the test population. However, customers can show significant heterogeneity in response to treatments. The problem of how to create the treatment policy, using the customer chatacteristics, is called an uplift modelling.

Classical uplift models are mazimize a single target metric. Unfortunately, we cannot limit ourselves with a single target metric. For example, we want to optimize metrics that exhibit inverse trends. The problem can then be naturally formulated as maximizing the average value of one metric while simultaneously fixing the average value of the other.

In this article, we develop an approach to solve the problem of conditional optimization in uplift modeling. We discuss how this approach performs on one of T-Bank's products and whether its assumptions are valid.

2 Conditional optimization

We use the mathematical formulation of uplift problem proposed in [1]. Let (X, T, Y^1, Y^2) be a random element, where the random vector $X \in \mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, the characterizes customer, the random integer $T \in \{0, \dots, K\}$ indicates the treatment, and the random variables $Y^1, Y^2 \in \mathbb{R}$ denote the values of target metrics.

The measurable function $h : \mathcal{X} \to \{0, \dots, K\}$ is called a *policy*. We denote by ν the set of all policies. The key performance metric of an uplift model is the expected

value of the response if the policy h is used to assign the treatment,

$$\mathbb{E}_h Y^j := \mathbb{E} f^j(X, h(X)), \quad f^j(x,t) = \mathbb{E}\left(Y^j \mid X = x, T = t\right), \quad j \in \{1,2\}.$$

There is no analytic experession for the function f^j, but we can build its approximation $\tilde{f}^j(x,t) \approx f^j(x,t)$, using machine learning algorithms. Set

$$h_w^*(x) := \arg\max_{t \le K} \left(f^1(x,t) + w f^2(x,t)\right), \quad x \in \mathcal{X},$$

where $w > 0$. The optimal expected response of $\mathbb{E}_h(Y^1 + wY^2)$ is achieved by the policy h_w^*.

Theorem 1. *Let $w > 0$ be a fixed weight. Set $C_w := \mathbb{E}_{h_w^*} Y^2$. Then the policy h_w^* is a solution of the conditional optimization problem*

$$\max_{h \in \nu} \mathbb{E}_h Y^1, \quad \mathbb{E}_h Y^2 \ge C_w.$$

We can use Theorem 1 to justify the following conditional optimization algorithm.

1. Train model $\tilde{f}^j(x,t) \approx f^j(x,t)$ for each $j \in \{1,2\}$.
2. Fix the boundary value C for the average value of the target metric Y^2.
3. Find $\tilde{h}_w^*(x) := \arg\max_{t \le K}(\tilde{f}^1(x,t) + w\tilde{f}^2(x,t))$ for each $w > 0$.
4. Estimate $\tilde{C}_w := \mathbb{E}_{\tilde{h}_w^*} Y^2$.
5. Find $w_0 > 0$ such that $\tilde{C}_{w_0} \approx C$.

The policy $\tilde{h}_{w_0}^*$ is the desired solution of the conditional optimization problem.

3 Estimate of the average value

Assume that we have the sample

$$(X_i, T_i, Y_i^1, Y_i^2), \quad i \in \{1, \ldots, N\},$$

of independent random elements, whose distribution coincides with the distribution of the random element (X, T, Y^1, Y^2). We can use an uplift model to estimate $\mathbb{E}_h Y^2$:

$$\mathbb{E}_h Y^2 \approx \sum_{i=1}^{N} \frac{\tilde{f}^2(X_i, h(X_i))}{N}.$$

However, this estimate depends on the quality of the uplift model. Most likely, we get a biased estimate.

We need the following assumption to build the unibiased estimate.

Assumption 1. For any $t \in \{0, \ldots, K\}$

$$\mathbb{P}(T = t \mid X) = p_t$$

a.s., where $p_t > 0$, $p_0 + \ldots + p_K = 1$.

In a randomized experiment Assumption 1 holds.

Theorem 2 (Theorem 2.1, [1])**.** *Let the random element (X, T) satisfies Assumption 1. We fix $h \in \nu$, $j \in \{1, 2\}$. Set*

$$Z^j_{h;i} := Y^j_i \frac{\mathbb{I}\{h(X_i) = T_i\}}{p_{h(X_i)}}, \quad i \in \mathbb{N}.$$

Then

$$\overline{Z}^j_h = \sum_{i=1}^{N} \frac{Z^j_{h;i}}{N}$$

is an unbiased and consistent estimate of $\mathbb{E}_h Y^j$.

Theorem 2 allows us to construct the estimate of $\mathbb{E}_h Y^2$ in step 4 of the above algorithm. However, we can use this algorithm to estimate $\mathbb{E}_h Y^1$. It allows us to evaluate the quality of the policy $\tilde{h}^*_w$.

4 Algorithm application

We applied the algorithm for optimization to one of T-Bank's products. This product is aimed at increasing a customer engagement in the ecosystem, but the bank spends a lot of money on this product. Therefore, we want to increase customer happiness, keeping the money to be spent on the product.

We applied the algorithm to the data from an earlier experiment, where $K = 3$, $N \propto 10^6$. The response Y^1 is the customer happiness, the response Y^2 is the income. Since the experiment is randomized, Assumption 1 is fulfilled.

To evaluate

$$\Delta f^j(x, t) := f^j(x, t) - f^j(x, 0),$$

we used the causal random forest [2] for each $j \in \{1, 2\}$, $t \in \{1, \ldots, K\}$. We validated the quality of the pairwise models $\Delta \tilde{f}^j(x, t) \approx \Delta f^j(x, t)$ using the AUUC [3].

Note that the curve based on the uplift model is higher than the curve for constant policies. We developed several policies for different exchanges between customer happiness and our income.

We conducted two experiments to test the quality of the choosen policies. The experimental results confirmed the applicability of the developed algorithm. We have managed to maximize one of the metrics by controlling the average of the other metric.

5 Discussion

In this article, we have developed and applied approach of conditional optimization in uplift modeling. The algorithm was theoretically validated and applied to a real-world business problem.

We want to continue developing the methodology. Firstly, the approach relies on Assumption 1. We want to solve the problem of conditional optimization with heterogeneous data or data from several experiments. Secondly, the approach is only effective when the number of treatments is small. We want to extend the approach to the case of multidimensional or continuous treatment. Thirdly, the approach produces a constant-time treatment. We want to develop a methodology that allows us to retrain and issue a time-varying policy. These improvements will allow us to expand an algorithm to more cases.

References

1. Zhao Y., Fang X., Simchi-Levi D. (2017). Uplift Modeling with Multiple Treatments and General Response Types *Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics* pp. 588-596.

2. Athey S., Imbens G. (2016). Recursive partitioning for heterogeneous causal effects *Proceedings of the National Academy of Sciences* Vol. **113**, Num. **27**, pp. 7353-7360.

3. Gutierrez P., Grardy J.Y. (2017). Causal inference and uplift modelling: A review of the literature *International conference on predictive applications and APIs. PMLR.* pp. 1-13.

RATIONAL-INFINITE DIVISIBILITY OF MIXTURE PROBABILITY LAWS WITH DOMINATED CONTINUOUS SINGULAR PARTS

A.A. Khartov[1]

[1]*Institute for Information Transmission Problems of Russian Academy of Sciences*
[1]*Smolensk State University*
[1]*ITMO University*
Moscow, Smolensk, Saint Petersburg, RUSSIA
e-mail: [1]khartov.a@iitp.ru, [1]alexeykhartov@gmail.com

We consider a new class $\boldsymbol{Q}$ of distribution functions F that have the property of rational-infinite divisibility: there exist some infinitely divisible distribution functions F_1 and F_2 such that $F_1 = F * F_2$. Characteristic functions of such probability laws admit the Lévy-type representation with "signed spectral measures". We propose criteria for a distribution function F to belong to the class $\boldsymbol{Q}$ for the unexplored case, where F may have a continuous singular part.

Keywords: infinite divisibility, rational-infinite divisibility, quasi-infinite divisibility, the Lévy-type representation, continuous singular part

1 Introduction

Let $\boldsymbol{I}$ denote the class of all infinitely divisible distribution functions on the real line. This class is naturally extended by the following way. We call a distribution function F *rational-infinitely divisible* if there exist some infinitely divisible distribution functions F_1 and F_2 such that $F_1 = F * F_2$. In terms of characteristic functions, this definition is equivalent to the formula $f(t) = f_1(t)/f_2(t)$, $t \in \mathbb{R}$, for the characteristic function f of F, where f_1 and f_2 are the characteristic functions of some infinitely divisible distribution functions F_1 and F_2, respectively. We denote by $\boldsymbol{Q}$ the class of all rational-infinitely divisible distribution functions. Since F_2 may be degenerate, it is seen that $\boldsymbol{I} \subset \boldsymbol{Q}$. The class $\boldsymbol{Q}$ coincides with the class of *quasi-infinitely divisible distribution functions*, in which the characteristic function f of any representative F admits *the Lévy-type representation*:

$$f(t) = \exp\biggl\{ it\gamma - \frac{\sigma^2 t^2}{2} + \int_{\mathbb{R}\setminus\{0\}} \bigl(e^{itx} - 1 - it\sin(x)\bigr) dL(x) \biggr\}, \quad t \in \mathbb{R},$$

with some *shift parameter* $\gamma \in \mathbb{R}$, *the Gaussian variance* $\sigma^2 \geqslant 0$, and *the Lévy-type spectral function* $L : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, which has a bounded total variation on $\mathbb{R} \setminus O_\delta$ for every $\delta > 0$, where $O_\delta := (-\delta, 0) \cup (0, \delta)$, and, in general, it is non-monototic on the intervals $(-\infty, 0)$ and $(0, +\infty)$. The function L also satisfies the conditions $L(\pm\infty) = 0$ and

$$\int_{O_\delta} x^2 d|L|(x) < \infty \quad \text{for any} \quad \delta > 0.$$

The function L is assumed to be right-continuous at every point of the real line. It is important to note that *the characteristic triplet* (γ, σ^2, L) is uniquely determined by f and hence by F.

The class $\boldsymbol{Q}$ and its multivariate analog are actively studied now (see [2] and [6]) and they find some interesting applications in probability limit and compactness theorems (see [1] and [5]), and in other areas. This class is seen to be rather wide. It is interesting and important to obtain criteria for belonging to it. There are finished results in [1, 6] for the discrete distributions and in [3, 4] for the mixtures of discrete and absolutely continuous laws. In this note, we propose a criterion, which generalizes these results for the case, when F may have a continuous singular part.

2 Results

Let F be an arbitrary distribution function on the real line. According to the Lebesgue decomposition theorem, F admits the representation:

$$F(x) = c_d F_d(x) + c_a F_a(x) + c_s F_s(x), \quad x \in \mathbb{R}, \tag{1}$$

where F_d, F_a, and F_s are discrete, absolutely continuous and continuous singular distribution functions, respectively. Here the coefficients c_a, c_a, and c_s are non-negative constants such that $c_d + c_a + c_s = 1$. Let f be the characteristic function of F. It is represented in the similar way:

$$f(t) = c_d f_d(t) + c_a f_a(t) + c_s f_s(t), \quad t \in \mathbb{R},$$

where f_d, f_a, and f_s are the characteristic functions corresponding to F_d, F_a, and F_s, respectively.

We consider only the case, when F has a non-zero discrete part, i.e. $c_d > 0$ in (1). Let

$$F_d(x) = \sum_{\substack{k \in \mathbb{N}_0: \\ x_k \leqslant x}} p_k, \quad x \in \mathbb{R},$$

where x_k are distinct reals associated with weights $p_k \geqslant 0$, $k \in \mathbb{N}_0$, $\sum_{k=0}^{\infty} p_k = 1$. We define the support of the distribution corresponding to F_d,

$$\mathcal{X} := \{x_k : p_k > 0, k \in \mathbb{N}_0\},$$

and the set of all finite $\mathbb{Z}$-linear combinations of elements from the set $\mathcal{X}$,

$$\langle \mathcal{X} \rangle := \left\{ \sum_{k=1}^{m} a_k z_k : \, a_k \in \mathbb{Z}, \, z_k \in \mathcal{X}, \, m \in \mathbb{N} \right\}.$$

Let us formulate the main result. For convenience, we preliminarily select the following property of distributions. Let $\mu_d := \inf_{t \in \mathbb{R}} |f_d(t)|$. We say that a distribution function F has *the dominated continuous singular part* if $c_s < c_d \mu_d$ for the case $\mu_d > 0$ and if $c_s = 0$ for the case $\mu_d = 0$.

Theorem 1. *Suppose that F has decomposition (1) with some $c_d > 0$, $c_a \geqslant 0$, $c_s \geqslant 0$, and F has the dominated continuous singular part. Then the following statements are equivalent:*

(i) $F \in \boldsymbol{Q}$,

(ii) $\inf_{t\in\mathbb{R}} |f(t)| > 0$,

(iii) $f(t) \neq 0$ *for any* $t \in \mathbb{R}$, *and* $\inf_{t\in\mathbb{R}} |f_d(t)| > 0$.

If one of the conditions is satisfied, and hence all, then f admits the following representation

$$f(t) = \exp\Biggl\{ it\gamma_0 + \sum_{u\in\langle\mathcal{X}\rangle\setminus\{0\}} \lambda_u\bigl(e^{itu}-1\bigr) + \int_{\mathbb{R}\setminus\{0\}} (e^{itx}-1)\biggl(v_a(x) + \operatorname{sign}(x)\,\frac{\mathfrak{m}_a\cdot e^{-|x|}}{|x|}\biggr)dx + \int_{\mathbb{R}\setminus\{0\}} (e^{itx}-1)\,dW(x)\Biggr\}, \quad t\in\mathbb{R}.$$

Here $\gamma_0 \in \langle\mathcal{X}\rangle$, $\lambda_u \in \mathbb{R}$ for all $u \in \langle\mathcal{X}\rangle \setminus \{0\}$, and $\sum_{u\in\langle\mathcal{X}\rangle\setminus\{0\}} |\lambda_u| < \infty$. Next, the function $v_a : \mathbb{R} \mapsto \mathbb{R}$ satisfies $\int_{\mathbb{R}} |v_a(x)|dx < \infty$, and, in the case $c_a = 0$, v_a is identically 0; the constant $\mathfrak{m}_a$ is an integer and $\mathfrak{m}_a = 0$ for the case $c_a = 0$. Next, the function $W : \mathbb{R} \to \mathbb{R}$ has a bounded total variation and it is always continuous on $\mathbb{R}$. If $c_s = 0$ then W is identically 0. If $c_s \neq 0$ then W is not absolutely continuous on $\mathbb{R}$, i.e. it always contains some continuous singular part. In addition, if all the functions F_s^{*k}, $k \in \mathbb{N}$, are continuous singular, then the function W is (pure) continuous singular.

Using Theorem 1, it is easy to construct a lot of particular examples of $F \in \boldsymbol{Q}$ with non-zero continuous singular parts. For instance,

$$F(x) := c_d\mathbb{1}_0(x) + c_aF_a(x) + c_sF_s(x), \quad x\in\mathbb{R},$$

with $c_d > c_s > 0$, $c_a \geqslant 0$, and $c_d + c_a + c_s = 1$. Here $\mathbb{1}_0$ denotes the distribution function of the degenerate law concentrated at the point $x = 0$. Let F_s be an arbitrary continuous singular function, but F_a be an absolutely continuous distribution function, whose characteristic function f_a is real and non-negative (for instance, f_a is a Pólya-type characteristic function or f_a is corresponded to a symmetric continuous stable distribution). Then it is not difficult to check *(ii)* and conclude that $F \in \boldsymbol{Q}$.

It should be noted that the condition of the dominated singular part can not be simply omitted and, moreover, it cannot be extended to the case $c_s = c_d\mu_d$ with $\mu_d > 0$ without certain additional assumptions.

References

1. Alexeev I.A., Khartov A.A. (2023). Spectral representations of characteristic functions of discrete probability laws. *Bernoulli* Vol. **29**, Num. **2**, pp. 1392–1409.

2. Berger D., Kutlu M., Linder A. (2021). On multivariate quasi-infinitely divisible distributions. *A Lifetime of Excersions Through Random Walks and Lévy Processes. A Volume in Honour of Ron Doney's 80th Birthday*. L. Chaumont, A.E. Kyprianou (eds.), Progress in Probability **78**, Birkhäuser, pp. 87–120.

3. Berger D. (2019). On quasi-infinitely divisible distributions with a point mass. *Math. Nachr.* Vol. **292**, pp. 1674–1684.

4. Berger D., Kutlu M. (2023). Quasi-infinite divisibility of a class of distributions with discrete part. *Proc. Amer. Math. Soc.* Vol. **151**, Num. **5**, pp. 2211–2224.

5. Khartov A.A. (2023). On weak convergence of quasi-infinitely divisible laws. *Pacific J. Math.* Vol. **322**, Num. **2**, pp. 341–367.

6. Lindner A., Pan L., Sato K. (2018). On quasi-infinitely divisible distributions. *Trans. Amer. Math. Soc.* Vol. **370**, pp. 8483–8520.

DISCRETIZATION OF DATA THAT DO NOT CHANGE THE LIMIT DISTRIBUTION OF ASYMPTOTICALLY NORMAL STATISTICS

E.V. Khil[1], A.V. Shklyaev[2]

[1,2]*Lomonosov Moscow State University*
Moscow, RUSSIA
e-mail: [1]elena.khil@math.msu.ru, [2]alexander.shklyaev@math.msu.ru

We show that the specific discretization of data doesn't change the limit distribution of a wide class of statistics. More specifically, we group the data in $o(\sqrt{n})$ bundles, where n is the sample size. We show that the limit distribution of a wide class of functionals of the empirical cumulative distribution function remains the same after discretization.

Keywords: discretization, asymptotical statistical tests, Hadamard differentiability, empirical process

1 Introduction

Let $X_1, \ldots, X_n \in \mathbb{R}$ be a sample of one-dimensional random variables.

Nowadays statisticians often work with big data. Many statistical tests are based on quite slow algorithms ($O(n^2)$ and more operations). As an example one can consider the adaptive chi-square test, proposed in [1] or MMD test (see [2]). This test is very slow but at the same time very powerful.

That is why one of the most important problems is to decrease the sample size without significant loss of power.

We propose a natural approach for an i.i.d. one-dimensional sample – let group the adjacent observations and construct a sample of the size $k = k(n)$. It's easy to see that for $k(n)/\sqrt{n} \to +\infty$, $n \to \infty$, the weak limit of the empirical process is still the Brownian bridge. Therefore, discretization doesn't change the asymptotical distribution for a wide class of test statistics.

2 Results

Significant part of asymptotic statistics is based on convergence theorems for empirical processes, particularly on the fact that $\sqrt{n}(\overline{f(X)} - \mathbf{E}_F f(X))$ converges in distribution to a tight gaussian process Y_f for different classes $f \in \mathcal{F}$ (see [3]). The multiplier $\sqrt{n}$ gives an idea to discretize data by replacing $o(\sqrt{n})$ adjacent elements with one value, representing the whole group. In this case the functional limit theorems still holds.

Let $\hat{F}$ be the empirical cumulative distribution function of the sample $X_1, \ldots, X_n$ and $X_{(i)}$, $i \leq n$, – the order statistics of the sample.

We fix some k and compute $l = [n/k]$, $r = n - kl$. Let

$$Y_i = \begin{cases} \text{MED}\left(X_{((i-1)(l+1)+j)}, j \le l+1\right), & i \le r, \\ \text{MED}\left(X_{(r+(i-1)l+j)}, j \le l\right), & i \in (r,k). \end{cases} \tag{1}$$

Let

$$(Z_1, Z_2, \ldots, Z_n) = (Y_1, Y_1, \ldots, Y_1, Y_2, Y_2, \ldots, Y_2, \ldots, Y_k, Y_k, \ldots, Y_k),$$

where $Y_1, \ldots, Y_r$ are repeated $l+1$ times and $Y_{r+1}, \ldots, Y_k$ are repeated l times.

Definition 1. The sample $(Z_1, Z_2, \ldots, Z_n)$ is called the Discretized sample.

Let $\hat{F}^{(k)}$ be the empirical cumulative distribution function of the Discretized sample.

We consider the class $\mathcal{F}$ of Hadamard differentiable functionals (and some other classes too) and show that $f(\hat{F}^{(k)})$ has the same asymptotical distribution as $f(\hat{F})$ for $f \in \mathcal{F}$.

We use the synthetic data to show (empirically) that our $o(\sqrt{n})$-discretization does not decrease the power of asymptotical statistical tests, does not change the asymptotical distribution of tests statistics and significantly accelerates the work of these tests. Our empirical experience propose to use $O(n^{1/4})$-discretization for the sample size $n \approx 10^3$–10^4 and $O(n^{1/3})$-discretization for n more than 10^6. In this case the computational power decreases from $O(n^\alpha)$ to $O(n^{3\alpha/4})$ and $O(n^{2\alpha/3})$ correspondingly.

In our report we will describe a number of applications of presented approach to different statistical problems, show the results of computer modeling and discuss the further development of the method.

References

1. Heller R. [et. al.] (2016). Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Machine Learning Research*. Vol. **17**, No. **29**. P. 1–54.

2. Gretton A. [et. al.] (2012). A kernel two-sample test. *J. Machine Learning Research*. Vol. **13**, No. **1**. P. 723–773.

3. Shorack G., Wellner J. (2009). *Empirical processes with applications to statistics*. Wiley & Sons.

ON PROBABILITY DISTRIBUTION ASSOCIATED BY TODA CHAIN

M.K. Khomidov[1]

[1] *University of Exact and Social Sciences*
Tashkent, UZBEKISTAN
e-mail: [1]mkhomidov0306@mail.ru

We study Gibbs distributions associated with the Toda chain. We consider the following Hamiltonian defined on the phase space $\mathbb{R}^{2(m+l+1)}$:

$$H_{m,l} = \sum_{j=-m}^{l} \frac{p_j^2}{2} + \sum_{k=-m}^{l-1} e^{q_{k+1}-q_k}.$$

It is well known chat the discrete Toda chain is one of classical completely integrable models. We prove that for fixed values of parameters β and μ the limit Gibbs distribution defined by Hamiltonian H there exists. We study the probability properties of limit Gibbs distribution.

Keywords: Gibbs distribution, Toda chain, completely integrable, Lax pair

1 Introduction

The main goal of this work is to study the limit Gibbs distribution of Toda Chain and their probability properties.

The Toda chain (see [5]) is a model consisting of a chain of particles with nearest-neighbor interactions, described by the Hamiltonian $H_{m,l}$ and the corresponding equations of motion:

$$\begin{cases} \dfrac{dq_k}{dt} = \dfrac{\partial H_{m,l}}{\partial p_k} = p_k, & -m \le k \le l, \\ \dfrac{dp_j}{dt} = -\dfrac{\partial H_{m,l}}{\partial q_j} = -e^{q_j - q_{j-1}} + e^{q_{j+1}-q_j}, & -m \le j \le l. \end{cases}$$

Here, $q_j(t)$ denotes the displacement of the j-th particle from its equilibrium position, and $p_j(t)$ is its momentum (assuming unit mass, i.e., $m = 1$).

Let us introduce new variables:

$$a_j = -\frac{p_j}{2}, \qquad -m \le j \le l,$$

$$b_k = \frac{1}{2} e^{\frac{q_{k+1}-q_k}{2}}, \qquad -m \le k < l,$$

which are called the **Flaschka variables**.

The Hamiltonian system is equivalent to the matrix equation

$$\dot{L} = [L, A] := AL - LA,$$

where

$$L = \begin{bmatrix} a_1 & b_1 & 0 & \cdots & 0 \\ b_1 & a_2 & b_2 & \ddots & \vdots \\ 0 & b_2 & a_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & 0 & b_{n-1} & a_n \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & b_1 & 0 & \cdots & 0 \\ -b_1 & 0 & b_2 & \ddots & \vdots \\ 0 & -b_2 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & 0 & -b_{n-1} & 0 \end{bmatrix}.$$

It is well known that the functions $F_k = \operatorname{Tr} L^k$, for $k = 1, \ldots, m + l + 1$, are the first integrals of the Toda chain; that is, each F_k is a constant of motion:

$$F_k = \text{const}.$$

It is clear that

$$F_1 = \operatorname{Tr} L = \sum_{j=-m}^{l} a_j = -\frac{1}{2}\sum_{j=1}^{n} p_j,$$

$$F_2 = \operatorname{Tr} L^2 = \sum_{j=1}^{n} a_j^2 + 2\sum_{j=1}^{n-1} b_j^2 = \frac{1}{4}\sum_{j=1}^{n} p_j^2 + \frac{1}{2}\sum_{j=1}^{n-1} e^{q_{j+1}-q_j},$$

$$F_3 = \operatorname{Tr} L^3 = -\frac{1}{8}\sum_{j=1}^{n} p_j^3 + \frac{3}{16}\sum_{j=1}^{n-1}(p_j + p_{j+1})e^{q_{j+1}-q_j},$$

$$\begin{aligned} F_4 = \operatorname{Tr} L^4 = = \frac{1}{16}\sum_{j=1}^{n} p_j^4 + \sum_{j=1}^{n-1}\left(\frac{1}{4}(p_j^2 + p_{j+1}^2) + \frac{1}{4}p_j p_{j+1}\right) e^{q_{j+1}-q_j} \\ + \frac{1}{8}\sum_{j=1}^{n-1} e^{2(q_{j+1}-q_j)} + \frac{1}{4}\sum_{j=1}^{n-2} e^{q_{j+1}-q_j} e^{q_{j+2}-q_{j+1}}. \end{aligned}$$

It is well known that any linear combination of the first integrals F_1, F_2, F_3, and F_4 is also a first integral. In what follows, we are particularly interested in studying linear combinations of the form

$$\tilde{H}_{m,l} = F_4 + J_3 \cdot F_3 + J_2 \cdot F_2 + J_1 \cdot F_1,$$

where the parameters J_1, J_2, $J_3 \in \mathbb{R}$. In the last expression, if m and l tend to infinity, we obtain an infinite-dimensional Hamiltonian $\tilde{H}$.

2 Main part

The configuration model with Hamiltonian $\tilde{H}$ is defined on countable, locally finite subsets $X \subset \mathbb{R}^1$, where for each $q \in X$, its right and left "neighbors" q^r and q^l, respectively, are defined. It is not required that $q^l < q < q^r$; however, it is required that:

(i) $(q^l)^r = q = (q^r)^l$;
(ii) the natural graph with edges $q \to q^l$ and $q \to q^r$ is connected;
(iii) if an ordering $q < q^r$ is introduced on X, then

$$\lim_{m\to\infty} q_{-m} = +\infty, \qquad \lim_{l\to\infty} q_l = -\infty.$$

We denote the space of all configurations X by Ω. Consider the mapping $\tau : X \to \mathbb{R}^2$, given by $\tau(q) = (q, p) \in \mathbb{R}^2$ for any $q \in X$. A point in the phase space is defined as the set $Y = \tau(X)$, together with the graph induced by X.

The phase space is defined as the set

$$M = \bigcup_{X\in\Omega} \bigcup Y,$$

We need the following definitions.

Definition 1. The conditional Gibbs distribution at inverse temperature $\beta > 0$, given the boundary conditions Y^l and Y^r, is a probability distribution on the set $M(Y^l, Y^r)$, such that for $k \geq 0$, its restriction to $M_k(Y^l, Y^r)$ has the density

$$P_{m,l}(q_{-m}, p_{-m}, \ldots, q_l, p_l) = \Xi^{-1}_{\beta,\mu}(s) \cdot \exp\left\{-\beta\left(\tilde{H}_{m+2,l+2}(s) + \mu k\right)\right\}, \tag{1}$$

where $(q_{-m-1}, p_{-m-1}) = z_1$, $(q_{-m-2}, p_{-m-2}) = z_1^l$, $(q_{l+1}, p_{l+1}) = z_2$, and $(q_{l+2}, p_{l+2}) = z_2^r$. Here, m and l denote the numbers of particles to the left and right of the zero particle, respectively, so that $m + l + 1 = k$. The term $\Xi_{\beta,\mu}(s)$ is the normalizing factor (partition function).

Definition 2. The limiting Gibbs distribution $\nu_{\beta,\mu}$, for $\beta > 0$ and $\mu \in \mathbb{R}$, is a probability measure on γ such that, for any set

$$M(\Delta_{-2}, \Delta_{-1}, \Delta_0, \Delta_1, \Delta_2),$$

its induced conditional distribution on the σ-subalgebra

$$\gamma^t(\Delta_{-2}, \Delta_{-1}, \Delta_0, \Delta_1, \Delta_2)$$

coincides almost surely (a.s.) with the distribution defined by (1).

Now we formulate our main results.

Theorem 1. Let the parameters $\beta > 0$ and $\mu \in \mathbb{R}$ satisfy the following condition

$$e^{\beta\mu} \int \exp\{-\beta p^4\}\, dp \int \exp\{-\beta(e^{2y} - y)\}\, dy < 1.$$

Then, for the probability measures with density (1), as $s \to +\infty$, there exists at least one limitic Gibbs measure $\nu_{\beta,\mu}$ on the phase space M.

We define the sequence $(z_k)_{k\in\mathbb{Z}}$ by

$$z_k := q_{k+1} - q_k.$$

Let $\sum$ denote the **symbolic space**, defined as:

$$\sum := \left\{\underline{y} = (\dots, y_{i-1}, y_i, y_{i+1}, \dots) \,:\, y_i \in \mathbb{R},\, i \in \mathbb{Z}\right\} =: \mathbb{R}^{\mathbb{Z}}.$$

Let $\tau : \Sigma \to \Sigma$ be the **shift map**, defined as

$$(\tau(\underline{y}))_i = y_{i+1}, \quad i \in \mathbb{Z}.$$

For any lattice interval $\Lambda := [m, k] \subset \mathbb{Z}$, we define a **cylinder set**, or simply a **cylinder**, as follows:

$$C[m,k] := \left\{\underline{y} :\ \underline{y} = (\dots, y_{i-1}, y_i, y_{i+1}, \dots),\ y_i \in [a_i, b_i],\ m \le i \le k\right\},$$

where $\mathbb{B}(\mathbb{R})$ is the Borel σ-algebra of subsets of the real line $\mathbb{R}$.

Let $\mathbb{B}(\mathbb{R}^{\mathbb{Z}})$ denote the smallest σ-algebra containing all possible cylinder sets.

Theorem 2. For any cylinder C defined on an lattice interval $\Lambda \subset \mathbb{Z}$ hold the following inequalities:

$$c_1\theta_1^{|\Lambda|} \le \nu_{\beta,\mu}(C) \le c_2\theta_2^{|\Lambda|}$$

where c_1, $c_2 > 0$ and θ_1, $\theta_2 \in (0,\, 1)$ only depend on the Gibbs measure $\nu_{\beta,\mu}$.

Theorem 3. $C_1 \subset C$ be cylinders defined on intervals Λ_1, Λ; then

$$c_1\theta_1^{|\Lambda_1|-|\Lambda|} \le \frac{\nu_{\beta,\mu}(C_1)}{\nu_{\beta,\mu}(C)} \le c_2\theta_2^{|\Lambda_1|-|\Lambda|}.$$

Remark 1. Analogues results are true for other integrals of motion of Toda Chain.

References

1. Sarig O. (1999). Thermodynamic formalism for countable Markov shifts. *Ergodic Theory and Dynamical Systems*. Vol. **19**, Num. **6**, pp. 1565-1593.

2. Van Morbeke P. (1976). The spectrum of Jacobi Matrices. *Inventions math.* Num. **37**, pp. 45-81.

3. Ruelle D. (2004). *Thermodynamic Formalism*. Cambridge: Cambridge University Press, second edition.

4. Herbert S. (2019). Generalized Gibbs ensembles of the classical Toda chain. *arXiv:190207751v4*.

5. Manakov S. V. (1974). O polnoy integriruyemosti i stoxastizattsii v diskretnix sistemax *Journal of Experimental and Theoretical Physics*, Vol. **67**, Num. **2**. pp. 543-555.

TENSORS FOR SIGNAL AND FREQUENCY ESTIMATION IN SUBSPACE-BASED METHODS: WHEN THEY ARE USEFUL?

N.A. Khromov[1], N.E. Golyandina[2]

[1,2]*Saint Petersburg State University*
Saint Petersburg, RUSSIA
e-mail: [1]hromovn0@gmail.com, [2]n.golyandina@spbu.ru

Tensor modifications of singular spectrum analysis for signal extraction and frequency estimation problems in a noisy sum of exponentially modulated sinusoids are reviewed. Modifications using Higher-Order SVD are considered. Numerical comparisons are carried out. It is shown numerically that for, the signal extraction problem, tensor methods generally perform worse than matrix methods for a single-channel series, but can outperform multi-channel SSA for a series system. For frequency estimation, tensor modifications are generally advantageous.

Keywords: time series, signal, frequency estimation, tensor, singular spectrum analysis

1 Introduction

Singular spectrum analysis (SSA) is one of the methods used for time series analysis [2], in which the original time series is transformed into a matrix, called the trajectory matrix, using a given window length L. The singular value decomposition (SVD) of this matrix is then analyzed. When the objective is to estimate the signal and its properties from an observed noisy series, the first r components of the SVD are considered, where r is the rank of the signal trajectory matrix. Based on the selected components, the signal estimation is constructed. A distinctive feature of the method is that it does not require the specification of a signal model. However, SSA can also handle a parametric signal model in the form of a sum of products of polynomials, exponentials and sinusoids. The frequency estimation problem plays a special role. The ESPRIT method uses the estimation of the signal subspace based on the r leading left singular vectors of the trajectory matrix SVD to estimate the frequencies present in the signal. The least squares (LS) version of ESPRIT [5] is also known as Hankel SVD (HSVD), and the total least squares (TLS) version is known as HTLS [7].

A number of works propose tensor modifications of the SSA and ESPRIT methods, where the original series is transformed into a tensor, usually of 3rd order, instead of a matrix [1, 3, 6]. One of the common variants of tensor decompositions is Higher-Order SVD (HO-SVD), which generalizes the matrix SVD.

This work aims to compare the performance of matrix and tensor modifications of SSA in solving signal extraction and frequency estimation problems. We will consider the tensor modifications proposed in [3] and [4], which have been adapted for signal extraction.

2 Methods description

2.1 Tensor SSA algorithm layout for signal extraction

The general structure of tensor SSA algorithms based on HO-SVD is as follows (Basic SSA is a special case). Let X be the observed object. The tensor dimensions I, L and K are considered as the window length; some of these dimensions are expressed in terms of the others, or are fixed. The parameters of the algorithm are the values R_1, R_2 and R_3. These are often chosen to be equal to r, but not always.

1. Embedding to the trajectory tensor $\mathbf{X} = \mathcal{T}(\mathsf{X})$.
2. Tensor decomposition $\mathbf{X} = \sum_{i=1}^{I}\sum_{l=1}^{J}\sum_{k=1}^{K} \mathcal{Z}_{ilk} U_i^{(1)} \circ U_l^{(2)} \circ U_k^{(3)}$.
3. Grouping $\hat{\mathbf{X}} = \sum_{i=1}^{R_1}\sum_{l=1}^{R_2}\sum_{k=1}^{R_3} \mathcal{Z}_{ilk} U_i^{(1)} \circ U_l^{(2)} \circ U_k^{(3)}$.
4. Obtaining from $\hat{\mathbf{X}}$ the signal estimate $\hat{\mathsf{X}}$ based on the structure of the trajectory tensor and the operation that is inversed to embedding.

We will further consider two types of input object: single-channel and multi-channel time series.

2.2 Trajectory tensors

Let $\mathsf{X} = (x_1, \ldots, x_N)$ be a single-channel time series of length N, $x_n \in \mathbb{C}$.

Definition 1. The tensor embedding operator for a single-channel time series with window lengths I and L (then $K = N - I - L + 2$) such that $1 < I, L < N$, $I + L < N + 1$ is a mapping $\mathcal{T}_{I,L}$ that transfers the series X into the tensor $\mathcal{X} \in \mathbb{C}^{I \times L \times K}$ as follows: $\mathcal{X}_{ilk} = x_{i+l+k-2}$, where $i \in \overline{1:I}$, $l \in \overline{1:L}$, $k \in \overline{1:K}$.

Let $\mathsf{X} = \left(\mathsf{X}^{(1)}, \ldots, \mathsf{X}^{(P)}\right)$ be a multi-channel time series consisting of P single-channel series, also called channels.

Definition 2. The tensor embedding operator for a multi-channel time series with window length L (that is, $I = P$, L, $K = N - L + 1$) such that $1 < L < N$ is a mapping $\mathcal{T}_L$ that translates the P-channel time-series X into the tensor $\mathcal{X} \in \mathbb{C}^{P \times L \times K}$ as follows $\mathcal{X}_{plk} = x_{l+k-1}^{(p)}$, where $p \in \overline{1:P}$, $l \in \overline{1:L}$, $k \in \overline{1:K}$.

2.3 Algorithm for signal parameters estimation

Consider the P-channel time series (including the single-channel case $P = 1$) with elements

$$x_n^{(p)} = \sum_{r=1}^{R} a_r^{(p)} e^{\alpha_r n} e^{\mathrm{i}\left(2\pi\omega_r n + \varphi_r^{(p)}\right)},$$

where the model parameters are the amplitudes $a_r^{(p)} \in \mathbb{C}\setminus\{0\}$, phases $\varphi_r^{(p)} \in [0, 2\pi)$, the frequencies $\omega_r \in [0, 1/2]$, and the damping factors $\alpha_r \in \mathbb{R}$. The HO-ESPRIT algorithm that estimates the frequencies and damping factors of a time series is defined as follows. After the embedding step the matrix $\mathbf{U} = \mathbf{U}_d = \left[U_1^{(d)} : U_2^{(d)} : \ldots : U_{R_d}^{(d)}\right]$ for $d \in \{1, 2, 3\}$ is constructed and the following matrix equation

$$\mathbf{U}^{\uparrow} = \mathbf{U}_{\downarrow}\mathbf{Z}$$

is solved with respect to matrix $\mathbf{Z}$, where the up and down arrows placed behind the matrix $\mathbf{U}$ stand for deleting its first and last rows accordingly. The R largest eigenvalues of the matrix $\mathbf{Z}$ are considered to be the estimates of the poles $\lambda_r = e^{\alpha_r + 2\pi \mathrm{i}\omega_r}$, from which the parameters α_r and ω_r can be obtained.

2.4 Dstack modifications

In the paper [4], to improve the speed of the method, it is proposed to transform a single-channel series into a multi-channel series before applying the tensor modification: $x_m^{(d)} = x_{(m-1)D+d}$, where $m \in \overline{1 : (N/D)}$. In that paper only the ESPRIT modification called HTLSDstack is considered, but we will apply this time series transformation for the signal estimation problem as well, and will call the resulting method SSADstack. Tensor modifications are constructed as for a multi-channel series.

3 Comparison of tensor and matrix methods

All numerical comparisons are made using time series that are expressed as sums of sinusoids.

The following methods were compared for single channel time series and signal extraction problem: SSA, HO-SSA, SSADstack, HO-SSADstack with $R_1 = r$ and HO-SSADstack with $R_1 = 1$. It has been shown that, in most cases, the SSA method significantly outperforms other methods in terms of accuracy. When the SSA method is less accurate, the difference is negligible and only occurs for a very narrow range of parameters. This minor disadvantage is therefore not a practical consideration. Of the Dstack methods, SSADstack and HO-SSADstack are the most accurate, with a small difference in accuracy when $R_1 = r$.

For single-channel time series and frequency estimation problem, a signal in the form of two sinusoids with close frequencies was considered. The ESPRIT, HO-ESPRIT, HTLSDstack, HO-HTLSDstack with $R_1 = r$ and HO-HTLSDstack with $R_1 = 1$ methods were compared. It was found that ESPRIT performs more accurately at a low noise level. However, at a medium or high noise level, HO-ESPRIT with optimal parameter selection becomes more accurate. Furthermore, HO-HTLSDstack with $R_1 = 1$ outperforms all methods.

For multi-channel time series, it has been demonstrated that, when all channels are expressed as a sum of sinusoids with equal frequencies, tensor modifications provide more accurate results for both signal extraction and frequency estimation.

4 Conclusions

Numerical comparisons revealed the varying effects of the HO-SVD tensor modifications on different time series problems. For signal extraction from a single-channel time series, the basic matrix method is certainly more accurate. However, for multi-channel time series with an equal set of frequencies across the channels, and for frequency estimation problems, the tensor methods can offer improved accuracy.

References

1. De Lathauwer L. (2011). Blind separation of exponential polynomials and the decomposition of a tensor in rank-$(L\_r, L\_r, 1)$ terms. *SIAM Journal on Matrix Analysis and Applications*. Vol. **32**, Num. **4**, pp. 1451-1474.

2. Golyandina N.E., Nekrutkin V.V., Zhigljavsky A.A. (2001). *Analysis of Time Series Structure*. Chapman and Hall/CRC: Boca Raton.

3. Papy J.M., De Lathauwer L., Van Huffel S. (2005). Exponential data fitting using multilinear algebra: the single-channel and multi-channel case. *Linear Algebra with Applications*. Vol. **12**, Num. **8**, pp. 809-826.

4. Papy J.M., De Lathauwer L., Van Huffel S. (2009). Exponential data fitting using multilinear algebra: the decimative case. *Journal of Chemometrics*. Vol. **23**, Num. **7-8**, pp. 341-351s.

5. Roy R., Kailath T. (1989). ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. Vol. **37**, Num. **7**, pp. 984-995.

6. Trung Le T., Abed-Meraim K., Trung N.L. (2024). Higher-order singular spectrum analysis for multichannel biomedical signal analysis. *2024 32nd European Signal Processing Conference (EUSIPCO), Lyon, France*. Pp. 1337-1341.

7. Van Huffel S., Chen H., Decanniere C., van Hecke P. (1994). Algorithm for time-domain NMR data fitting based on total least squares. *Journal of Magnetic Resonance, Series A*. Vol. **110**, Num. **2**, pp. 19-36.

ON TRUE VALUE AND UNCERTAINTY OF A PHYSICAL QUANTITY

E.Y. Kolesnikov[1]

[1]*Peter the Great Saint Petersburg Polytechnic University*
[1]*Civil Engineering Institute*
[1]*Higher School of Technospheric Security*
Saint Petersburg, RUSSIA
e-mail: [1]e.konik@list.ru

In this paper, we consider the problem of quantitative estimation in the interval statement of: a) uncertainty of the results of indirect observations; b) true value of a physical quantity. It is known, that in the case of indirect measurements with more than two input parameters, statistical processing of observation results in a probabilistic statement is possible only by statistical modeling. Estimation of the uncertainty value of the result of indirect measurements is performed in an interval statement using the example of the simplest three-parameter problem. It is shown that the idea of representing the true value of a physical quantity using an interval number to be promising. Measurement models are described for two types of physical quantity: the rod length and the weight of the load, and quantitative estimates of the width of the intervals of their true values are performed.

Keywords: physical quantity, indirect measurements, uncertainty of the measurement result, true value of the quantity, interval statement

1 Introduction

At the beginning of the twentieth century, thanks to the efforts of many researchers in the field of the processing of observation results, a classical approach was developed, including the concepts of the observation result, the true value of a quantity and the error, which was differentiated into random and systematic. Within the framework of probability theory, a powerful mathematical approach was developed that allowed us to evaluate various statistical parameters of samples of measured values of a quantity obtained during direct measurements.

For indirect measurements, this scheme worked much worse, since, as is known, the problem of finding statistical moments of a quantity that is an analytic function of two or more input parameters with known distribution laws generally has no analytical solution. This problem can be solved using statistical modeling (Monte Carlo) methods with considerable computational resources.

The basic idea of the approach, the difference between the true value and the measured value of a quantity, seemed perfectly natural and suitable for everyone. However, at the end of the seventies of the last century, British metrologists doubted the very existence of the true value of a quantity (in the classical point formulation): how can one operate (in the mathematical sense) with a value that, for a number of circumstances, will never be known?

In the end, these ideas were approved by all the leading metrological organizations of the world, which resulted in the expulsion of the concept of error from the international metrological circulation and its replacement by the new concept of measurement uncertainty, understood as an interval that can be assigned to the value of the measured quantity on the basis of all available information. The first version [1] of the GUM is "Guidelines for expressing Measurement Uncertainty"; it was approved (currently the eth modified edition [2] is in force).

We would like to point out a number of works [3, 4, 5] (not exhaustive, of course) that have questioned the main provisions of the probabilistic paradigm for processing observational results. Meanwhile, this paradigm, which in the 60s of the last century seemed unshakable, as if cast in bronze, on the periphery of scientific research (if the classical probabilistic statement is still considered mainstream), alternatives were developed, one of which was the interval approach.

As its forerunner we can consider the Soviet mathematician Vladimir Modestovich Bradis (known to the older generation from the tables of his name), who developed interval ideas in his works at the beginning of the last century, [6]. Interval analysis was further developed in the work of Rosalind Young [7], (1931), and in the works of many researchers from the Soviet Union, the United States, Great Britain, Germany, Poland and Japan, [6]. The work of academician Kantorovich [8], (1962), which, among other things, proposed the use of two-way estimates in processing the results of observations, had a serious impact on the research of Soviet mathematicians in the field of interval analysis.

In recent decades, research in the field of interval data statistics has continued and turned out to be very successful; some of its results can be found in the dissertation [9] and two collective monographs [6, 10]. Particularly important is the first book, published by well-known Russian experts in 2024.

2 Statistics of indirect measurement results

Consider the simplest three-parameter problem of finding the frequency of oscillation of a physical pendulum on a free suspension consisting of two point masses m_1 and m_2, kg for kilograms, connected by a rod of length l, m for meters (Figure 1). Model assumptions: the rod is weightless and absolutely rigid, and the rod can move on a horizontal surface without friction.

In the case of small oscillations (angle $\alpha \ll 1$, rad), ignoring the terms of the second order of smallness, the circular frequency ω, Hz, of the pendulum oscillations can be found, [11], by the relation

$$\omega = \sqrt{\left(1 + \frac{m_2}{m_1}\right)\frac{g}{l}}. \tag{1}$$

In the limit, for $m_2/m_1 \to 0$ (1), equation (1) becomes the classical equation of a mathematical pendulum.

Linear frequency ν, Hz, of the pendulum oscillations

$$\nu = \frac{1}{2\pi}\sqrt{\left(1 + \frac{m_2}{m_1}\right)\frac{g}{l}}. \tag{2}$$

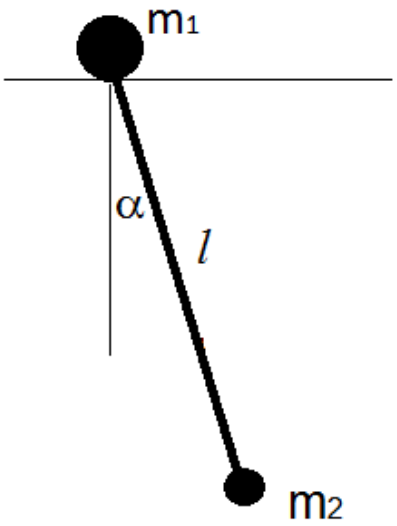


Figure 1: Pendulum

Let us assume that the values of all three input parameters of equation (2) are obtained on the basis of several direct measurements: the mass of the loads is determined using electronic scales of accuracy class 2 weighings each), the length of the rod is determined using a metal ruler 1000 mm long with a graduation of 1 mm (seven measurements), Table 1. The limits of the absolute error Δ of the measurement results for scales are calculated taking into account the accuracy class of the equipment, when measuring the length Δ is assumed to be equal to half of the scale interval of the measuring instrument.

Let us present the measurement results of the three input parameters of the problem in the interval formulation

For the purpose of visualization, the intervals will be displayed graphically (see Figure 2), with the example of the mass of the load m_1 (monographs [6] refer to such figures as scattering diagrams). The mean value is indicated by the dotted line.

In accordance with the terminology outlined in [6], it can be concluded that all three samples (m_1, m_2, l) are covering samples. This means that most of the experimental data contained in them in the form of interval numbers contains the true value of the measured values. The interval shells of the interval samples of the measured values of the input parameters of the problem are as follows:

a) mass of the first load: $m_1 \in [2.4549, 2.6531]$ kg;

b) mass of the first load: $m_2 \in [0.0189, 0.0227]$ kg;

c) the rod lengths: $l \in [0.6514, 0.6546]$ m.

Table 1: Measurement results

Load weight $m_1 \pm \Delta$ kg	Load weight $m_2 \pm \Delta$, kg	Rod length $l \pm \Delta$, m
$2,582 \pm 0,051$	$0,0212 \pm 0,0004$	$0,6540 \pm 0,0005$
$2,579 \pm 0,052$	$0,0212 \pm 0,0004$	$0,6540 \pm 0,0005$
$2,597 \pm 0,052$	$0,0200 \pm 0,0004$	$0,6540 \pm 0,0005$
$2,577 \pm 0,052$	$0,0203 \pm 0,0004$	$0,6540 \pm 0,0005$
$2,594 \pm 0,052$	$0,0202 \pm 0,0004$	$0,6520 \pm 0,0005$
$2,518 \pm 0,050$	$0,0198 \pm 0,0004$	$0,6530 \pm 0,0005$
$2,555 \pm 0,051$	$0,0207 \pm 0,0004$	$0,6530 \pm 0,0005$
$2,551 \pm 0,051$	$0,0194 \pm 0,0004$	
$2,505 \pm 0,050$	$0,0222 \pm 0,0004$	
$2,564 \pm 0,051$	$0,0199 \pm 0,0004$	
$2,523 \pm 0,050$	$0,0201 \pm 0,0004$	
$2,601 \pm 0,052$	$0,0211 \pm 0,0004$	
$2,531 \pm 0,051$	$0,0201 \pm 0,0004$	
$2,573 \pm 0,051$	$0,0199 \pm 0,0004$	

Table 2: Statistical parameters of loads and rods based on measurement results

Weight of load m_1 kg		Weight of load m_2, kg		Rod length l, m	
Average value	of SKO	Average value	of SKO	Average value	of SKO
2,561	0.031	0.0204	0.0008	0.6534	0.0005

Table 3: Measurement results in interval representation

Weight of cargo $m_1 \pm \Delta$, kg	Mass of the load $m_2 \pm \Delta$, kg	Length of the rod $l \pm \Delta$, m
$[2,5309,\ 2,6331]$	$[0,0207,\ 0,0217]$	$[0,6535,\ 0,6545]$
$[2,5270,\ 2,6311]$	$[0,0207,\ 0,0217]$	$[0,6535,\ 0,6545]$
$[2,5449,\ 2,6491]$	$[0,0195,\ 0,0205]$	$[0,6535,\ 0,6545]$
$[2,5249,\ 2,6291]$	$[0,0198,\ 0,0208]$	$[0,6535,\ 0,6545]$
$[2,5419,\ 2,6460]$	$[0,0197,\ 0,0207]$	$[0,6515,\ 0,6525]$
$[2,4679,\ 2,5681]$	$[0,0193,\ 0,0203]$	$[0,6525,\ 0,6535]$
$[2,5040,\ 2,6061]$	$[0,0202,\ 0,0212]$	$[0,6525,\ 0,6535]$
$[2,5000,\ 2,6021]$	$[0,0189,\ 0,0199]$	
$[2,4549,\ 2,5551]$	$[0,0217,\ 0,0227]$	
$[2,5129,\ 2,6151]$	$[0,0194,\ 0,0204]$	
$[2,4729,\ 2,5731]$	$[0,0196,\ 0,0206]$	
$[2,5489,\ 2,6531]$	$[0,0206,\ 0,0216]$	
$[2,4799,\ 2,5821]$	$[0,0196,\ 0,0206]$	
$[2,5219,\ 2,6241]$	$[0,0194,\ 0,0204]$	

The frequency, ν, of the pendulum oscillations in the interval statement can be calculated from relation (2) by direct substitution of the values of the input parameters

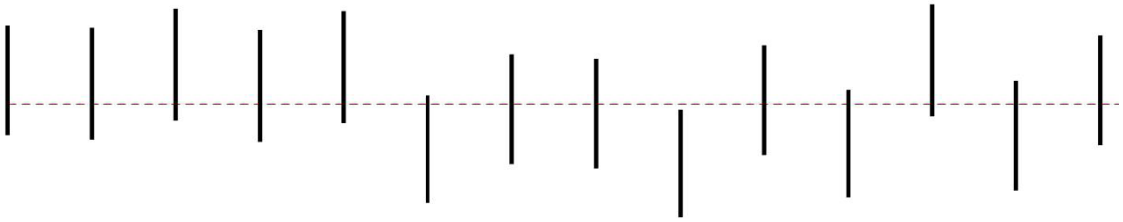


Figure 2: Results of measuring the mass of the load m_1 in an interval representation

into it. Interval calculations were conducted using the INTLab program [12], an interval application (toolbox) within the MATLAB package. In this particular instance, it was deemed unnecessary to implement preventive measures to mitigate widening of the intervals. This is because each input parameter is only used once in the calculated ratio. It is important to acknowledge that the aforementioned effect has historically served as a deterrent, hindering the extensive use of interval statement in engineering problems and interval calculations. Effective methods for its suppression have now been developed.

The result of calculating the linear oscillation frequency of the pendulum in the interval setting: $\nu \in [0.6183, 0.6205]$ Hz. This is an interval with an average value of 0.6194 Hz and a radius of 0.0011 Hz. Thus, the width of the interval that characterizes the uncertainty value of the oscillation frequency was equal to 0.0022 Hz.

For the purpose of comparison, please refer to Table 4, which presents the results of processing indirect measurement data obtained by three methods: statistical modeling, linear approximation and interval statement.

Table 4: The estimate uncertainty of the results of indirect measurements (linear frequency, Hz, oscillations of a physical pendulum)

Parameter	Statistical modeling	Linear approximation	Interval statement
Average value	0.6191	0.6195	0.6194
Interval nedefiniteness and	[0.6186; 0.6196]	[0.6188, 0.6202]	[0.6183, 0.,6205]
Width of the uncertainty interval	0.001	0.0014	0.0022

3 On true value of a physical quantity

The realization that the true value of any physical quantity cannot be determined through measurement led to the introduction of the concept of its actual value by Russian metrology. This is defined as a value “found experimentally and so close to its true value that it can be used instead of it for the problem being solved”.

In other words, even under ideal conditions (an impeccable measurement tool, measurement procedure, and method of processing the results obtained) the value obtained in the experiment will not coincide with the true value of the physical quantity. The reason for this lies in the inherent uncertainty of the measurement model, its attribute property, which consists in the fact that this model always idealizes and simplifies the measured object to one degree or another.

In the point setting, this led to the conclusion that the true value of a physical quantity in reality does not exist at all, since it is a kind of idealization. Meanwhile, the interval paradigm allows us to consider the true value of a physical quantity as a limited set of its actual point values, continuously filling its specific interval.

Let us demonstrate this thesis using the example of the input parameters of the problem we considered above, i.e. the length of the rod and the mass of loads:

1) the measuring model of a rod is a cylinder with plane-parallel ends orthogonal to its axis, made of steel. The length of such a cylinder is a point number.

Meanwhile, for any real rod (cylinder) it is evident that the end surfaces are not perfectly flat, and are not parallel to each other or orthogonal to the axis. Furthermore, the length of the object varies with temperature due to thermal expansion. The result of the thought experiment would be that a set of point values would be continuously generated, filling a certain interval. These values would be the result of measuring the length of the rod under study at different points on the end surface and parallel to the axis, at different temperatures. It is evident that the convex set is an interval number, and thus should be considered as the true value of the length of the rod in question.

2) the measuring model for determining mass of a load is predicated on the measurement of its weight. The process entails the measurement of its gravitational mass, which, as has been repeatedly and rigorously demonstrated through experimental means, is equivalent to its inertial mass with a high degree of precision. In the event of the gravitational mass of a load being measured with absolute accuracy at differing geographical locations and over a specified period of time, the resultant data would be a set of point values. This is due to the fact that the result of measuring weight is proportional to the acceleration of free fall, which, as is known, varies within certain limits on the Earth surface, both in spatial and temporal terms. If the number of these points being directed to infinity, the resulting set of point values of the load mass becomes continuous, and its true value becomes an interval number.

Let us try to quantify the width of such an interval in a rough approximation.

Assuming that:

a) the coefficient of linear expansion of structural steel $\alpha \in [11, 12]$ is $10^{-6}\ K^{-1}$, and the temperature range $T \in [288, 301]$ K. Then the estimation of the width of the interval of the true value of its length, due to the influence of temperature, will be as follows:

$$\Delta_T = l\alpha\Delta T = 2.9 \cdot 10^{-4} m; \tag{3}$$

b) the end surfaces of the rod are made with a tolerance of 0.8 microns, or $8 \cdot 10^{-7}$ m, and their parallelism: with a tolerance of 0.02 mm, or $2 \cdot 10^{-5}$ m,

c) the dependence of the acceleration of free fall g, m/s^2, on the geographical

latitude φ, °, point and height h, m, its location above sea level, can be calculated by the empirical formula, [13]:

$$g = 9.780327 \cdot (1 + 0.0053024 \cdot \sin^2 \varphi - 0.0000058 \cdot \sin^2 2\varphi) - 3.086 \cdot 10^{-6} \cdot h. \quad (4)$$

Considering, that the southernmost city in Russia is Derbent (42° N., height 5 m), and the northernmost city is Pevek (69° N., height 100 m); we get estimates of the width (radius) of the intervals of the true values:

- of the rod length $3.1 \cdot 10^{-4}$ m,
- of weight of the first load 0.0113 kg (taking into account the width of the interval $g \in [9.8035, 9.8252]$, equal to 0.0217 m/s^2);
- the second load is $9.0 \cdot 10^{-5}$ kg.

4 Conclusion

Notwithstanding its centuries-old history, the problem of statistical processing of the results of observations remains relevant topic of research, and an alternative to the classical probabilistic formulation of the problem is a new direction that has been actively developed recently: the interval approach to the problem, or statistics of interval data.

Within the interval paradigm, the imaginary problem of non-existence of the true value of a physical quantity is removed.

Using the example of two physical quantities, the study demonstrated that their true values form bounded continuous sets that can be represented as interval numbers. In addition, quantitative estimates of the width of these intervals were performed.

References

1. BIPM (1980), Rapport BIPM-80/3, Report on the BIMP enquiry n error statements, Bur. Intl. Poids et Mesures (Sevres, France).
2. JCGM 100:2008 GUM 1995 with minor corrections Evaluation of measurement data. Guide to the expression of uncertainty in measurement.
3. Alimov Yu.I., Kravtsov Yu.A. (1992). Is probability a normal physical quantity? *Advances in physical Sciences*. Vol. **162**, Num. **7**, pp. 149-182 (in Russian).
4. Gorban I.I. (2016). *Randomness and hyper-randomness*. Naukova dumka, Kiev. (In Russian)
5. Tutubalin, V.N. (1977). *Limits of applicability (probabilistic statistical methods and their capabilities)*. Moscow: Znanie. (In Russian)

6. Bazhenov A.N., Zhilin S.I., Kumkov S.I., Sharyi S.P. (2024). Processing and analysis of interval data. *SIC Regular and Chaotic Dynamics*. (In Russian)

7. Young, R.C. (1931). The algebra of many-valued quantities. Mathematische Annalen. 104. pp. 260–290.

8. Kantorovich L.V. (1962). On some new approaches to computational methods and processing of observations. *Siberian Mathematical Journal*. Vol. **3**, Num. **5**, pp. 701-709. (In Russian)

9. Postovalov S.N. (1997). Statistical analysis of interval observations of one-dimensional continuous random variables. Diss. Candidate of Technical Sciences-Novosibirsk. (In Russian)

10. Hung T. [et. al.] (2012). Computing Statistics under Interval and Fuzzy Uncertainty. *Springer-Verlag Berlin Heidelberg*.

11. Samarskiy A.A., Mikhailov A.P. (2012). *Mathematical modeling: Ideas. Methods. Examples*. FIZMATLIT, Moscow. (In Russian)

12. Rump S.M. (1999). INTLAB – INTerval LABoratory. In Tibor Csendes, editor. *Developments in Reliable Computing – Dordrecht: Kluwer Academic Publishers*. pp. 77–104.

13. Acceleration of free fall: physical encyclopedia [in 5 volumes]. Ch. ed. A.M. Prokhorov. M: Great Russian Encyclopedia, 1998. (In Russian)

ASYMPTOTIC ANALYSIS OF EXPECTED REVENUES IN G-NETWORK WITH UNRELIABLE LINES SERVICE AND LIMITED WAITING TIME POSITIVE AND NEGATIVE CUSTOMERS

D.Y. Kopats[1]

[1] *Yanka Kupala State University of Grodno*
Grodno, BELARUS
e-mail: [1]dk80395@mail.ru

We study G-network consisting of multiline unreliable queuing systems (QS) that receive impatient positive and negative requests and previously obtained results on asymptotic analysis and research of similar networks. The impatience of negative requests is manifested in the destruction of positive requests by them not immediately, but after a random time, and the impatience of positive ones is manifested in limiting the waiting time for the start of servicing positive requests, after which it can move along the network systems or leave the network. Using asymptotic analysis for a large but limited number of requests operating in the network, expressions for the expected income of this network are found. A system of difference-differential equations is obtained, which are satisfied by the expected incomes of the network systems. Next, an equation in partial derivatives is obtained for the income distribution density and a return to the expression for expected incomes is carried out. In conclusion, the findings of the work are presented and the prospects for studying queuing networks (QSN) using this method are noted.

Keywords: G-network, asymptotic analysis, expected revenues

1 Introduction

G-networks as a type of queueing networks (QN) were first introduced in the article [1]. In the article [2] a G-network with unreliable service lines (SL) was studied in the transient mode, in the case when the SL failed due to reasons not related to computer viruses. In the article [3] G-network with impatient positive and negative customers is studied in the transient regime. For negative customers, impatience is understood as the ability to harm the QS not immediately, but after a certain period of time.

The first work on the asymptotic analysis of QN with a large but limited number of requests functioning in the network is the work of Medvedev [4]. The method of asymptotic analysis was first applied to G-networks in [5]. This method was not applied to G-networks with impatient positive and negative requests, and for a G-network with unreliable QS, an asymptotic analysis was performed in [6]. For a network with unreliable LO and impatient requests, an asymptotic analysis was applied in [7]. This work generalizes the results of [6,7] for finding the expected income of network systems in the case where the network parameters depend on the network state and time, that is, the results of [7] are generalized to the case of functioning of negative requests in

the network, and the results of [6] to the case where the network parameters depend on the network state, time, as well as the impatience of positive and negative requests and the unreliability of the LS.

2 Network description

We consider closed G-network [1] with $n+1$ queuing systems (QS) $S_0, \ldots, S_n$. In earch QS consist m_i line service, $i \in \{0, \ldots, n\}$. In network moves K positive and K negative customers.Independent Poisson flow of positive customers with rate λ_{0i}^{+} and Poisson flow of negative customers with rate λ_{0i}^{-} arrive to QS S_i from outside (system S_0), $i \in \{1, \ldots, n\}$. All Suppose S_0 handles are reliable, and in other systems $S_1, \ldots, S_n$ lines can be damaged. The servicing time for QS customers has an exponential distribution with a parameter $\mu_i(d_i, k_i, l_i)$, where d_i is number undamaged lines in S_i, k_i, l_i are numbers of positive and negative customers in i-th QS; the operation time of each line without damage in this system has an exponential distribution with the parameter $\beta_i(d_i, k_i, l_i)$, $i \in \{1, \ldots, n\}$. After the damage of the line immediately begins repairing it, the repair time also has an exponential distribution with the parameter $\gamma_i(d_i, k_i, l_i)$, $i \in \{1, \ldots, n\}$. Positive customer being served in S_i is moved to QS S_j with probability p_{ij}^{+} as a positive customer, and with probability p_{ij}^{-} as a negative customer, and with probability $p_{i0} = 1 - \sum_{j=1}^{n}(p_{ij}^{+} + p_{ij}^{-})$, $i, j \in \{1, \ldots, n\}$, come out from the network to external environment. Each positive customer located in i-th QS, stay in the queue random time according to a Poisson process of rate $\theta_i(d_i, k_i, l_i)$, $i \in \{1, \ldots, n\}$. By the end this time, positive customer is moved to j-th QS as positive customers with probability q_{ij}^{+}, with probability q_{ij}^{-} as negative customers, and with probability $q_{i0} = 1 - \sum_{i=1}^{n}(q_{ij}^{+} + q_{ij}^{-})$, $i \in \{1, \ldots, n\}$. Negative customer is arrived to QS increases the length of the queue of negative customers for one, and requires no service. Each negative customer, located in i-th QS, stay in the queue random time according to a Poisson process of rate $\mu_i^{-}(d_i, k_i, l_i)$, $i \in \{1, \ldots, n\}$. By the end this time, negative customer destroy one positive customer in the QS S_i and leave the network. With used asymptotic analysis finding expected revenues of this network.

Let $(\vec{d}, \vec{k}, \vec{l}, t) = (d_1, \ldots, d_n, k_1, \ldots, k_n, l_1, \ldots l_n, t)$. Same as [6] it can be shown that network expected revenues satisfy different-difference equations (DDE) system:

$$
\begin{aligned}
\frac{dV(\vec{d}, \vec{k}, \vec{l}, t)}{dt} = R + \sum_{i=1}^{n} \Bigg[& \lambda_{0i}^{+}(V(\vec{d}, \vec{k} + I_i, \vec{l}, t) - V(\vec{d}, \vec{k}, \vec{l}, t)) + \\
& + u(k_i) p_{i0} \Big(\mu_i(d_i, k_i - 1, l_i) V(\vec{d}, \vec{k} - I_i, \vec{l}, t) - \mu_i(d_i, k_i, l_i) V(\vec{d}, \vec{k}, \vec{l}, t) \Big) \\
& + \gamma_i(d_i + 1, k_i, l_i) V(\vec{d} + I_i, \vec{k}, \vec{l}, t) - \gamma_i(d_i, k_i, l_i) V(\vec{d}, \vec{k}, \vec{l}, t) \\
& + \mu_i^{-}(d_i, k_i - 1, l_i - 1) V(\vec{d}, \vec{k} - I_i, \vec{l} - I_i, t) - \mu_i^{-}(d_i, k_i, l_i)) V(\vec{d}, \vec{k}, \vec{l}, t) \\
& + u(k_i) q_{i0} \Big(\theta_i(d_i, k_i - 1, l_i) V(\vec{d}, \vec{k} - I_i, \vec{l}, t) - \theta_i(d_i, k_i, l_i) V(\vec{d}, \vec{k}, \vec{l}, t) \Big) \\
& + \lambda_{0i}^{+} r_{0i} - \mu_i(d_i, k_i - 1, l_i) u(k_i) p_{i0} R_{i0} - \theta_i(d_i, k_i - 1, l_i) u(k_i) q_{i0} H_{i0}
\end{aligned}
$$

$$- g_i\gamma_i(d_i+1,k_i,l_i) - \mu_i^-(d_i,k_i-1,l_i-1)R_i^-\Bigg]$$

$$+\sum_{i,j=1}^{n}\Bigg[u(k_i)p_{ij}^+\Big(\mu_i(d_i,k_i-1,l_i)V(\vec{d},\vec{k}-I_i+I_j,\vec{l},t)-\mu_i(d_i,k_i,l_i)V(\vec{d},\vec{k},\vec{l},t)\Big)$$

$$+u(k_i)q_{ij}^+\Big(\theta_i(d_i,k_i-1,l_i)V(\vec{d},\vec{k}-I_i+I_j,\vec{l},t)-\theta_i(d_i,k_i,l_i)V(\vec{d},\vec{k},\vec{l},t)\Big)$$

$$+u(k_i)p_{ij}^-\Big(\mu_i(d_i,k_i-1,l_i)V(\vec{d},\vec{k}-I_i,\vec{l}+I_j,t)-\mu_i(d_i,k_i,l_i)V(\vec{d},\vec{k},\vec{l},t)\Big)$$

$$+u(k_i)q_{ij}^-\Big(\theta_i(d_i,k_i-1,l_i)V(\vec{d},\vec{k}-I_i,\vec{l}+I_j,t)-\theta_i(d_i,k_i,l_i)V(\vec{d},\vec{k},\vec{l},t)\Big)$$

$$\mu_i(d_i,k_i-1,l_i)u(k_i)p_{ij}^+R_{ij}^+ + \theta_i(d_i,k_i-1,l_i)u(k_i)q_{ij}^+H_{ij}^+$$

$$-\mu_i(d_i,k_i-1,l_i)u(k_i)p_{ij}^-R_{ij}^- - \theta_i(d_i,k_i-1,l_i)u(k_i)q_{ij}^-H_{ij}^-\Bigg]. \quad (1)$$

The solution of the system (1) in an analytic form is difficult task. Therefore, we shall consider the asymptotic case of a big number of customers in the network, that is, we assume that $K \gg 1$. To find the probability distribution of the random vectors $\vec{k}(t),\vec{l}(t),\vec{d}(t)$, we move on to the relative variables and consider the vector $\eta(t)K^{-1} = (d_1K^{-1},\dots,d_nK^{-1},k_1K^{-1},\dots,k_nK^{-1},l_1K^{-1},\dots,l_nK^{-1})$. Possible values of this vector at a fixed t belong to a bounded closed set

$$G = \Bigg\{(\vec{y},\vec{x},\vec{z},t) = (y_1,\dots,y_n,x_1,\dots,x_n,z_1,\dots,z_n,t):$$

$$x_i,z_i \geq 0, \quad \sum_{i=0}^{n} x_i = K, \quad \sum_{i=0}^{n} z_i = K, \quad 0 \leq y_i \leq m_iK^{-1}\Bigg\}. \quad (2)$$

We can introduce the distribution density function of expected income in the region:

$$\rho(\vec{y},\vec{x},\vec{z},t) = \lim_{\epsilon\to 0} V(y_1 \leq \xi_1 \leq y_1+\epsilon,\dots,y_n \leq \xi_n \leq y_n+\epsilon, x_1 \leq \xi_{n+1} \leq x_1+\epsilon,\dots,$$
$$x_n \leq \xi_{2n} \leq x_n+\epsilon, z_1 \leq \xi_{2n+1} \leq z_1+\epsilon,\dots,z_n \leq \xi_{3n} \leq z_n+\epsilon,t)\epsilon^{-3n}. \quad (3)$$

Same as [6] it can be shown that the distribution density function of expected income with precision $\mathcal{O}(K^{-2})$, satisfy partial difference equations:

$$\frac{\partial\rho(\vec{y},\vec{x},\vec{t},t)}{\partial t} = -\frac{\partial}{\partial y_i}\sum_{i=1}^{n} A_i^{(1)}(\vec{y},\vec{x},\vec{t},t)\rho(\vec{y},\vec{x},\vec{t},t) - \frac{\partial}{\partial x_i}\sum_{i=1}^{n} A_i^{(2)}(\vec{y},\vec{x},\vec{t},t)\rho(\vec{y},\vec{x},\vec{t},t)$$

$$-\frac{\partial}{\partial z_i}\sum_{i=1}^{n} A_i^{(3)}(\vec{y},\vec{x},\vec{t},t)\rho(\vec{y},\vec{x},\vec{t},t) + \sum_{i=1}^{n}\Bigg[\lambda_{0i}^+r_{0i}^{(1)} - u(k_i)p_{i0}r_{i0}\frac{\partial\mu_i(y_i,x_i,z_i)}{\partial x_i}$$

$$-\frac{\partial\theta_i(y_i,x_i,z_i)}{\partial x_i}u(k_i)q_{i0}h_{i0} - g_i^{(1)}\frac{\partial\gamma_i(y_i,x_i,z_i)}{\partial y_i} - \frac{\partial\mu_i^-(y_i,x_i,z_i)}{\partial x_i}r_i^-\Bigg]$$

$$- \sum_{i,j=1}^{n} \Bigg[u(k_i) p_{ij}^{+} r_{ij} \frac{\partial \mu_i(y_i, x_i, z_i)}{\partial x_i} + \frac{\partial \theta_i(y_i, x_i, z_i)}{\partial x_i} u(k_i) q_{ij}^{+} h_{ij}^{+} + u(k_i) p_{ij}^{-} r_{ij} \frac{\partial \mu_i(y_i, x_i, z_i)}{\partial x_i} + \frac{\partial \theta_i(y_i, x_i, z_i))}{\partial x_i} u(k_i) q_{ij}^{-} h_{ij}^{-} \Bigg], \quad (4)$$

where

$$\begin{aligned}
A_i^{(1)}(\vec{y}, \vec{x}, \vec{t}, t) &= \gamma_i(\vec{y}, \vec{x}, \vec{t}, t), \\
A_i^{(2)}(\vec{y}, \vec{x}, \vec{t}, t) &= -\lambda_{0i}^{+} + u(x_i)\mu_i(y_i, x_i, z_i) p_{ij}^{+*} \\
&\quad + \mu_i^{-}(y_i, x_i, z_i) + u(x_i)\theta_i(y_i, x_i, z_i) q_{ij}^{+*}, \\
A_i^{(3)}(\vec{y}, \vec{x}, \vec{t}, t) &= \mu_i^{-}(y_i, x_i, z_i) - u(x_i)\mu_i(y_i, x_i, z_i) p_{ij}^{-*} \\
&\quad + \mu_i^{-}(y_i, x_i, z_i) - u(x_i)\theta_i(y_i, x_i, z_i) q_{ij}^{-*}, \\
p_{ij}^{+*} &= \begin{cases} 1 - p_{ij}^{+}, & i = j \\ p_{ij}^{+}, & i \neq j, \end{cases}
\end{aligned}$$

and similarly p_{ij}^{-*}, $q_{ij}^{\pm *}$ are derived from p_{ij}^{-}, $q_{ij}^{\pm}$ respectively.

3 Conclusion

The article presents a G-network with unreliable LS and impatient positive and negative orders in the case when the number of orders operating in the network systems is large but limited. Asymptotic analysis is used to solve the DDE system. In the future, it is planned to consider a similar network with incomes and various features.

References

1. Gelenbe, E. Product form queueing networks with negative and positive customers // Journal of Applied Probability. 1991. V. 28. P. 656–663.

2. Matalytski, M. , Naumenko, V. Non-stationary analysis of queueing network with positive and negative messages // Journal of Applied Mathematics and Computational Mechanics. 2013. V. 12, No 2. P. 61–71.

3. Kopats, D., Naumenko, V., Matalytski M. Finding expected revenues with randon waiting time positive and negative customers [In Russian] // Vesnik GrSU. 2017. V. 7, No 1. P. 147–153.

4. Medvedev, G. Closed queueing systems and their optimisation // Izvestia NAN USSR. Technicheskaya kibernetika. 1978. No 6. P. 199–203.

5. Kopats, D. Asymptotic analysis of G-networks with many-lines queueing systems [In Russian] // Vesnik GrSU. 2021. V. 11. No 3. P. 138-146.

6. Monko, V., Kopats, D., Statkevich S. Asymptotic analysis exponentional network with limited waiting time and unreliable queueing systems . [In Russian] // Vesnik GrSU. 2015. No 1. P. 138–147.

7. Kopats, D. Asymptotic analysis G-network with unreliable many-lines queueing systems // Information technology and mathematical modeling. Tomsk, 1-5 dec. 2021 . P. 42-48.

ON THE CONVERGENCE OF A TRAINING ALGORITHM BASED ON THE BOLTZMANN ANNEALING OPTIMIZATION SCHEME

V.V. Krasnoproshin[1], V.V. Matskevich[2]

[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]krasnoproshin@bsu.by, [2]matskevich1997@gmail.com

The paper deals with a state-of-art problem, related to neural network training with using algorithms based on random search. One of the main problems while training using such algorithms is convergence problem. An original algorithm, based on modification of Boltzmann annealing scheme is proposed, for which theoretical convergence by probability to optimal solution from any initial is proved.

Keywords: annealing method, training, optimization, convergence

1 Introduction

At the present time a wide class of applied problems solved using neural network technologies. In this case, the effectiveness of their solution significantly depends on the quality of neural network training. One of the main tools for training neural networks are gradient algorithms. They have a number of undeniable positive properties, due to which they are widely used in practice [1]. At the same time, some algorithms features have been identified that may limit their use in practice. In particular, they do not guarantee convergence to an optimal solution [2]. As the range of applied problems using neural networks expands, many shortcomings may become critical [3]. Therefore, there is a need to search for alternative approaches to training [4]. Random search is such an approach in relation to gradient (directed) methods. In turn, random search-based training algorithms, on the contrary, can ensure convergence [5]. The main problem of the algorithms is their slow convergence [6]. However, at present, there are virtually no theoretical studies of such algorithms convergence. The paper examines the specifics of the optimization space in neural network training problems. It has been shown that it has a unique property of heterogeneity, taking into account which can significantly speed up the training process. A training algorithm based on a modification of the Boltzmann annealing optimization scheme that takes this property into account is proposed. Research is conducted to study the nature of its convergence.

2 Neural networks training

Training neural networks is a typical conditional optimization problem [7]. The features of the training process are considered and it is shown that the solution space has the property of a peculiar heterogeneity. In particular, it is shown that in any parameter's variation range there is a relatively small segment of values, the probability of finding

an optimal solution in which is significantly higher than in the remaining range [8]. This means that the power of the set in which the optimal solution is located is much less than the power of the entire solution space. It was also shown that taking this property into account in the random search algorithm allows for a significant reduction in training time.

One of the effective random search algorithms, both from a theoretical and practical point of view, is Boltzmann annealing. The main advantage of this algorithm is its guaranteed convergence in probability to the optimal solution. However, the convergence rate of Boltzmann annealing is logarithmic, which is critical for training large networks. An approach to constructing a training algorithm based on a modification of the Boltzmann annealing optimization scheme is proposed, which takes into account the parameter space heterogeneity.

Let us assume that the objective function F is defined on a finite set of feasible solutions Ω, for example, mean square error function, and for each element $x \in \Omega$ it is possible to build neighbor elements' vicinity $N(x) \subset \Omega$. Then the conditional optimization problem can be defined as a triple (Ω, F, N).

Let us use an algorithm based on the Boltzmann annealing optimization scheme to solve this problem, which can be briefly described as follows.

Initial iteration. The initial solution x_0 and the sequence of temperatures $T_k = T_0/\ln(k+2)$ are given (T_0 is initial temperature).

General k-th iteration.

Step 1. Generation of a new solution y based on the current solution x. To do this, generate a multidimensional increment r to the vector x. Each increment coordinate is specified by the realization of a uniformly distributed random variable on a segment centered at zero. New solution is determined as $y = x + r$.

Step 2. The objective function estimate for the new solution is calculated ($F(y)$).

Step 3. New solution is accepted with probability

$$P(x' = y|x) = \min\{1, \exp((F(x) - F(y))/T_k)\}$$

Stop criteria. If the time for neural network training has expired, the algorithm terminates. Otherwise, move to the next iteration is performed.

Let us consider the issues of this training algorithm convergence. First, we introduce a number of known definitions that will be needed later [9].

A path connecting solution $x \in \Omega$ with solution $y \in \Omega$, is a sequence $x_1, x_2, , x_n$:

$$\begin{cases} x_1 = x, x_n = y & \\ x_i \in \Omega, & i = \overline{1,n} \\ x_i \in N(x_{i-1}), & i = \overline{2,n} \end{cases}$$

A solution $y \in \Omega$ is reachable from $x \in \Omega$ at height h if: $x = y, F(y) \leq h$, or there is a finite sequence of solutions $x = x_0, x_1, x_2, ..., x_p = y, p > 0$ such that:

$$\begin{cases} x_{k+1} \in N(x_k), & \forall k = \overline{0,p-1} \\ x_k \in \Omega, F(x_k) \leq h, & k = \overline{0,p} \end{cases}$$

Local minima depth of x is called the smallest value d, for which condition $\exists y \in \Omega : F(y) < F(x)$, y is reachable from x at height $F(x) + d$ is satisfied.

According to convergence theorem [9] algorithm converges in probability to optimal solution, iff the series $\sum_{k=1}^{+\infty} \exp(-d^*/T_k) = +\infty$ diverges, where T_k is a temperature series and d^* is the biggest local minimum depth.

Other theorem limitations and their feasibility were shown in [10]. In [10] it was shown that to ensure convergence it is necessary and sufficient that $T_0 >= d^*$. However, the value of the parameter d^* is not known. Let us conduct a study of the algorithm's efficiency with a limited number of iterations and determine the dependence of d^* value on the algorithm's parameters. For simplicity of research, we will consider the solution space for the one-dimensional case. After that, the obtained results will be formulated for the general case. For the solution space, the following statements are proved.

Statement 1. *Let y be the global minimum point. The equality $d^* = d$ holds iff the following conditions are met:*

$$\begin{cases} \exists x \in \Omega : \forall \epsilon > 0, \exists [a,b] \in \Omega, |a-b| = l, [a,b] \subseteq [x,y] : \min\limits_{g \in [a,b]} F(g) \geq F(x) + d - \epsilon; \\ \forall x \in \Omega : \nexists [a,b] \in \Omega, |a-b| = l, [a,b] \subseteq [x,y] : \min\limits_{g \in [a,b]} F(g) > F(x) + d. \end{cases}$$

From these conditions it is clear that the sought value d^* can be interpreted as the maximum value at which there is a segment of length l that separates the point x from the point of the global minimum y. I.e., a segment of length l must have two properties:

a) the objective function minimum value on the segment must be greater than its value at point x by the amount d^*;

b) points x and y must be located on opposite sides of this segment.

Note. There may be several dividing segments with the specified properties.

The neural network training process can be represented as an optimization problem in a multidimensional solution space Ω. Let us reformulate statement 1 for the multidimensional case. To do this, we introduce a special set Ω_3 that depends on the selected point x and the d value: $\Omega_3 = \{g \in \Omega | F(g) > F(x) + d\}$ Let us also introduce the set $\Omega_2 = \Omega \setminus \Omega_3$.

Statement 2. *Let y be the global minimum point. The equality $d^* = d$ holds if and only if, when the following conditions are met:*

a) for $\forall d < d^*, \exists x \in \Omega$, in set Ω_2 does not exist path from x to y;

b) $\forall d \geq d^*$, $\forall x \in \Omega$, in set Ω_2 there exists path from x to y.

Thus, the value of d^* depends only on objective function F, solution searching space Ω and sets of neighboring points $N(x)$.

It follows that the function estimating the training algorithm convergence rate under consideration is a unimodular function with respect to the power of the set $N(x)$. I.e.:

The smaller the power of the set, the higher the value of d^* and the higher the value of T_0 required for convergence to the optimal solution vicinity, and, consequently, a large number of iterations are required for convergence. The greater the power of the set,

the smaller d^* value and the lower T_0 value. However, the large power of the set $N(x)$ reduces the probability of generating a solution in the vicinity of the optimal solution, since a multidimensional uniform distribution is specified.

References

1. Zhang, C. [et. al.] (2021). Gradient descent optimization in deep learning model training based on multistage and method combination strategy. *Security and Communication Networks*. Vol. **2021**. P. 9956773–9956787.

2. Arora, S., Li, Z., Panigrahi, A. (2022). Understanding gradient descent on the edge of stability in deep learning. *Proc. Machine Learning Research*. P. 948–1024.

3. Tan, S.W. [et. al.] (2021). The estimation life cycle of lithium-ion battery based on deep learning network and genetic algorithm. *Energies*. Vol. **14**, No. **15**. P. 4423–4443.

4. Bu, S.J., Kim, H.J. (2022). Optimized URL Feature Selection Based on Genetic-Algorithm-Embedded Deep Learning for Phishing Website Detection. *Electronics*. Vol. **11**, No. **7**. P. 1090–1101.

5. He, F., Ye, Q. (2022). A bearing fault diagnosis method based on wavelet packet transform and convolutional neural network optimized by simulated annealing algorithm. *Sensors*. Vol. **22**, No. **4**. P. 1410–1426.

6. Bandyopadhyay, R. [et. al.] (2021). Harris Hawks optimisation with Simulated Annealing as a deep feature selection method for screening of COVID-19 CT-scans. *Applied Soft Computing*. P. 111.107698–107712

7. Nakamura, K., Hong, B.W. (2019). Adaptive weight decay for deep neural networks. *IEEE Access*. Vol. **7**. P. 118857–118865.

8. Krasnoproshin, V.V., Matskevich, V.V. (2024). Random search in neural networks training. *Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications*. Vol. **34**, No. **2**. P. 309–316.

9. Hajek, B. (1988). Cooling schedules for optimal annealing. *Mathematics of operations research*. Vol. **13**, No. **2**. P. 311–329.

10. Krasnoproshin, V.V., Matskevich, V.V. (2022). Random search in neural networks training. *Proc. Computer Data Analysis and Modeling (CDAM-2022)*. P. 96–99.

ADAPTIVE METHOD OF DYNAMIC LOCAL APPROXIMATION IN IT SERVICE SCALING PROBLEMS

V.V. Krasnoproshin[1], A.A. Starovoitov[2]

[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]krasnoproshin@bsu.by, [2]starovoytovaa@bsu.by

The proposed method enables adaptation to new workload patterns associated with uncertainty by approximating real-time data using neural network models in IT service scaling problems.

Keywords: decision-making, proactive management, external load uncertainty, neural networks, multi-model system

1 Introduction

Uncertainty in external load and failures of computational equipment lead to disruptions in operation and degradation of performance in critical IT systems. As a result, the timeliness of information processing and execution of banking and other operations is lost, which in turn can have serious consequences (financial losses, major accidents, etc.).

With the advent of the cloud computing era, it became possible to automatically adapt the volume of computational resources of critical IT systems to the current load. Autonomous solutions emerged, capable of promptly managing the scaling of critical IT services without human involvement. In general, scaling is a task of automated management of computational resources under changing, highly dynamic, complex load with uncertainty. An ideal automatic scaling mechanism is capable of minimizing both costs and violations of service quality.

The literature [1] describes many different research solutions using predictive models (predictors based on neural networks, autoregressive models, etc.), which enable operational decision-making. However, there are certain features that prevent these systems from achieving maximum efficiency under complex, dynamic load with uncertainty.

The key feature of any automated control system lies in its predictive capabilities (forecast accuracy and timeliness). Predictions are made based on a sufficiently large set of previously obtained data. Training on these samples takes a considerable amount of time and requires high-performance computational resources to effectively train models within an acceptable timeframe. An example of historical data (average CPU utilization across computational modules over 4 hours of system operation) is shown in Figure 1.

A model trained on one dataset (associated with a specific load profile) that demonstrates good predictive performance on a similar profile may show weak results on a different profile. Essentially, each load profile may correspond to its own predictive model. Alternatively, the training dataset must contain a sufficiently large number

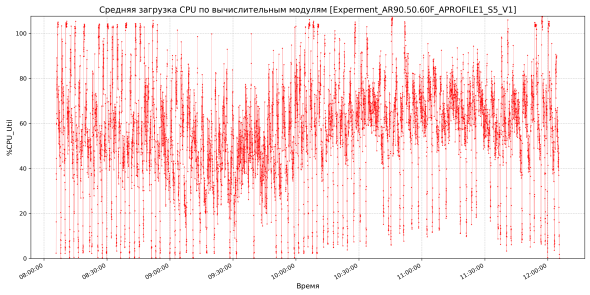


Figure 1: Example of system historical operation data (mean CPU utilization across compute modules over 4 hours)

of features that the model can recognize and memorize, enabling some level of generalization when making predictions. A certain contradiction arises due to the lack of adaptation of a pre-trained model to new changes in the external load profile, which are associated with uncertainty. Clearly, adaptation is possible if changes in the profile are used to improve the existing model or create a new one.

This study proposes a method that enables adaptation to new load patterns associated with uncertainty by constructing an approximation for data obtained in real time using neural network models. These models, along with metadata (training quality metrics, data characteristics), are stored in a model library during operation and are subsequently used for forecasting when the system transitions to new states.

2 Key Features of Real-Time Management

The key challenge lies in ensuring the timeliness and accuracy of decision-making. The forecast must be generated within seconds and predict changes tens of seconds ahead. At the same time, the system faces a number of constraints: data arrives with uncertainty, the volume of historical samples is small, and their collection frequency is limited (for example, one sample every 2 seconds). Moreover, it is impossible to pre-train the model: adaptation must occur in real time.

The system's flexibility plays a crucial role. Since the nature of the load can change, the algorithm must adapt to new conditions using minimal data. Traditional approaches requiring preliminary training or large datasets are unsuitable. Instead, a mechanism capable of adapting during operation is necessary to ensure stability and efficiency in resource management.

3 Idea and Method

This ultimately led to the idea that what happened to the system over a long period of time is not so important. What matters is the ability to make adequate predictions based on a small amount of data accumulated in real time. Training the model on this data and subsequent predictions should be performed quickly enough, in parallel with the accumulation of new data.

In essence, the idea boils down to the fact that during system operation, the load profile can be divided into small regions. For each region, during its existence, a model can be built that describes the behavior in that region and allows making a set of predictions, based on which the system can transition to a new state. Thus, models will be created that will have better accuracy within their domain of competence compared to a global model trained on data prepared over a longer period of time. An illustration of this idea is presented in Figure 2.

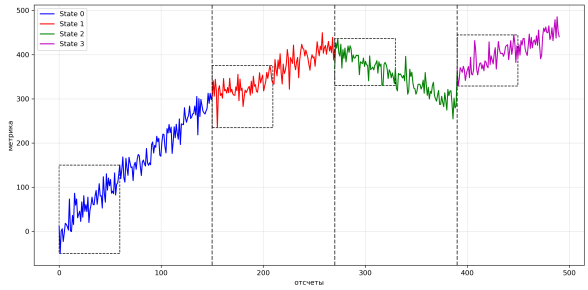


Figure 2: Segmentation of the load profile into states and selection of state-specific data for local model training

Forecasting within the local time region leads to the following challenge. Let's consider a critical system that can be in a finite set of states $S = \{S_1, S_2, \ldots, S_n\}$, determined by external load on the system and transition levels. In each state i the system generates a discrete signal in the form of a short time series $X_t^{S_i}$ (or a set of time series). The task is to build a predictor based on a neural network using a portion of this signal, denoted as $\tilde{X}_t^{S_i}$, which forecasts the system's transition to a new state S_{i+1}.

To solve this problem, we developed the Dynamic Local Approximation by Neural Network models (DLANN) method. Its essence lies in constructing a simple neural network model (with one hidden and one output layer) for each system state during operation. This model trains on partial data from the local region, with the assumption that a model trained on partial data will adequately forecast for the entire local region (the concept originates from the Pareto principle). After training, the model uses newly arriving data to generate forecasts of average %CPU values across a set of computing modules with specified lead time. These forecasts are compared against state transition thresholds, and when reached, trigger control decisions. The process then repeats.

To simplify implementation, the parameters that determine the size of the region for training and the forecasting lead time are predefined and may be tied to the specifics of a particular system. Automatic parameter selection based on signal characteristics is possible. Neural networks with identical architecture are used. An option with automatic architecture selection is possible (for example, changing the number of neurons in the hidden layer or adding hidden neuron layers), depending on signal complexity, which can be assessed using any metric or their combination (variance, entropy, various dimensionality measures, etc.).

Based on the DLANN method, a multi-model method was developed that utilizes a model library (Dynamic Local Approximation by Neural Network models with Library

- DLANNLIB). In the previously described process, various neural network models are created, each associated with a specific state determined by the number of computational modules. Each model captures the pattern of a particular system state. These models, along with metadata (training quality metrics, data characteristics), are stored in the model library during operation and are subsequently used for forecasting when the system transitions to new states.

The set of models is saved and accumulated in the form of a library. Multiple models may exist for a single state. For each model, the training quality metric (validation error) and signal complexity assessment are stored. The model library can be used for predictions before training a new model in various ways (not all options are listed):

A. Using the model for the previous state. The simplest model selection method. No metrics of the current signal need to be calculated. If no model was built for the previous state, the predictor from library models is not used.

B. Selecting the best model for the state based on validation error. When the system transitions to one of the known states, the model with the lowest validation error is selected from the library.

C. Creating an ensemble from a set of models corresponding to one state. In this case, an ensemble (composition without training) is created from predictions of existing N models for one state, with weights related to validation errors.

D. Creating an ensemble from a set of models corresponding to different states. In this case, an ensemble (composition without training) is created from existing models for different states, with weights related to signal complexity.

All these options can be used during training of a new model, as an additional predictor when the model for the current signal is not yet ready.

Creating a predictor from a set of models can be considered as an independent task. As shown above, there can be quite many variants of this approach. In particular, one could use a separate neural network that would perform pattern classification and select appropriate models to create the predictor. Over time, the model library grows, and one could implement a forgetting mechanism that retains only the best models.

Using the model library improves forecasting responsiveness, as it enables obtaining control decisions before the model trained on data for a specific state is ready. This represents a step toward reactive-type responsiveness while maintaining the ability to utilize knowledge accumulated by the system during previous states.

The experimental results demonstrate that the proposed method improves the quality of control for the modeled IT service.

References

1. Singh, P. [et. al.] (2019). Research on auto-scaling of web applications in cloud: survey, trends and future directions. *Scalable Computing: Practice and Experience*. Vol. **20**, No. **2**. P. 399–432.

SUFFICIENT CONDITIONS FOR ASYMPTOTIC NORMALITY OF NUMBER OF MULTIPLE REPETITIONS OF CHAINS IN MARKED COMPLETE TREES AND FORESTS

V.I. Kruglov[1]

[1]*Steklov Mathematical Institute of Russian Academy of Sciences*
Moscow, RUSSIA
e-mail: [1]kruglov@mi-ras.ru

We consider complete q-ary trees of height H with vertices marked by random independent marks taking values from the set $\{1, 2, \ldots, N\}$ and forests of such trees. For both cases we investigate the number of sets of $r \geq 2$ paths with fixed length s such that corresponding s-chains of marks of vertices are identical. We propose three theorems on sufficient conditions for the asymptotic normality for considered random values as $H \to \infty$ and possibly varying parameters s and q.

Keywords: marked trees, forests of trees, chains of marks, repetitions of chains, conditions of asymptotic normality

1 Introduction

Studies of the probabilities of repetitions in sequences of independent random variables had started with investigations of repetitions of chains in sequences of random symbols (see, for example, [9], [4]). A natural development of these studies had led to problems associated with repetitions of sequences in trees with randomly marked vertices; problems of this kind arise in computer science (see [7] and [8]) in the analysis of algorithms or, for example, in connection with the tree structure of XML documents; such problems can also arise in connection with the construction of statistical criteria and the analysis of genetic sequences.

Poisson limit theorems for the number of coincidences of labels of chains in a binary or q-ary tree whose vertex marks are independent and have an equiprobable distribution over a finite alphabet were obtained in [10] and [3], and a Poisson limit theorem for the number of coincidences of sequences in a q-ary tree with equiprobable vertex labels was proved in [2].

In this talk we consider complete q-ary rooted trees of height H and forests composed of such rooted trees. Non-root vertices of trees are assigned random marks chosen independently from the set $\{1, 2, \ldots, N\}$ in accordance with some probability distribution. We consider such paths of fixed lengths on these trees that transitions on a path occur in the direction from the root of the corresponding tree.

In the first section of this talk we consider the number of sets of $r \geq 2$ paths on (one) tree that consist of the same number s vertices and for which the corresponding s-chains of vertex marks coincide. We obtain sufficient conditions for the asymptotic normality of this random variable for height $H \to \infty$.

In the second section we study repetitions of chains in a forest: it is assumed that there are r trees that can have different heights $H_1, \ldots, H_r$, vertices of these

trees are assigned independent random marks labels. The number of sets of paths of equal length s is studied, one path in each tree, for which the corresponding s-chains of vertex labels coincide, and sufficient conditions for asymptotic normality for this random variable are also obtained.

2 Repetitions of chains on a tree

Let $Tr(H)$ be a complete q-ary tree of height H. We will denote the root of the tree by the symbol $*$. Let for vertices of this tree be assigned random marks independently chosen from the set $\{1, 2, \ldots, N\}$ according to probabilities $p_1, \ldots, p_N$, where $p_1 + \ldots + p_N = 1$.

Consider such paths in the tree $Tr(H)$ that each path consists of s vertices, where $s < H$, and each next vertex of a path is connected by an edge to the previous vertex, so value s is the length of the path. Now we define random value $\xi_r(H, s)$ that is equal to the number of sets of r such ways on the tree $Tr(H)$ with equal values of s-chains of marks.

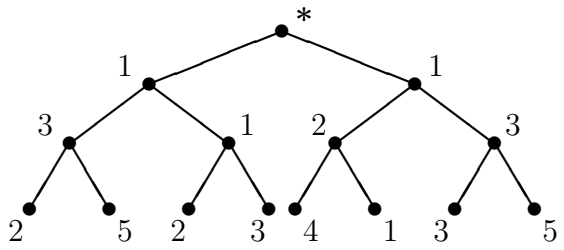
By $W(H, s)$ we denote the set of chains of length s in tree $Tr(H)$: the heigth of the first vertices of such chains is not greater than $H - s$. It is easy to show that

$$|W(H, s)| = \frac{q^{H-s+1} - 1}{q - 1} \cdot q^s = \frac{q^{H+1} - q^s}{q - 1}.$$

Let us enumerate elements of the set $W(H, s)$ by numbers from 1 to $|W(H, s)|$, so we denote by ω_u the path with the number u, where $1 \le u \le |W(H, s)|$, and s-chain of marks of vertices on this path we denote by $Y(\omega_u)$, thus

$$\xi_r(H, s) = \sum_{1 \le u_1 < \ldots < u_r \le |W(H,s)|} I\{Y(\omega_{u_1}) = \ldots = Y(\omega_{u_r})\}.$$

Example 1. In the tree below: $q = 2$, $H = 3$, $N = 5$. For $r = 2$, $s = 2$ we have $|W(H, s)| = 12$, $\xi_r(H, s) = 2$, and for $r = 2$, $s = 2$ we have $|W(H, s)| = 8$, $\xi_r(H, s) = 1$.



Theorem 1. *Let $H \to \infty$ and parameters $s = s(H)$ and $q = q(H)$ vary in such a way that $s/H \to 0$ and there exist such positive numbers C and $\varepsilon \in (0, 1]$ that for any sufficiently large H the following inequality holds:*

$$\mathbf{D}\xi_r(H, s) \ge C \left(\frac{q^{H+1} - q^s}{q - 1} \right)^{2(r-1)+\varepsilon}.$$

Then the distribution function and moments of random variable

$$\tilde{\xi}_r(H,s) = \frac{\xi_r(H,s) - \mathbf{E}\xi_r(H,s)}{\sqrt{\mathbf{D}\xi_r(H,s)}}$$

converge to distribution function and moments of standard normal distribution.

3 Repetitions of chains in forests of trees

Consider similar problem for the number of repetitions of chains in the set of trees. In this case formulae for expectation and variance of investigated random value have significantly simplier forms.

Let $Tr_1(H_1), \ldots, Tr_r(H_r)$ be full q-ary trees with roots of heights $H_1, \ldots, H_r$ respectively, let for vertices of these trees be assigned random marks that are independently chosen from the set $\{1, 2, \ldots, N\}$ according to probabilities $p_1, \ldots, p_N$, where $p_1 + \ldots + p_N = 1$.

We study random value $\xi_{(r)}(H_1, \ldots, H_r; s)$ which is equal to the number of such sets of r paths of length s that each path belongs to different tree and all paths have coinciding s-chains of marks of vertices:

$$\xi_{(r)}(H_1, \ldots, H_r; s) = \sum_{\omega_{u_1} \in W(H_1,s)} \ldots \sum_{\omega_{u_r} \in W(H_r,s)} I\{Y(\omega_{u_1}) = \ldots = Y(\omega_{u_r})\}.$$

Denote $H_{\min} = \min\{H_1, \ldots, H_r\}$. For any natural l define value $P_l = \sum_{k=1}^{N} p_k^l$, this value is equal to the probability that any l different vertices from the same tree or different trees have coinciding marks.

We propose the following sufficient conditions for asymptotic normality of this random variable.

Theorem 2. *Let $H_1, \ldots, H_r \to \infty$ and parameters $s = s(H_1, \ldots, H_r)$ and $q = q(H_1, \ldots, H_r)$ vary in such a way that $s/H_{\min} \to 0$ and there exist such positive numbers C and $\varepsilon \in (0, 1]$ that for any sufficiently large $H_{\min}$ the following inequality holds:*

$$\mathbf{D}\xi_{(r)}(H_1, \ldots, H_r; s) \geq C q^{2(H_1+\ldots+H_r)-(2-\varepsilon)H_{\min}}.$$

Then the distribution function and moments of random variable

$$\tilde{\xi}_{(r)}(H_1, \ldots, H_r; s) = \frac{\xi_{(r)}(H_1, \ldots, H_r; s) - P_r^s \prod_{k=1}^{r} \frac{q^{H_k+1}-q^s}{q-1}}{\sqrt{\mathbf{D}\xi_{(r)}(H_1, \ldots, H_r; s)}}$$

converge to distribution function and moments of standard normal distribution.

Theorem 3. *Let the distribution defined by probabilities $p_1, \ldots, p_N$ differ from the equiprobable distribution on set $\{1, 2, \ldots, N\}$. Let $H_1 = \ldots = H_r = H$, let $H \to \infty$ and parameters $s = s(H)$ and $q = q(H)$ vary in such a way that $s/H \to 0$.*

Then exists such $C \in (0, \infty)$ *that for* $H \to \infty$

$$\mathbf{D}\xi_{(r)}(H, \ldots, H; s) = Cq^{(2r-1)H}(1 + o(1))$$

and the distribution function and moments of random variable

$$\tilde{\xi}_{(r)}(H, \ldots, H; s) = \frac{\xi_{(r)}(H, \ldots, H; s) - P_r^s \left(\frac{q^{H+1}-q^s}{q-1}\right)^r}{\sqrt{\mathbf{D}\xi_{(r)}(H, \ldots, H; s)}}$$

converge to distribution function and moments of standard normal distribution.

Theorems 1, 2 and 3 had been proven in [6]. Proofs of all presented theorems are based on Janson's method [1] in the form, proposed by V.G. Mikhailov in [5].

References

1. Janson S. (1988). Normal convergence by higher semiinvariants with applications to sums of dependent random variables and random graphs. *Ann. Probab.* Vol. **16**, Num. **1**, pp. 306-312.

2. Kruglov V., Zubkov A. (2017). Number of pairs of template matchings in q-ary tree with randomly marked vertices. *Analytical and Computational Methods in Probability Theory, Lecture Notes in Comput. Sci.* Vol. **10684**, pp. 336-346.

3. Kruglov V.I. (2018). On coincidences of tuples in a q-ary tree with random labels of vertices. *Discrete Math. Appl.* Vol. **28**, Num. **5**, pp. 293-307.

4. Mikhailov V. G. (2002). Estimate of the accuracy of the compound poisson approximation for the distribution of the number of matching patterns. *Theory Probab. Appl.* Vol. **46**, Num. **4**, pp. 667-675.

5. Mikhailov V. G. (1991). On a theorem of Janson. *Theory Probab. Appl.* Vol. **36**, Num. **1**, pp. 173-176.

6. Mikhailov V. G., Kruglov V.I. (2023). Conditions for asymptotic normality of number of multiple repetitions of chains in marked complete trees and forests. *Mathematical Aspects of Cryptography.* Vol. **14**, Num. **1**, pp. 85-97.

7. Singh G., Smolka S.A., Ramakrishnan I.V. (1988). Distributed algorithms for tree pattern matching. *Lect. Notes Comput. Sci.* Vol. **312**, pp. 92-107.

8. Tahraoui M.A., Pinel-Sauvagnat K., Laitang C., Boughanem M., Kheddouci H., Ning L. (2013). A survey on tree matching and XML retrieval. *Computer Science Review.* Vol. **8**, pp. 1-23.

9. Zubkov A.M., Mikhailov V. G. (1974). Limit distributions of random variables connected with long duplications in a sequence of independent trials. *Theory Probab. Appl.* Vol. **19**, Num. **1**, pp. 172-179.

10. Zubkov A.M., Kruglov V.I. (2016). On coincidences of tuples in a binary tree with random labels of vertices. *Discrete Math. Appl.* Vol. **26**, Num. **3**, pp. 145-153.

SECTORAL STRUCTURE AND PROFITABILITY OF GRP: REGIONS OF RUSSIAN FEDERATION

A.V. Kudrov[1]

[1] *Central Economics and Mathematics Institute of the Russian Academy of Sciences*
Moscow, RUSSIA
e-mail: [1]kovlal@inbox.ru

The article will analyze the factors influencing the Gross Regional Product (GRP) and the profitability of the economies of Russian regions. These factors are represented by the main economic characteristics of the regional economy, reflecting its structure, level of development, capital-labor ratio, etc. The analysis methodology involves two stages: first, identifying variables directly related to GRP and the profitability of the gross regional product; second, constructing a nonlinear regression model using the identified directly related variables as explanatory factors. As a result, a sufficiently accurate nonlinear regression model was obtained for the profitability of the regional economy. This model includes indices for the extractive and manufacturing industries, as well as GRP per capita. The model allows us to identify the conditions affecting the industrial indices and GRP per capita under which profitability will decrease. This demonstrates that at different stages of economic development, maximizing profitability requires balancing the industrial structure and the size of the regional economy. In particular, in highly developed regions, the development of traditional production sectors is associated with diminishing profitability. The results show that when focusing on the profitability of the regional economy, economic strategies adaptive to specific regional characteristics are necessary.

Keywords: Gross Regional Product (GRP), regional economy profitability, extractive and manufacturing industries indices, economic complexity, nonlinear regression

1 Introduction

Profitability of the GRP, measured as the ratio of net profit earned by companies, enterprises, and other organizations registered in a region to its Gross Regional Product (GRP), is a crucial indicator of a region's economic health. The advantage of this metric compared to commonly used characteristics in regional economic studies (such as GRP growth rates or GRP per capita) lies in the fact that GRP profitability reflects the final economic outcome (profit), rather than intermediate measures of economic activity.

The interrelationship between profitability and a regional economy's structure, its capital-labor ratio, scale, and other socio-economic factors represents a highly relevant research direction in economics, especially amidst rapid structural transformation.

Economists reached a consensus relatively long ago that a country's ability to generate and distribute income depends on its production structure, as discussed in works such as [1-3]. However, using only indices of sectoral specialization as characteristics

of a regional economy's structure is insufficient. It is necessary to account for the interconnections between different types of economic activity, the complexity, and the sophistication of production activities. To quantify the degree of interconnectedness among various economic sectors within an economy's structure, the Economic Complexity Index (ECI) was proposed, see [4–6]. This report will present a modified Economic Complexity Index capable of capturing the most significant interdependencies.

High values of the Economic Complexity Index indicate that the economy's structure is dominated by interconnected sectors. For instance, industries with extended production cycles, such as electronics and machinery manufacturing, require a higher level of coordination and knowledge, thus exhibiting high economic complexity. Conversely, economic structures dominated by raw material and agricultural sectors yield low economic complexity values. The relationship between economic complexity and GRP is nonlinear, see Fig. 1.

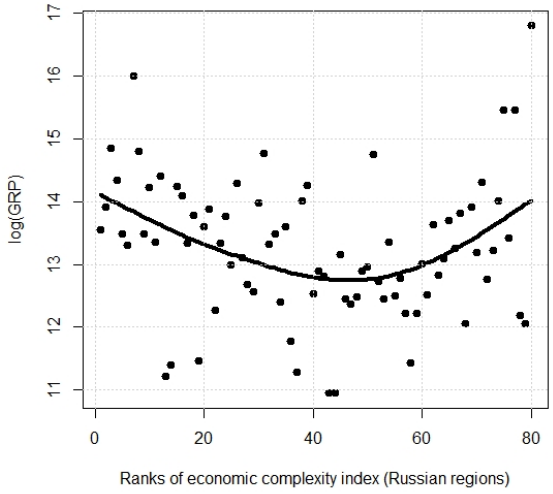


Figure 1: GRP and ranks of economic complexity (Russian regions)

The U-shaped relationship between economic complexity and GRP implies that both very low and very high levels of economic complexity correspond to high GRP, whereas a medium level of economic complexity corresponds to lower GRP values. Thus, according to Fig. 1, two distinct pathways to higher GRP can be identified: 1. through natural resource extraction, or 2. through developing a more complex industrial economy.

It should be noted that a statistically significant direct relationship between GRP and economic complexity (when controlling for sectoral specialization indices) is observed only at high complexity levels exceeding a certain threshold (for the concept of direct relationship, see [7-8]).

2 Model Specification and Results

Based on this identified threshold-based direct relationship, the conceptualization of an extended production function for GRP has been generalized, see [7]:

$$Y = c \cdot K^{\beta_1(S_1,S_2)} L^{\beta_2(S_1,S_2,T)} P^{\gamma} + \epsilon, \tag{1}$$

where

$$\beta_1(S_1, S_2) = \frac{\mu_1 e^{(\mu_2 \cdot S_1 + \mu_3 \cdot S_2)}}{1 + \mu_1 e^{(\mu_2 \cdot S_1 + \mu_3 \cdot S_2)}}, \quad \beta_2(S_1, S_2, T) = \frac{\lambda_1 e^{\left(\lambda_2 \cdot S_1 + \lambda_3 \cdot S_2 + \lambda_4 \cdot T^2\right)}}{1 + \lambda_1 e^{(\lambda_2 \cdot S_1 + \lambda_3 \cdot S_2 + \lambda_4 \cdot T^2)}};$$

$$T = \begin{cases} ECI & \text{if } ECI \geq 0.45 \\ 0 & \text{otherwise} \end{cases};$$

$c, \gamma, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \mu_1, \mu_2, \mu_3$ - parameters; Y - gross regional product; K - capital stock (fixed assets); L - average annual number of employees; P - number of researchers; ECI - economic complexity index; S_1 and S_2 - indices of sectoral specialization; ϵ - errors of model (1).

Table 1: Parameter Estimates for GRP Model (1)

Parameter	Estimate	Std. Error	t-value	p-value
c	6.77	0.42	4.53	0.00
μ_1	1.79	0.21	2.72	0.01
μ_2 (extractive)	0.01	0.00	3.53	0.00
μ_3 (manufacturing)	-0.02	0.01	-3.68	0.00
λ_1	0.33	0.26	-4.35	0.00
λ_2 (extractive)	-0.01	0.01	-1.96	0.05
λ_3 (manufacturing)	0.05	0.01	3.83	0.00
λ_4 (complexity)	3.34	1.16	2.89	0.01
γ (researchers)	0.05	0.02	2.81	0.01

As seen in 2, decreasing returns to scale are characteristic of regions with high concentrations of extractive industries in their economic structure. Decreasing returns to scale imply that a proportional increase in labor and capital results in a less-than-proportional increase in output. This may be attributed to the fact that extractive industries (e.g., mining, oil and gas) are often capital-intensive and may face challenges such as resource depletion, environmental regulations, or high operational costs.

Fig. 3 demonstrates that increasing returns to scale are typical for regions with substantial manufacturing sector concentration and high economic complexity values. Economic complexity levels exceeding the threshold of 0.45 correspond to higher returns to scale.

A nonlinear regression model has been developed for regional GRP profitability, capturing the influence of economic structure and regional development level (GRP per capita), see [8]:

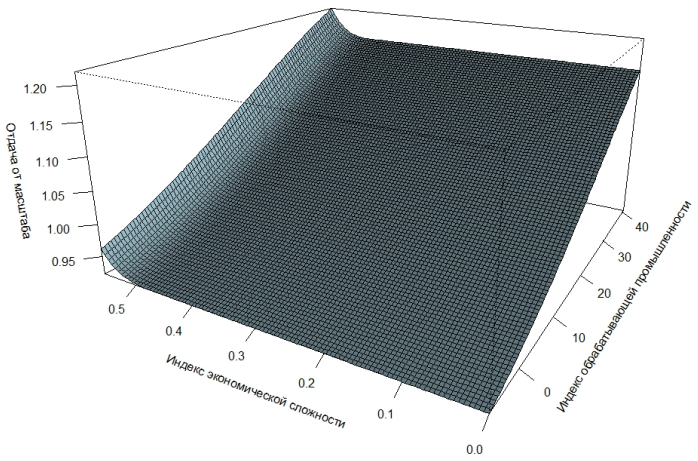


Figure 2: Returns to Scale: Economic Complexity Index vs. Manufacturing Sectors index

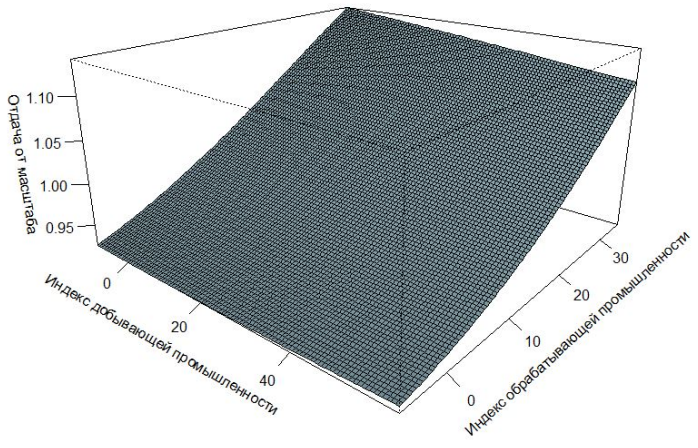


Figure 3: Returns to Scale: Extractive Sector vs. Manufacturing Sector

$$\frac{NI}{Y} = \frac{\mu^{(1)} e^{\left(\mu^{(2)} \cdot S_1 + \mu^{(3)} \cdot S_2\right)} \left(\frac{Y}{P}\right)^{\left(\mu^{(4)} + \mu^{(5)} \cdot S_1 + \mu^{(6)} \cdot S_2\right)}}{1 + \mu^{(1)} e^{\left(\mu^{(2)} \cdot S_1 + \mu^{(3)} \cdot S_2\right)} \left(\frac{Y}{P}\right)^{\left(\mu^{(4)} + \mu^{(5)} \cdot S_1 + \mu^{(6)} \cdot S_2\right)}} + \epsilon_{NI}, \tag{2}$$

where $\mu^{(1)}, \mu^{(2)}, \mu^{(3)}, \mu^{(4)}, \mu^{(5)}, \mu^{(6)}$ - parameters; NI - net income; S_1 – extractive sectors index, S_2 – manufacturing sectors index; Y/P - GRP per capita; ϵ_{NI} - errors.

Table 2: Parameter Estimates for GRP Profitability Model (2)

Parameter	Estimate	Std. Error	t-value	p-value
$\mu^{(1)}$	0.002	0.83	9.96	0.00
$\mu^{(2)}$	0.13	0.03	7.11	0.00
$\mu^{(3)}$	0.11	0.06	2.50	0.01
$\mu^{(4)}$	0.70	0.14	7.45	0.00
$\mu^{(5)}$	-0.02	0.00	-7.35	0.00
$\mu^{(6)}$	-0.02	0.01	-2.23	0.03

The model demonstrates high accuracy (concordance correlation coefficient equals 0.95). Based on Model (2), constraints were derived for GRP per capita and extractive/manufacturing sector indices that are associated with declining GRP profitability.

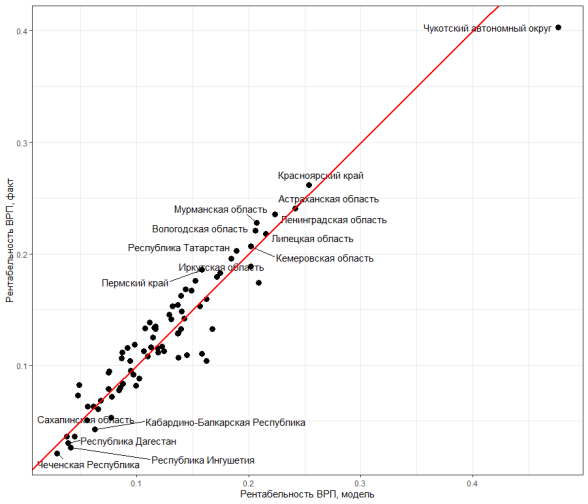


Figure 4: Comparison of Actual GRP Profitability Values vs. Model (2) Estimates

- The expression $\left(\mu^{(4)} + \mu^{(5)} S_1 + \mu^{(6)} S_2\right)$ determines whether an increase in GRP per capita leads to growth or decline in profitability. If this value is positive, profitability tends to increase with rising GRP per capita, whereas if negative, it tends to decrease.

- GRP profitability increases with growth in S_1 (extractive sectors index) at fixed GRP per capita (Y/P), when $\mu^{(5)} \log\left(\frac{Y}{P}\right) + \mu^{(2)} > 0$, and decreases when $\mu^{(5)} \log\left(\frac{Y}{P}\right) + \mu^{(2)} < 0$.

- GRP profitability increases with growth in S_2 (manufacturing sectors index) at fixed GRP per capita (Y/P) when $\mu^{(6)} \log\left(\frac{Y}{P}\right) + \mu^{(3)} > 0$, and decreases when $\mu^{(6)} \log\left(\frac{Y}{P}\right) + \mu^{(3)} < 0$.

Fig. 5 illustrates changes in GRP profitability estimates based on Model (2) across different combinations of the extractive sector index and manufacturing sector index, assuming a region with mean GRP per capita.

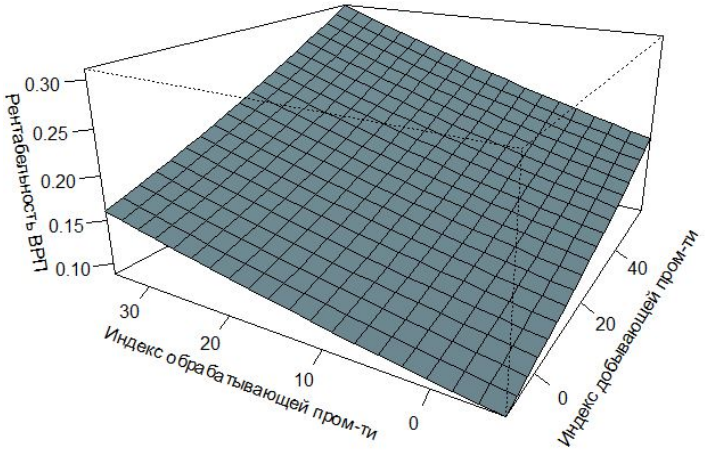


Figure 5: Dependence of GRP Profitability on Extractive and Manufacturing Sectors Indices According to Model (2) (for Mean GRP per Capita)

As seen from Fig. 5, the profitability estimate demonstrates an upward tendency with increases in both the extractive sector index and manufacturing sector index. Notably, the growth in GRP profitability is more pronounced in response to the extractive sector index compared to the manufacturing sector index. The highest GRP profitability estimates occur when both indices reach maximum values, indicating that deeper mineral processing substantially enhances GRP profitability.

Fig. 6 shows the relationship between the logarithm of GRP per capita, the extractive sector index, and GRP profitability when manufacturing sectors index equals to zero ($S_2 = 0$).

In regions characterized by low GRP per capita, there is a strong positive relationship between the extractive sectors index and GRP profitability. This indicates that developing new deposits significantly enhances regional GRP profitability. However, as easily accessible mineral deposits are depleted, substantial investments in more complex extraction become necessary, potentially leading to declining GRP profitability. This tendency is shown in Fig. 6, where at high values of the extractive sectors index, increases in the logarithm of GRP per capita correspond to reduced profitability.

Regions with low GRP per capita also exhibit a positive relationship between the manufacturing sectors index and GRP profitability (Fig. 7). This suggests that in less developed regions, expanding manufacturing industries contributes to higher GRP profitability. In regions with high GRP per capita, the correspondence between the manufacturing sector index and profitability is less pronounced. At a given GRP level, profitability increases much more slowly as the manufacturing sector index grows.

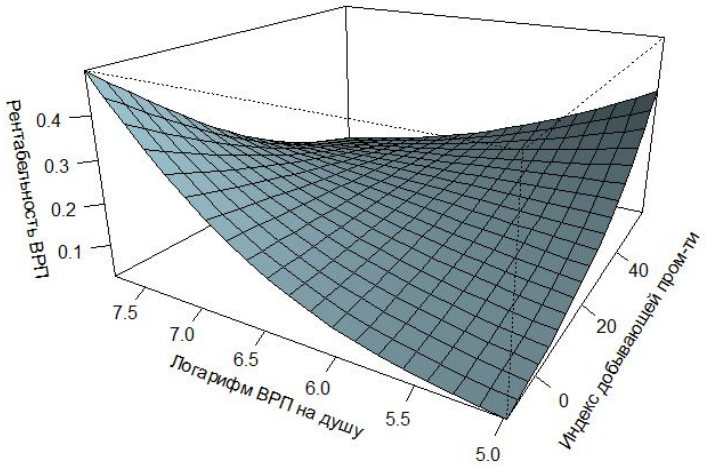


Figure 6: Dependence of GRP Profitability on the Extractive Sectors Index and GRP per Capita According to Model (2) (when Manufacturing Sectors Index equals to zero)

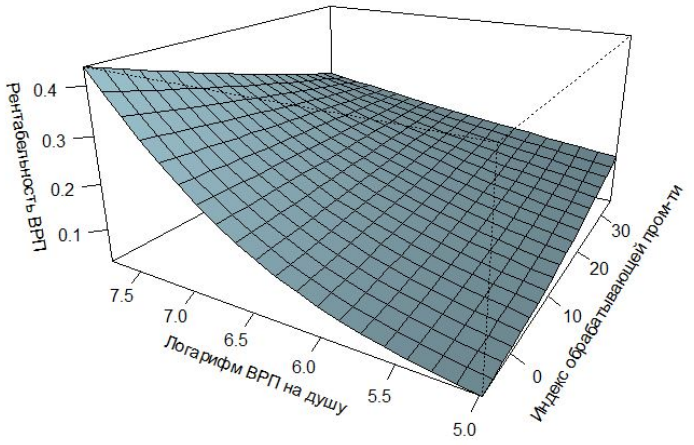


Figure 7: Dependence of GRP Profitability on the Manufacturing Sector Index and GRP per Capita According to Model (2) (when Extractive Sectors Index equals to zero)

Moreover, regions with high GRP per capita and low values of both extractive and manufacturing sectors indices demonstrate the highest levels of GRP profitability (Figs. 6 and 7). Economic activity in these regions is predominantly oriented toward construction, trade, and services – including sectors such as information-communication, finance, and other non-production subsectors characterized by high profitability. This economic structure combined with high GRP per capita is typically found in major metropolitan areas.

3 Conclusion

The key substantive economic findings include:

- Development of a nonlinear regression model for regional GRP profitability, capturing the influence of economic structure and regional development level (GRP per capita).
- Identification of constraints for GRP per capita and extractive/manufacturing sector indices associated with declining GRP profitability. This underscores the critical importance of balancing economic development and economic structure.
- In highly developed regions, further expansion of traditional extractive and manufacturing industries may prove ineffective for boosting GRP profitability. Such regions should prioritize transitioning toward service-oriented and knowledge-intensive sectors. This aligns with the observed tendency of advanced industrial economies shifting toward service- and knowledge-based industries.
- In economically underdeveloped regions, establishing and expanding traditional extractive and manufacturing industries remains an effective pathway to enhance GRP profitability.
- Maximizing profitability requires an optimal balance between GRP per capita and sectoral structure. This equilibrium evolves with economic development. Excessive concentration in either extractive or manufacturing sectors can reduce GRP profitability.

Based on these findings, we propose recommendations for a more flexible tax system accounting for regional economic structures and development levels to foster nationwide economic growth and reduce regional disparities:

- Regions with strong manufacturing sectors should implement/expand R&D tax incentives to stimulate innovation, technological advancement, and emerging knowledge-intensive industries.
- Regions with higher GRP per capita but low industrial share could sustain moderately higher tax rates, while less developed regions would benefit from lower rates to accelerate growth.

- Introduce special tax provisions for major metropolitan areas that account for agglomeration advantages.
- Regions with low capital-labor ratios require both accelerated depreciation and investment tax credits to incentivize capital accumulation and productivity gains.

In conclusion, our probabilistic-statistical analysis of regional economic profitability reaffirms the significance of adaptive economic strategies that evolve across development stages. The results highlight the need for regionally tailored policies rather than universal approaches. Strategies effective for boosting GRP profitability in less developed regions may prove inefficient or even counterproductive in more advanced economies.

References

1. Hirschman A.O. (1958) *The strategy of economic development*. New Haven: Yale Univ. Press.
2. Rosenstein-Rodan P.N. (1943) Problems of industrialization of eastern and south-eastern Europe. *The Economic Journal*, vol. 53(210/211), pp. 202–211.
3. Teece D., Rumelt R., Dosi G., Winter S. (1994) Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior & Organization*, vol. 23(1), pp. 1–30.
4. Hausmann R., Hidalgo C.A., Bustos S., Coscia M., Chung S., Jimenez J., Simoes A. (2007) The Building Blocks of Economic Complexity. *PLoS One*, vol. 2(1), e268.
5. Hausmann R., Hidalgo C.A. (2017) *Atlas of economic complexity: Mapping paths to prosperity*. MIT Press.
6. Hidalgo C.A., Hausmann R. (2009) The building blocks of economic complexity. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367(1897), pp. 1817–1825.
7. Kudrov, A. V. The impact of economic complexity and industry specialization on the gross regional product of Russian regions // Business Informatics. – 2023. – Vol. 17, No. 4. – P. 25-40. – DOI 10.17323/2587-814X.2023.4.25.40.
8. Kudrov A.V. (2024). Impact of the sectoral structure of regional economy profitability on GRP. *Economics of Contemporary Russia*, No. 4 (107), . 60–76. DOI 10.33293/1609-1442-2024-4(107)-60-76

ON USING OF ARTIFICIAL NEURAL NETWORKS FOR APPROXIMATION OF BINARY FUNCTIONS

K.V. Latushkin[1], Yu.S. Kharin[2]

[1,2]*Research Institute for Applied Problems of Mathematics and Informatics*
[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]LatushkinKV@bsu.by, [2]Kharin@bsu.by

The paper examines the properties of using artificial neural networks with one hidden layer in problems of approximating binary functions of many binary variables.

Keywords: artificial neural network, binary function, combinatorics, function approximation, pseudo-random sequence generators

1 Introduction

In recent years, artificial neural networks have begun to be widely used in cryptology and cybersecurity problems [1, 2, 3, 4]. Examples of such problems are: approximation of discrete functions in software pseudo-random sequence generators, assessment of the quality of s-blocks and other cryptographic primitives; recognition of computer attacks on information systems. Mathematically, these problems are reduced to the problem of approximation of binary functions of many binary variables. This publication is devoted to the study of the properties of this topical problem.

2 Mathematical model and problem statement

Let us introduce the following notation: $V = \{0, 1\}$ is the binary alphabet; s is a natural number; V^s is an s-dimensional binary hypercube; $x = (x_1, x_2, , x_s)^{'} \in V^S$ is a binary column vector or, in geometric interpretation, some vertex of the hypercube V^s; $\mathbb{1}\{B\} \in V$ is the indicator of event B, $\mathbb{1}\{B\} = \{1$ if B is true; 0 otherwise$\}$.

On the set V^s, some unknown binary function of s binary variables is defined:

$$y = f(x) = f(x_1, \ldots, x_s), \quad x \in V^s, \quad y \in V. \tag{1}$$

Let us consider the problem of approximating (restoring) the function (1) using a random sample of size n from V^s: $X = \left\{x^{(1)}, x^{(2)}, , x^{(n)}\right\} \subseteq V^s$ and known values $y^{(t)} = f(x^{(t)})$, $t = 1, , n$. To solve the problem, we use a two-layer (with s inputs, one hidden layer with m neurons and one output) artificial neural network (ANN), which is defined by the function $\hat{f}(x) = \hat{f}(x_1, x_2, , x_s) \in V$ with $N = m(s+2)+1$ parameters estimated using the sample X:

$$\hat{f}(x) = \sigma\left(b_0 + \sum_{j=1}^{m} b_j ReLU\left(a_{0j} + \sum_{i=1}^{s} a_{ij} x_i\right)\right), \tag{2}$$

here m is a natural number (the number of neurons in the hidden layer), $\{a_{il}\}$, $\{b_l\}$ are the parameters (coefficients, weights) of the model; $ReLU(z) = \max\{0, z\}$, $\sigma(z) = (1 + e^{-z})^{-1}$ are the so-called activation functions [5]. By the l-th neuron of model (2) we will understand the function $H_l(x) = ReLU\left(a_{0l} + \sum_{i=1}^{s} a_{il}x_i\right)$, $l \in \{1, ..., m\}$.

The approximation of $f(\cdot)$ using $\hat{f}(\cdot)$ is obtained as a result of the ANN training process (2), which consists of minimizing the loss function over $\{a_{il}\}$, $\{b_l\}$:

$$h(\hat{y}) = -\frac{1}{n}\sum_{t=1}^{n}\left(y^{(t)} log\left(\hat{y}^{(t)}\right) + \left(1 - y^{(t)}\right) log\left(1 - \hat{y}^{(t)}\right)\right) \to \min_{\{a_{il}\},\{b_l\}}. \tag{3}$$

The loss function $h(\hat{y})$ in (3) is usually called the "binary cross-entropy" [5]. In formula (3), the value $\hat{y}^{(t)} = \hat{f}\left(x^{(t)}\right)$ is the estimate for $y^{(t)}$ obtained during the training process, and "accuracy" is used to estimate the accuracy of training the proportion of correctly classified vertices:

$$\alpha = accuracy = \frac{1}{n}\sum_{t=1}^{n}\mathbb{1}\left\{\hat{y}^{(t)} = y^{(t)}\right\} \in [0, 1].$$

3 Properties of using ANN

When approximating binary functions of many binary variables based on ANN (2), two important properties arise:

1. a property associated with the presence of regions of piecewise constancy and multimodality of the objective function (3);
2. a property associated with the choice of the number of neurons m.

3.1 On piecewise constancy and multimodality of loss function

Lemma 1. *If the coefficients $\{a_{il}\}$ of some l-th neuron of the ANN (2) on the set of vertices X satisfy the condition*

$$a_{0l} + \sum_{i=1}^{s} a_{il}x_i < 0, \quad \forall x \in X,$$

then the derivative of the function $h(\cdot)$, defined by (3), with respect to the coefficients $\{a_{il}, b_l\}$ on the set X is equal to 0.

It follows from Lemma 1 that on the set of ANN parameters

$$A_l = \left\{(a_{0l}, a_{1l}, ..., a_{sl}) : a_{0l} + \sum_{i=1}^{s} a_{il}x_i < 0, \forall x \in X\right\} \subset R^{s+1}$$

the objective function is piecewise constant with respect to the parameters of the l-th neuron. On such sets of piecewise constancy, the use of gradient descent to solve the

minimization problem (3) leads to worse convergence. Another difficulty in solving problem (3) is the multimodality of the objective function in (3). To overcome these difficulties, it is proposed to specially select the initial values of the parameters $\{a_{il}, b_l\}$ as follows. First, we generate them as random variables according to [6, 7] from the probability distributions:

$$a_{il} \sim \mathcal{N}\left(0, \frac{2}{s}\right), \quad b_l \sim U\left[-\frac{\sqrt{6}}{\sqrt{m+1}}, \frac{\sqrt{6}}{\sqrt{m+1}}\right],$$

where $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mathematical expectation μ and variance σ^2; $U[a, b]$ is a uniform distribution defined on the interval $[a, b]$. If $(a_{0l}, a_{1l}, ..., a_{sl}) \in A_l$, then for the parameters $\{a_{0l}, ..., a_{sl}\}$ we repeat the generation until we obtain initial values outside the region of piecewise constancy A_l.

3.2 On evaluation of the number of neurons

When solving the problem of approximating the function $f(\cdot)$ (1), for each vertex $x \in X \subseteq V^s$ it is sufficient to use one neuron of the hidden layer. Therefore

$$1 \leq m \leq |X| \leq |V^s| = 2^s.$$

In the following we assume that $X = V^s$. In this case, the number of all possible different binary functions for approximation is limited and equals 2^{2^s}. Let $W(s, m)$ be the number of different binary functions for the approximation of which m neurons in the hidden layer of the ANN (2) are necessary and sufficient. It follows that

$$\sum_{m=1}^{2^s} W(s, m) = 2^{2^s}.$$

Let us construct a lower bound for the number $W(s, 1)$, i.e. the number of different functions defined on an s-dimensional hypercube, for the approximation of which an ANN (2) with one neuron on the hidden layer is sufficient.

Using combinatorics, the following lemmas are proved.

Lemma 2. *Any s-dimensional hypercube contains $2^{s-l}C_s^l$ faces of dimension l.*

Lemma 3. *Let A, B be two finite sets, k be some non-negative integer, $C_k(*)$ be the number of different ways to choose k elements from the set $*$. Then the inequality $C_k(A \cup B) \geq C_k(A) + C_k(B) - C_k(A \cap B)$ is true.*

Let us agree to say that k vertices ($1 \leq k \leq 2^s - 1$) are linearly distinguishable in a hypercube V^s if there exists an $(s-1)$-dimensional hyperplane that partitions V^s into two disjoint subsets of k and $2^s - k$ vertices.

Lemma 4. *If k vertices are linearly distinguishable on some face of a hypercube, then the same k vertices can be linearly distinguished in the entire hypercube.*

Based on Lemmas 2-4, the following recurrence relation is constructed:

$$\begin{cases} Q(s,k) = \sum_{i=1}^{s} (-1)^{i+1} 2^i C_s^{s-i} Q(s-i,k), \text{if } k \le 2^{s-1}, \\ Q(s,k) = 0, \text{if } k > 2^s, \\ Q(s,k) = Q(s,2^s-k), \text{if } k > 2^{s-1}, \\ Q(s,0) = Q(s,2^s) = 1. \end{cases} \tag{4}$$

Theorem 1. *The number $Q(s,k)$ obtained from the recurrence relation (4) is a lower bound for the number of ways to linearly select k vertices in an s-dimensional hypercube.*

Corollary 1. *To find a lower bound for the number $W(s,1)$, it is necessary to sum the values $Q(s,k)$ over k:*

$$W(s,1) \ge \sum_{k=0}^{2^s} Q(s,k).$$

4 Application of ANN to approximate the generating function of pseudorandom sequence generators

Let us consider the problem of approximating the generating functions of pseudorandom sequence generators based on linear feedback shift registers (LFSRs) and nonlinear feedback shift registers (NLFSRs), defined by the following general recurrence relation:

$$x_\tau = f\left(x_{\tau-1}, x_{\tau-2}, ..., x_{\tau-s}\right), \quad \tau > s.$$

To apply the ANN (2) the training sample was formed as follows:

$$x^{(t)} ::= \left(x_{t-1}, ..., x_{t-s}\right), \quad y^{(t)} ::= x_t, \quad t \ge s+1.$$

Two LFSRs [8] and six NLFSRs [9] were studied. For each of them, the smallest number of neurons m_{min} required for error-free approximation ($\alpha = 1$) was found. The results are presented in Table 1.

Table 1: Approximation of generating functions of LFSR and NLFSR

Num. variables s	Function f	Num. neurons m_{min}
7	$x_1 \oplus x_5$	2
15	$x_1 \oplus x_9$	2
17	$x_1 \oplus x_2 \oplus x_8 x_{11} \oplus x_{10} x_{16}$	6
17	$x_1 \oplus x_7 \oplus x_3 x_{10} \oplus x_8 x_{13}$	6
17	$x_1 \oplus x_2 \oplus x_4 \oplus x_{10} \oplus x_{13} \oplus x_8 x_{14}$	7
17	$x_1 \oplus x_2 \oplus x_8 \oplus x_{12} \oplus x_{14} \oplus x_7 x_{15}$	7
17	$x_1 \oplus x_4 \oplus x_9 \oplus x_{12} \oplus x_{13} \oplus x_4 x_{12}$	7
24	$x_1 \oplus x_2 \oplus x_9 \oplus x_{10} \oplus x_{16} \oplus x_8 x_{19}$	7

5 Conclusion

For the selected model, the regions of piecewise constancy of the loss function "binary cross-entropy" with respect to the coefficient are found. A lower bound is constructed for the number of functions that can be approximated by an ANN with one neuron on the hidden layer. Approximation of some generating functions of pseudorandom sequence generators (LFSR and NLFSR) is considered.

References

1. Gohr, A. (2019). Improving attacks on round-reduced speck32/64 using deep learning. *Advances in cryptology, CRYPTO-2019*. P. 150–179.

2. Picek, S. [et. al.] (2023). Deep learning-based physical side-channel analysis. *ACM Computing Surveys*. Vol. **55**, No. **11**. P. 1–35.

3. Boanca, S. (2024). Exploring patterns and assessing the security of pseudorandom number generators with machine learning. *Int. Conf. Agents and Artificial Intelligence*. Vol. **3**. P. 186–193.

4. Betelin, V.B., Galkin, V.A. (2021). Mathematical problems of artificial intelligence and artificial neural networks. *Advances in Cybernetics*. Vol. **2**, No. **4**. P. 6–14. (In Russian)

5. Nikolenko, S., Kadurin, A., Arkhangelskaya, E. (2018). *Deep Learning. Dive into the world of neural networks*. Piter: Saint Petersburg. (In Russian)

6. Glorot, X., Bengio, Y. (2010). Understanding the Difficulty of Training Deep Feedforward Neural Networks. *Int. Conf. Artificial Intelligence and Statistics*. P. 249–256.

7. Kaiming, H. [et. al.] (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proc. ICCV-2015*. P. 1026–1034.

8. Kharin, Yu.S. [et. al.] (2023). *Cryptology: textbook*. BSU: Minsk. (In Russian)

9. Dubrova, E. (2012). *A List of Maximum Period NLFSRs*. Cryptology ePrint Archive, Report 2012/166.

STATISTICAL FORECASTING OF PANEL DATA BASED ON STATE SPACE MODELS

V.I. Lobach[1], S.V. Lobach[2]

[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]Lobach@bsu.by

Panel (or longitudinal) data describes a set of objects which are observed during certain period of time, so they consist of repeated observations of the same objects in sequential time periods. The following examples of panel data can be mentioned: annual household studies, monthly performance indicators for economic institutions and many others. In this study we provide another approach to forecasting cross-sectional data based on state space models together with Kalman filtering procedure.

Keywords: forecasting, panel data, state space model, Kalman filtering

1 Introduction

In economic researches regression models are widely used within large number of applications [1]. Regression models for panel data allow usage of two indices to describe the data: $y_{i,t} = \alpha + X_{i,t}\beta + x_{i,t}$, where i defines object index (household, company, country, etc), t denotes timestamp of an observation, α is an unknown intercept, β is a $(n \times 1)$-vector of unknown parameters, $X_{i,t}$ is a known matrix denoting factors which influence observations. Uncontrollable factors $x_{i,t}$ are modeled with the following equation: $x_{i,t} = \mu_i + \epsilon_{i,t}$, where μ_i is an unobservable individual effect of $i-$th object, $\epsilon_{i,t}$ is a random variable which defines random uncontrollable effect.

Statistical analysis of panel data can be carried out using state space models. In order to express panel data in a state space form it is necessary to introduce one more index i for state parameters vector x_t in classic state space model formulation. This resutls in $x_{i,t}$, where $t = 1, \cdots, T_i$, $i = 1, \cdots, K$, t denotes timestamp, i denotes object index. It means that the mathematical model for panel data is a random field $\{x_{i,t}\}$, $t \in \{1, \ldots, T_i\}$, $i \in \{1, \ldots, K\}$.

Based on linear state space models [2] we express panel data in a state space form: $x_{i,t} = Fx_{i,t-1} + \omega_{1,t}$, $y_{i,t} = Hx_{i,t} + \omega_{2,t}$, where $x_{i,t}$ is an unobserved state of i-th object at moment t, $y_{i,t}$ is an observation for the object at the same moment. In common case $x_{i,t} \in R_1^n$, $y_{i,t} \in R_2^n$, $\{\omega_{1,t}\}$ and $\{\omega_{2,t}\}$ are sequences of i.i.d. random variables $\omega_{1,t}$, $\omega_{2,t} \sim \mathcal{N}(0, Q)$, $x_{i,0} \sim \mathcal{N}(\mu, P)$. The parameters of the model are F, H, μ, P. And the problem is to estimate future observations $x_{i,t+h}$, $y_{i,t+h}$ based on previous observations $y_{i,s}$, $s = 1, \cdots, t$, $h > 0$.

2 Kalman Filter

Kalman Filter [3] allows to build optimal in mean-squared sense forecasts if they are introduced in linear state space form. Let us consider the following $x^t_{i,t} = E\{x_{i,t}|y^t_{i,0}\}$, $P^t_{i,t_1,t_2} = E\{(x_{i,t_1} - x^t_{i,t_1})(x_{i,t_2} - x^t_{i,t_2})|y^t_{i,0}\}$, where $y^t_{i,0} = \{y_{i,j}, j = 1, \ldots, t\}$.

Kalman Filter can be expressed using the following equations [3, 2]

$$x^t_{i,t} = x^{t-1}_{i,t} + K_{i,t}(y_{i,t} - H_{i,t}x^{t-1}_{i,t}, \quad P^t_{i,t} = (1 - K_{i,t}H_{i,t})P^{t-1}_{i,t}, \tag{1}$$

$$K_{i,t} = P^{t-1}_{i,t}H^T_{i,t}(H_{i,t}P^{t-1}_{i,t}H^T_{i,t} + R)^{-1}, \tag{2}$$

where $i \in \{1, \ldots, K\}$, $t \in \{1, \ldots, T_i\}$, $x_{i,0} = \mu$, $P^0_{i,0} = P$.

In order to compute forecasts for $x_{i,t}$ for h lags forward equations (1)–(2) are used with initial values $x^T_{i,t}$, $P^T_{i,t}$ instead of $x^0_{i,0}$, $P^0_{i,0}$.

In order to predict observed values $y_{i,t}$ for h future lags we provide the following procedure: $y_{i,t+h} = E\{y_{i,t+h}|y^T_{i,0}\}$, $B^T_{i,T+h} = E\{(y_{i,T+h} - y^T_{i,T+h})|y^T_{i,0}\}$. Using Kalman Filter (1)–(2) the following equations for forecasting statistics are provided:

$$x^T_{i,T+h} = Fx^T_{i,T+h-1}, \quad y^T_{i,T+h} = H_i x^T_{i,T+h}, \tag{3}$$

$$P^T_{i,T+h} = FP^T_{i,T+h-1}F^T + Q, \quad B^T_{i,T+h} = H_i P^T_{i,T+h}H_i + R. \tag{4}$$

3 Panel data in linear state space form

Classic linear mixed regression model in a compact form can be expressed in the following way: $y = X\beta + Z\gamma + \epsilon$, $E\{\gamma, \epsilon)\} = (0, 0)$, $cov(\gamma, \epsilon) = \mathrm{diag}(Q, R)$, where y is observed variable with the following expectation and covariance $E\{y\} = XB$, $cov(y, y) = ZQZ^T + R$. Matrices X and Z describe determined and stochastic effects in observations respectively. For panel data modification of a linear mixed regression model observations for i-th object $y_i = (y_{i,1}|, \cdots, y_{i,T_i})^T$, $i \in \{1, \ldots, K\}$ are aggregated for $t \in \{1, \ldots, T_i\}$ which results in the following model: $y_i = X_i\beta + Z_i\gamma_i + \epsilon_i$, $\gamma_i \sim \mathcal{N}(0, G)$, $\epsilon_i = (\epsilon_{i,1}, \epsilon_{i,T_i})^T \sim \mathcal{N}(0, \Sigma)$, which leads to $y_i \sim \mathcal{N}(X_i\beta, Z_iGZ^T_i + \Sigma_i)$.

One of the possible ways of expressing longitudinal modification of mixed regression model in state space form can be expressing observations $y_{i,t}$ as a single vector of higher dimensionality, then the state and observation equations can be formulated as following

$$y_{i,t} = x^T_{i,t}\beta_{i,t} + Z^T_{i,t}\gamma + \epsilon_{i,t}, \quad \beta_{i,t} = \beta_{i,t-1}, \tag{5}$$

where $\epsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$.

Then we apply Kalman filtering procedure (1)–(2) to the panel data model (5) and finally construct forecasting statistics (3)–(4).

4 Computational experiments

Let us consider the case described with the model (5). Let the observation vector be a constant vector with additive errors defined by $AR(1)$ process: $y_{i,t} = \beta_i + \epsilon_i$,

$\epsilon_i \sim \mathcal{N}(0, \Sigma_t)$, $\Sigma_t(i,j) = \sigma^2\phi^{|i-j|}/(1-\phi^2)$, $|\phi| < 1$. One of possible state space models for this case can be the following: $y_{i,t} = \beta_i + \epsilon_t$, $\epsilon_t = \phi\epsilon_t + \omega_t, \omega_t \sim \mathcal{N}(0, \sigma^2)$ with the initial conditions $\epsilon_t = \mathcal{N}(0, \sigma^2/(1-\phi^2))$.

The task is to estimate model parameters which can be non-trivial due to nonlinear relationships between parameters. After parameters estimates are built they can be used to construct forecasts $x_{i,t+h}$, $y_{i,t+h}$. In order to avoid this problem we construct another state space form

$$y_{i,t} = \mu + \beta_i + \epsilon_i, \quad x_{i,t} = \begin{pmatrix}\epsilon_t \\ \beta_i\end{pmatrix} = \begin{pmatrix}\phi & 1 \\ 0 & 1\end{pmatrix}\begin{pmatrix}\epsilon_{t-1} \\ \beta_i\end{pmatrix} + \begin{pmatrix}\omega_t \\ 0\end{pmatrix}, \quad y_{i,t} = (1,1)x_{i,t},$$

with $\omega_t \sim \mathcal{N}(0, \Omega)$ and the initial condition $(\epsilon_0, \beta_i)^T \sim \mathcal{N}(0, G)$, where $G = \mathrm{diag}(\sigma^2/(1-\phi^2), 0)$, $\Omega = \mathrm{diag}(\sigma^2, 0)$.

Finally this results in linear state space model and we can apply Kalman filtering procedure (1)–(4). For computational experiments we generated two-dimensional time series according to model described above. The experients were carried out with the following parameters: $\mu = 0$, $\sigma^2 = 1$, $\phi = 0.5$, $\beta_i = 1$. To construct forecasting statistics equations (3)–(4) were used. Forecasting horizon with $h = 10$ was used. We observed mean absolute percentage error below 2.1% which indicates possibility of modeling panel data using the described approach.

References

1. Ivchenko G.I. (1984). *Mathematical Statistics*. High School: Moscow (in Russian).

2. Harvey A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press: Cambridge.

3. Liptser R.S. (1974). *Stochastic Processes Statistics*. Nauka: Moscow (in Russian).

ON A SEQUENTIAL PROCEDURE FOR EARLY DETECTION OF CHANGE IN DISTRIBUTION

V.I. Lotov[1]

[1]*Sobolev Institute of Mathematics*
Novosibirsk, RUSSIA
e-mail: [1]lotov@math.nsc.ru

The paper is devoted to a sequential statistical procedure for early change detection in a probability distribution of observations.

Keywords: change point problem, probabilistic inequalities, stopping time

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables,

$$W_{n+1} = \max\{0, W_n + X_{n+1}\}, \quad W_0 = 0,$$
$$T = \inf\{n \geq 1 : W_n \geq b\}, \quad b > 0.$$

We obtain two-sided inequalities for the mean stopping time $\mathbf{E}\{T\}$ under conditions $\mathbf{E}\{X_1\} > 0$ and $\mathbf{E}\{X_1\} < 0$. These bounds are then used to characterize the quality of the sequential procedure of cumulative sums (CUSUM procedure) for the early detection of change in distribution.

ON STATISTICAL ESTIMATION OF S-DIMENSIONAL PROBABILITY DISTRIBUTION FOR BINARY RANDOM SEQUENCES

M.V. Maltsew[1]

[1]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]maltsew@bsu.by

The article is devoted to generalization of frequency estimators for the probabilities of s-tuples in binary random sequences. Such estimatoros are widely used for random numbers testing. Expectation and covariances of the generalized frequencies are calculated.

Keywords: binary sequence, frequency estimator, randomness testing

1 Introduction

Random numbers are required in many areas: statistics, physics, computer science and others. For example, in statistical physics Monte Carlo method is used for molecular modeling [1]. Random numbers are of particular importance in cryptography. They are needed to form keys, initialization vectors, initial values of variables in algorithms and for other tasks [2]. Specialized software or hardware devices (generators) are used to obtain random numbers. Typically cryptographic generators produce binary (bit) sequences. Generators used in information security systems must meet strict requirements for the quality of output sequences. Using vulnerable generators leads to key compromise and disclosure of confidential information. Therefore a thorough analysis of the reliability of the developed and used generators is necessary.

2 Frequency estimators for the probabilities of s-tuples

Binary output sequence of a secure generator must be indistinguishable from equiprobable Bernoulli sequence with a success probability of $1/2$ (null hypothesis H_0). Statistical testing of the generator's output sequences is used to verify this property. Statistical tests are combined into batteries (sets). Each test verifies a specific property of a truly random sequence. For example, the number of zeros in the sequence should not differ significantly from the number of ones, there should not be Markov dependence, and many other properties. In practice such test batteries as NIST, Diehard, TestU01 are widely used. However these and other test batteries have a number of shortcomings and limitations: they test a simple null hypothesis, the family of alternatives is not fixed, relatively simple alternatives may not be found. For example, in [3] generator built on the basis of a combination of 27-bit and 16-bit linear-feedback shift registers

successfully passed all NIST tests. In [4] it is shown that the NIST battery may fail to reject cryptographically weak binary sequences containing repeating blocks of large length. The examples given show that development of new methods and algorithms for analyzing the quality of cryptographic generators is an important task.

If the null hypothesis for binary sequence $\{x_t \in \{0,1\} : t \in \mathbb{N}\}$ is true then the s-dimensional probability distribution (for any $s \in \mathbb{N}$) will be uniform:

$$p_{J_1^s} = \mathrm{P}\{x_t = j_1, \ldots, x_{t+s-1} = j_s\} = \frac{1}{2^s},\ J_1^s = (j_1, \ldots, j_s) \in \{0,1\}^s. \tag{1}$$

Let us consider a binary sequence of length $T = ms$: $X = (x_1, \ldots, x_T) \in \{0,1\}^T$. To check the s-tuples uniformity (1) it is necessary to construct statistical estimates $\hat{p}_{J_1^s}$ for probabilities $p_{J_1^s}$. There are two approaches to calculating these frequencies. The first approach uses overlapping fragments of X, the second uses non-overlapping ones. In the first case the estimators are as follows:

$$\nu(J_1^s) = \sum_{t=1}^{T-s+1} \mathbf{1}(x_t = j_1, \ldots, x_{t+s-1} = j_s),$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Frequencies for the second case calculated using non-overlapping fragments, have the form:

$$\mu(J_1^s) = \sum_{t=1}^{m} \mathbf{1}(x_{(t-1)s+1} = j_1, \ldots, x_{ts} = j_s).$$

Under the null hypothesis frequencies $\mu(J_1^s)$ are sums of independent random variables with a uniform probability distribution on $\{0,1\}^s$, which simplifies the construction of statistical tests based on these frequencies. Using frequencies $\nu(J_1^s)$ allows one to use more information about the binary sequence X, but their use requires additional calculations.

In this paper the following generalization of frequencies $\nu(J_1^s)$ and $\mu(J_1^s)$ is proposed for statistical testing:

$$\nu_\Delta(J_1^s) = \sum_{t=1}^{T'} \mathbf{1}(x_{(t-1)\Delta+1} = j_1, \ldots, x_{(t-1)\Delta+s} = j_s),$$

where $T' = \left[\frac{T-s}{\Delta} + 1\right]$, $\Delta = 1, \ldots, s.0$ If $\Delta = 1$ then frequencies $\nu_\Delta(J_1^s)$ coincide with $\nu(J_1^s)$, if $\Delta = s$ then $\nu_\Delta(J_1^s)$ coincide $\mu(J_1^s)$.

Thus parameter Δ specifies the size of the shift between adjacent fragments when calculating frequencies. The choice of this parameter allows to maintain a compromise between the degree of fragment dependence and the accuracy of calculating the probability estimators of s-tuples.

Under the null hypothesis the expectation of $\nu_\Delta(J_1^s)$ is

$$\mathrm{E}\{\nu_\Delta(J_1^s)\} = T' p_{J_1^s},\ p_{J_1^s} = p = 2^{-s},\ \forall J_1^s \in \{0,1\}^s. \tag{2}$$

Frequency covariances are:

$$cov\{\nu_\Delta(I_1^s), \nu_\Delta(J_1^s)\} = T'p\mathbf{1}\{I_1^s = J_1^s\} + \\ +p\sum_{k=1}^{K}(T'-k)2^{-k\Delta}\left(\mathbf{1}\{I_{k\Delta+1}^s = J_1^{s-k\Delta}\} + \mathbf{1}\{J_{k\Delta+1}^s = I_1^{s-k\Delta}\}\right) - \\ -p^2\left(T'K + (T'-K)(K-1)\right), \tag{3}$$

where $T' = \left[\frac{T-s}{\Delta}\right] + 1$, $K = \left[\frac{s-1}{\Delta}\right]$.

References

1. Izaguirre J.A., Hampton S.S. (2004). Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules. *Journal of Computational Physics*. Vol. **200**, Num. **2**, pp. 581–604.

2. Schneier B. (1996). Applied Cryptography: Protocols, Algorithms, and Source Code in C (2nd ed.). John Wiley.

3. Zubkov A.M., Serov A.A. (2019). Testing the NIST Statistical Test Suite on artificial pseudorandom sequences. *Mathematical Aspects of Cryptography*. Vol. **10**, Num. **2**, pp. 89–96.

4. Zubkov A.M., Serov A.A. (2023). Experimental study of NIST Statistical Test Suite ability to detect long repetitions in binary sequences. *Mathematical Aspects of Cryptography*. Vol. **14**, Num. **2**, pp. 137–145.

MIXED-FREQUENCY DATA MODELS AND THEIR APPLICATION TO REAL-TIME ANALYSIS AND FORECASTING

V.I. Malugin[1]

[1]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]malugin@bsu.by

This paper presents monthly mixed-frequency models MIDAS, MIDAS-GETS and MF-VAR describing the dependence of the producer price index in industry of the Republic of Belarus on the Belarusian consumer price index and producer price index in industry of the Russian Federation. Daily growth rates of the Belarusian ruble against the US dollar and the Russian ruble are used as real-time data. The models demonstrate significantly higher accuracy of out-of-sample short-term forecasts compared to ARX and VARX models with the similar exogenous structure and average monthly exchange rates. The causal analysis based on the MF-VAR and VARX models allows us to conclude that the Russian index of industrial producer prices has a leading influence on the Belarusian indices.

Keywords: mixed-frequency data models; monthly industrial producer price index; monthly consumer price index; daily exchange rates; short-term forecasting; nowcasting; causal analysis.

1 Relevance of the problem

Econometric models based on mixed-frequency data have emerged and have been intensively developed in the last two decades. They provide a new methodology for joint analysis and forecasting of time series with different observation frequencies, which is relevant in the context of the growing volume and diversity of information available from various sources in real time. Thus, the use of variables with a higher frequency (e.g., week, day, hour, etc.) in models for low-frequency indicators (quarterly, monthly) made it possible to create automated platforms for monitoring macroeconomic and financial processes in real time [1]. In [2], the first such strictly substantiated platform was developed, combining models based on mixed-frequency data, including large arrays of macroeconomic, financial and news data. Machine learning and artificial intelligence methods are used in the implementation of such approach. The relevance of mixed-frequency models is related to the need for early assessment (nowcasting) of key macroeconomic and financial indicators before their official values appear with a significant delay. Compared with traditional models based on aggregated data, mixed-frequency models allow taking into account the dynamics of high-frequency variables arriving in real time within the low-frequency interval of the modeled indicator.

2 Models based on mixed frequency data

In this paper, the following models are constructed using high-frequency real-time data:

- MIDAS/PDL (*Mixed Data Sampling with Polynomial Distributed Lags*) – a model with restrictions on the lags structure determined by the Almon polynomial distributed lags [3];
- MIDAS-GETS – a model with lags structure optimization based on machine learning algorithms according to the *General-To-Specific approach* (GETS) [4];
- MF-VAR (*Mixed-Frequency Vector Autoregression*) [5] – a vector autoregression model on mixed-frequency data for joint forecasting of several low-frequency variables using high-frequency data.

The effectiveness of these models is compared with traditional univariate and vector autoregression models ARX and VARX on aggregated high-frequency exogenous variables [6].

The MIDAS modeling. Historically, the first model for mixed data is the MIDAS regression model [7]. By now, its various modifications are known, see the review [5]. The MIDAS model uses stationary representations of economic and financial time series.

Let us give an analytical description of the MIDAS/PDL model. For a discrete moment $t = 1, \ldots, T$ on a low-frequency time scale, the superscripts M and D correspond to monthly and daily time series:

- Y_t^M and Y_{t-1}^M – time series of endogenous variable and its lags;
- $Z_{k,t-1}^M$ $(k = 1, \ldots, K)$ – time series of leading monthly exogenous economic variables;
- $d_{l,t}^M$ $(l = 1, \ldots, L)$ – time series of monthly dummy variables to account for structural changes;
- η_t – independent and identically distributed random errors of observations according to the normal law.

The MIDAS model with Almon polynomial distributed lags for exogenous high-frequency variables and polynomial order p is denoted by MIDAS/PDL/p, S. The lag structure of this model for high-frequency variables is determined by a weighting function with coefficients $\theta^{(s)} = (\theta_0^{(s)}, \ldots, \theta_p^{(s)})'$:

$$w_\tau(\theta^{(s)}) = \sum_{j=0}^{p} \theta_j^{(s)} \tau^j, \ \tau = 0, 1, 2, \ldots, \ s = 1, \ldots, S. \tag{1}$$

The MIDAS PDL/p, S model in the accepted notations allows for the following representation:

$$Y_t^M = \mu + \alpha_1 Y_{t-1}^M + \sum_{k=1}^{K} \gamma_k Z_{k,t-1}^M + \sum_{l=1}^{L} \gamma_l d_{l,t}^M + \sum_{s=1}^{2} \left(\sum_{j=0}^{p} \theta_j^{(s)} \overline{X}_{s,j,t}^M \right) + \eta_t,\ t = 1, \ldots, T, \quad (2)$$

where

$$\overline{X}_{s,j,t}^M = \sum_{\tau=0}^{q_{X_s}^M - 1} \tau^j X_{s,(t-\tau)/m}^D \quad (3)$$

– projection of the values of the s-th daily variable at time t onto the monthly interval; $X_{s,(t-\tau)/m}^D$ is the value of the s-th daily variable for the time interval $(t-\tau)/m$, where hyperparameter m ($m \leq N_D$ or $m > N_D$, where N_D is the number of days in a month) is the number of lags of the daily variable in the low-frequency interval; $q_{X_s}^M$ the order of lags for the s-th aggregated high-frequency exogenous variable in the low-frequency equation.

A non-parsimonious alternative to the MIDAS/PDL model is the U-MIDAS (*Unrestricted MIDAS*) model [5], without restrictions on the parameters with the same set of low-frequency variables as model (2), which has the form:

$$Y_t^M = \mu + \alpha_1 Y_{t-1}^M + \sum_{k=1}^{K} \gamma_k Z_{k,t-1}^M + \sum_{l=1}^{L} \gamma_l d_{l,t}^M + \sum_{s=1}^{2} \left(\sum_{j=1}^{m} \theta_j^{(s)} X_{s,t,j}^D \right) + \eta_t,\ t = 1, \ldots, T. \quad (4)$$

According to (5), each of the m values of the daily variable in the U-MIDAS model corresponds to a separate low-frequency (monthly) variable. Due to this, the U-MIDAS model is linear in parameters and can be estimated using the linear least squares method. Obviously, the nonlinear MIDAS/PDL/p, S model is more economical in the number of estimated parameters than U-MIDAS. This advantage is more significant the greater the difference in the frequency of observation of low-frequency and high-frequency variables.

MIDAS-GETS model. The method of construction the MIDAS-GETS model uses an algorithm for optimizing the lags structure. In final the model includes statistically significant lags for each high-frequency variable. The MIDAS-GETS model, like the U-MIDAS, has a linear lag structure, but allows for a significant reduction in the number of parameters. Due to linearity, it is more convenient for interpretation than the MIDAS/PDL model. In addition, in the paper we use its modification, which carries out an automatic search for anomalous observations in low-frequency time series with the inclusion of corresponding dummy variables in the model.

MF-VAR model. In the model MF-VAR(p) of order p, high-frequency variables are used linearly, as in U-MIDAS [5]. This means that if the model uses m previous days to construct the current monthly forecast, then value for each day is included in the model as endogenous variable jointly with the original set of endogenous variables.

In the simplest case, for one endogenous variable and one high-frequency variable with m values $X_{t,1}^M, \ldots, X_{t,m}^M$ in the low-frequency interval the MF-VAR(p) model takes

the form:

$$
\begin{aligned}
Y_t^M &= c_t + \sum_{j=1}^{m}\sum_{l=1}^{p} \beta_l^{1,j} X_{t-l,j}^M + \sum_{l=1}^{p} \beta_l^{1,m+1} Y_{t-l}^M + \eta_{t,1},\\
X_{t,1}^M &= c_t + \sum_{j=1}^{m}\sum_{l=1}^{p} \beta_l^{2,j} X_{t-l,j}^M + \sum_{l=1}^{p} \beta_l^{m+1,2} Y_{t-l}^M + \eta_{t,2},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
X_{t,m}^M &= c_t + \sum_{j=1}^{m}\sum_{l=1}^{p} \beta_l^{m,j} X_{t-l,j}^M + \sum_{l=1}^{p} \beta_l^{m+1,m} Y_{t-l}^M + \eta_{t,m},
\end{aligned}
\tag{5}
$$

where

$$
c_t = \mu + \sum_{k=1}^{K} \gamma_k Z_{k,t-1}^M + \sum_{l=1}^{L} \gamma_l d_{l,t}^M. \tag{6}
$$

Model (5), (6) is a linear VAR model in terms of parameters and can be estimated by traditional methods for relatively small values of the triple of parameters (m, p, S). It allows the use of the *Granger causality test* [8] to analyze the nature of the causal relationship between endogenous variables, as well as conducting an *Impulse Response Analysis* [9]. To construct and select the best models, it is necessary to specify hyperparameter m, the values of m significantly affect the statistical properties, forecast accuracy and performance of the models, which requires further optimization of the models by hyperparameters.

3 The problems and data used

Problems of the study. In [10] we proposed monthly and quarterly MIDAS models for the consumer price index (CPI) in the Belarusian economy based on the effect of exchange rate pass-through to inflation. As real-time data in these models we use daily grows rates of the exchange rates of the Belarusian ruble for major currencies.

Due to the high degree of integration of the Belarusian and Russian economies [11], as well as the significant share of industrial production in the structure of GDP, the producer price index in industry is of greatest interest [12]. In this regard, the purpose of this study is to solve the following problems of analyzing and forecasting main inflation indicators in the Belarusian economy on the models with mixed and aggregated data:

1) causal analysis of the producer price index of industrial products PPI_RB and the consumer price index CPI_RB using daily growth rates of the exchange rates of the Belarusian ruble to major currencies;

2) nowcasting and short-term forecasting of the target indicator PPI_RB based on univariate models as well as jointly forecasting of all indices PPI_RB, CPI_RB, PPI_RU using multivariate models;

3) analysis of the impact of industrial producer prices in the Russian economy on the considered inflation indicators in the Belarusian economy.

Data used. To build the models we use time series of price indices provided by the National Statistical Committee of the Republic of Belarus, the Federal State Statistics Service of the Russian Federation, as well as official exchange rates of the Belarusian ruble of the National Bank of the Republic of Belarus for the observation period from June 2017 to December 2024:

- PPI_RB_t – seasonally adjusted time series of the producer price index for industrial products of the Republic of Belarus (month to month in %);
- CPI_RB_t – seasonally adjusted time series of the consumer price index of the Republic of Belarus (month to month in %), time series;
- PPI_RU_t – time series of the producer price index of industrial products of the Russian Federation (month to month in %) without seasonal adjustment;
- RUR_BYN_D_t and USD_BYN_D_t – time series of daily growth rates (day to day, in %) of the official exchange rates of the Belarusian ruble against the Russian ruble and the US dollar.

4 Forecasts accuracy evaluation and causality analysis

The forecast accuracy indicators such as root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are used for assessing retrospective and out-of-sample one-step forecasts for ARX, VARX models and nowcasts for MIDAS/PDL, MIDAS-GETS, MF-VAR. The retrospective forecasts are constructed for the entire models estimation period. The out-of-sample forecasts and nowcasts are obtained during one year by means of an expanding window algorithm that sequentially excludes the forecast month from the model's estimation period, beginning from January 2024 and up to November 2024.The results of the accuracy assessment of retrospective one-step forecasts and nowcasts for the target indicator PPI_RB on the base of univariate models are presented in Table 1, and for multivariate models – in Table 2.

Table 1: Forecasts accuracy evaluation for target indicator PPI_RB

Retrospective forecasts			
Indicators	MIDAS	MIDAS-GETS	ARX
RMSE	0,2973	0,1753	0,2883
MAE	0,2411	0,1284	0,2352

MAPE	0,2396	0,1276	0,2338
Out-of-sample forecasts			
Indicators	MIDAS	MIDAS-GETS	ARX
RMSE	0,2438	0,2382	0,2812
MAE	0,2166	0,2031	0,2350
MAPE	0,2156	0,2020	0,2336

Table 2: Out-of-sample forecast accuracy for three indices

MF-VAR			
Indicators	CPI RB	PPI RB	PPI RU
RMSE	0.156525	0.232963	1.342147
MAE	0.121449	0.198602	0.973767
MAPE	0.120920	0.197601	0.976148
VARX			
Indicators	CPI RB	PPI RB	PPI RU
RMSE	0.131484	0.284091	1.146631
MAE	0.107709	0.236294	0.932707
MAPE	0.107213	0.234952	0.929154

5 Causality analysis based on multivariate models

Using the Granger causality test [7] and Impulse Response Analysis [8], the leading nature of both CPI_RB and PPI_RU indices was established for the PPI_RB index: the CPI_RB index has a lead of 1 lag, while the PPI_RU index has a lead of 2 lags. At the same time, the effects of a significant increase in industrial producer prices in Russia have a significantly longer attenuation period than the effects of consumer price growth in the Belarusian economy.

Figure 1 illustrates the responses of the industrial producer price index PPI_RB to impulse shocks from consumer prices in Belarus and industrial producer prices in Russia. An impulse shock in the form of a one-time increase in the CPI_RB consumer price index causes an increase in the PPI_RB index, which reaches its maximum value within one quarter and then fades away rather quickly. While similar shocks from producer prices in the Russian industry PPI_RU are maximally manifested in the PPI_RB index over two quarters, and their extinction occurs over 2-3 quarters.

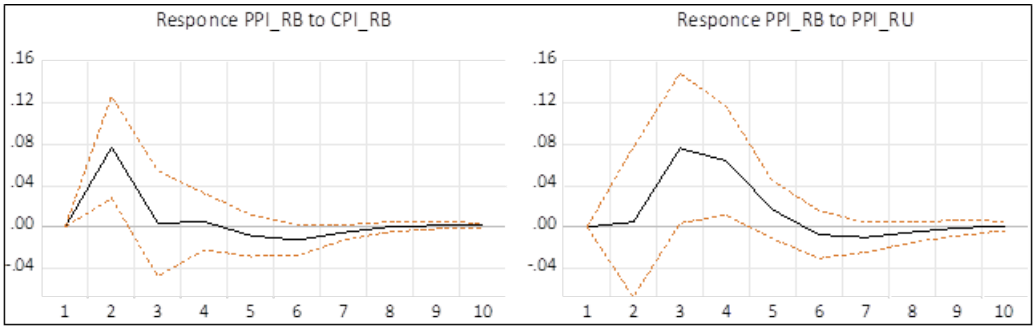


Figure 1: Responses of PPI_RB to impulse shock impacts from CPI_RB and PPI_RU

References

1. Giannone D., Reichlin L., Small D. (2008) Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*. 55(4). pp. 665–676.

2. Macroeconomic Nowcasting and Forecasting with Big Data (2017). Brandyn Bok, Daniele Caratelli, Domenico Giannone, Argia Sbordone, Andrea Tambalotti. *Federal Reserve Bank of New York*. Staff Reports. 830. 38 p.

3. Andreou E. Ghysels E., Kourtellos A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*. 58. pp. 246–261.

4. Campos J., Ericsson N., Hendry D. (2005). General-to-specific Modeling: An Overview and Selected Bibliography. *International Finance Discussion Paper*. 838. pp. 1–94.

5. Foroni, C., Marcellino M. (2013). A survey of econometric methods for mixed frequency data. *Working Paper. Norges Bank*. 6, 45 p.

6. Kharin Yu.S., Malugin V.I., Kharin A.Yu. (2003). Econometric modeling. *Minsk: BSU*, 318 p.

7. Ghysels E. Santa-Clara P., Valkanov R. (2002). The MIDAS touch: Mixed data sampling regression models / E. Ghysels, // *Working paper, UNC and UCLA*, 33 p.

8. Granger C.W.J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*. 37 (3). pp. 424–438.

9. Simes C.A., Goldfeld S.M., Sachs J.D. (1982). Policy Analysis with Econometric Models. *Brookings Papers on Economic Activity*. 1. pp. 107–164.

10. Malugin V.I. (2024). Short-term forecasting and nowcasting of inflation growth rates based on mixed data models. *Banking Bulletin*. 1/726. pp. 23–36.

11. Malugin V.I., Novopoltsev A.Yu. (2022). The relationship between the growth rates of the economies of Belarus and Russia under shocks: econometric analysis and forecasting. *Economy. Modeling. Forecasting.* 16. pp. 236–250.

12. Kravtsov M.K., Kartun A. (2010). Econometric modeling and forecasting of the main price indices in Belarus. *Banking Bulletin.* 22/495. pp. 25–33.

STATISTICAL CLASSIFICATION OF STATIONARY TIME SERIES BY AUTOREGRESSIVE MODEL PARAMETERS AND ITS EFFICIENCY

G. Mikulich[1], E. Zhuk[2]
[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]aragornguga@gmail.com, [2]zhukee@mail.ru

The problem of statistical classification of stationary time series in the AR-model parameters space is considered. The decision rule based on the least squares method is proposed and its efficiency is analytically investigated.

Keywords: statistical classification, stationary time series, AR-model

1 Introduction

Let $X^n = (X_1, \ldots, X_n)$ be an observed random sample of n i.i.d. random vectors from $L \geq 2$ classes $\Omega_1, \ldots, \Omega_L$. Observation X_t is a member of the class with random, unknown number $d_t^0 \in S$, $S = \{1, \ldots, L\}$, $t \in \{1, \ldots, n\}$. If the class number is fixed: $d_t^0 = i$, $i \in S$, than X_t is a realization of length T_t ($X_t = (x_{t1} \ldots, x_{tT_t})' \in \mathbb{R}^{T_t}$, $'$ means transposing) of the time series which may be represented [1, 3] using autoregressive model $x^i = \{x_l^i\}_{l=-\infty}^{+\infty}$ of order p $p \geq 1$ (AR(p) for short)

$$x_l^i + \theta_{i1}^0 x_{l-1}^i + \ldots + \theta_{ip}^0 x_{l-p}^i = u_l^i, \quad l \in \mathbb{Z}, \tag{1}$$

where $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$, $\theta_i^0 \in \mathbb{R}^p$ are autoregressive parameter for class with number i, and $\{u_l^i\}_{l=-\infty}^{+\infty}$ are independent in total, identically distributed random values with a mathematical expectation of 0 and the same variation σ^2 for all classes Ω_i:

$$E\left\{u_l^i\right\} = 0, \; D\left\{u_l^i\right\} = \sigma^2, \; l \in \mathbb{Z}, \quad i \in S. \tag{2}$$

Classes Ω_i differ from each other by autoregressive coefficients θ_i^0, and by prior probability

$$P\left\{d_t^0 = i\right\} = \pi_i^0 > 0, \quad i \in S, \quad \sum_{i=1}^{L} \pi_i^0 = 1. \tag{3}$$

The problem is to classify the observations $X^n = (X_1, \ldots, X_n)$ between the classes Ω_i, so, to construct the decision rule (DR):

$$d = d(X_1, \ldots, X_n) \in S, \quad S = \{1, \ldots, L\}, \tag{4}$$

and then, to estimate efficiency of this rule using generalized risk.

2 Decision rule in the AR coefficients space

The model (1) can also be written as Wold decomposition [1, 2]:

$$x_l + \sum_{j=1}^{\infty} \theta_j x_{l-j} = u_l, \quad l \in \mathbb{Z} \tag{5}$$

Let us calculate the estimate for the first p autoregressive coefficients $\theta_{(p)} = (\hat{\theta_1}, \ldots, \hat{\theta_p})'$ using realization $X_T = \{x_t; t = 1, \ldots, T\}$:

$$\hat{\theta}_{(p)} = -\left(\sum_{t=p+1}^{T} X_t X_t' \right)^{-1} \sum_{t=p+1}^{T} x_t X_t, \tag{6}$$

where $X_t = (x_{t-1}, \ldots, x_{t-p})$, $t \in \{p+1, \ldots, T\}$.

The decision then is to take such a class that the distance between estimates and true autoregressive coefficients for this class is minimal across all the classes. So,

$$d = d(X) = \arg\min_{i \in S} |\hat{\theta} - \theta^{(i)}| = \arg\min_{i \in S} |\hat{\theta} - \theta^{(i)}|^2 =$$

$$= \arg\min_{i \in S} \left\{ \left| \hat{\theta}_{(p)} - \theta^{(i)}_{(p)} \right|^2 + \sum_{j=p+1}^{+\infty} \left(\theta_j^{(i)} \right)^2 \right\}, \tag{7}$$

where $\hat{\theta} = (\hat{\theta'}_{(p)}, 0, \ldots, 0, \ldots)'$, and $\theta^{(i)}_{(p)} = (\theta_1^{(i)}, \ldots \theta_p^{(i)})', i \in S$

If the $AR(p)$ model is such that $\theta^{(i)} = ((\theta^{(i)}_{(p)})', 0, \ldots, 0, \ldots)'$, then the decision rule from (7) can be simplified:

$$d = d(X) = \arg\min_{i \in S} \left| \hat{\theta}_{(p)} - \theta^{(i)}_{(p)} \right|. \tag{8}$$

3 Generalized risk for the decision rule

To measure the efficiency of the proposed decision rule, we will use generalized risk [4]:

$$r_T = P\{d(X) \notin D^0\}, \quad D^0 = \left\{ k : \left| \theta - \theta^{(k)} \right| = \min_{i \in S} |\theta - \theta^{(i)}| \right\} \tag{9}$$

where $D^0 \subseteq S$ is a set of the numbers of such classes, for which the time series from (5) is closer in the term of Euclidean distance. The generalization of risk allows us to handle the cases where there are multiple classes that are closest, when the function of decision rule returns a set instead of a number. The risk itself r_T is a probability ($0 \leq r_T \leq 1$) not to assign the time series by its realization $X = \{x_t\}_{t=1}^T$ to one of the closest classes. So, risk is used to measure the efficiency of the decision rules in the following way: the lower the risk, the more efficient is the decision rule.

To cover border cases, if $D^0 = S$, then $r_T = 0$, so the decision rule is not relevant. If $|D^0| = 1$, so the set is of one element, then we will get the classic risk and (9) can be simplified to

$$r_T = P\{d(X) \neq d^0\}, \quad d^0 = \arg\min_{i \in S} |\theta - \theta^{(i)}| \tag{10}$$

References

1. Anderson T. (1976). *The Statistical Analysis of Time Series*.

2. Kharin Yu.S., Zuev N.M., Zhuk E.E. (2011). *Probability Theory, Mathematical and Applied Statistics*. BSU, Minsk.

3. Borovkov A.A. (1984). *Mathematical Statistics*. Nauka, Moskow.

4. Zhuk E.E. (2013). Assignment of multivariate samples to the fixed classes by the maximum likelihood method and its risk. *Computer Data Analysis and Modeling: Theoretical and Applied Stochastics : Proc. of the Tenth Intern. Conf.* Vol. **1**, pp. 185-188.

ON THE GENERALIZED INVERSE MOORE-PENROSE MATRIX FOR MULTIDIMENSIONAL MATRICES

V.S. Mukha[1]

[1]*Belarusian State University of Informatics and Radioelectronics*
Minsk, BELARUS
e-mail: [1]mukha@bsuir.by

The article is devoted to dissemination of the known for the usual matrices generalized inverse Moore-Penrose matrix to the multidimensional matrices. At first, the generalized inverse Moore-Penrose matrix for the usual matrices is considered. Then the definition of the generalized Moore-Penrose inverse matrix for the multidimensional matrix is introduced and two theorems about the properties of this matrix are presented. The example on use the multidimensional generalized inverse Moore-Penrose matrix in the problem of the multidimensional-matrix polynomial regression is given.

Keywords: inverse Moore-Penrose matrix, multidimensional matrix, polynomial multidimensional-matrix regression

1 Introduction

The generalized inverse Moore-Penrose matrix is one of the amazing scientific achievements of the 20th century. It was proposed by E. H. Moore in 1920 [1], and more detail considered in 1955 by R. Penrose [2] without cite of the Moore's work. The mechanism of its properties has not been fully studied. The epithet pseudoinverse assigned to it restrained and restraints its widespread use. In this article, the sphere of application of the generalized inverse Moore-Penrose matrix is expanding to the multidimensional matrices.

2 The generalized inverse Moore-Penrose matrix for twodimensional matrices

Let us denote $M(R, m, n)$ the set of all $m \times n$ real matrices, and $A \in M(R, m, n)$ is the $m \times n$ matrix with real elements.

Let $A \in M(R, m, n)$. The matrix $A^+ \in M(R, n, m)$ satisfying the equalities

$$AA^+A = A, \tag{1}$$

$$A^+AA^+ = A^+, \tag{2}$$

$$(AA^+)^T = (AA^+) \quad \text{(the matrix AA}^+\text{ is symmetric)}, \tag{3}$$

$$(A^+A)^T = (A^+A) \quad \text{(the matrix A}^+\text{A is symmetric)}, \tag{4}$$

is called the generalized inverse Moore-Penrose matrix (MP-inverse matrix) to the matrix $A \in M(R, m, n)$ [2].

Theorem 1. Let the linear vector-matrix equation $AX = B$ be solved. The vector $X = A^{+}B$ provides the minimal value of the Euclidean norm $||AX - B||$ and has minimal value of the Euclidean norm $||X||$ [3, 4].

3 The generalized inverse Moore-Penrose matrix for multidimensional matrices

The question of the inversion of the multidimensional matrices is the key one in solution of many multidimensional-matrix problems. It is of interest the algorithms and programs for practical use, i.e. numerical algorithms for the matrices of the big sizes. Such algorithms and programs are developed for the twodimensional matrices, so it is advisably to consider the inverse of the multidimensional matrices as the inverse of their associated twodimensional matrices.

The matrix A is called k-dimensional, if its elements contain k indices $i_1, \ldots, i_n$ [5, 6]:

$$A = (a_{i_1,\ldots,i_n}), \;\; i_\alpha = 1, \ldots, n_\alpha, \;\; \alpha = 1, \ldots, k.$$

If $n_1 = n_2 = \cdots = n_k = n$, then the matrix A is called k-dimensional matrix of the order n.

The matrix denoted by $E(\lambda, \mu)$ is called (λ, μ)-unit matrix, if it satisfies the equalities

$$^{\lambda,\mu}(AE(\lambda, \mu)) = ^{\lambda,\mu} (E(\lambda, \mu)A) = A$$

for any $(\lambda + 2\mu)$-dimensional matrix $A = (a_{\overline{l}_{(\mu)},\overline{s}_{(\lambda)},\overline{c}_{(\mu)}})$ of the order n , where $\overline{l}_{(\mu)}$, $\overline{s}_{(\lambda)}$, $\overline{c}_{(\mu)}$ are $\mu-, \lambda-, \mu-$multiindices respectively [5, 6].

The matrix $^{\lambda,\mu}A^{-1}$ is called (λ, μ)-inverse to the $(\lambda + 2\mu)$-dimensional matrix $A = (a_{\overline{l}_{(\mu)},\overline{s}_{(\lambda)},\overline{c}_{(\mu)}})$ of the order n , if it satisfies the equalities:

$$^{\lambda,\mu}(A^{\lambda,\mu}A^{-1}) = ^{\lambda,\mu} (^{\lambda,\mu}A^{-1}A) = E(\lambda, \mu).$$

The matrix denoted by $E^s(\lambda, \mu)$ is called symmetrical (λ, μ)-unit matrix, if it satisfies the equalities

$$^{\lambda,\mu}(A_s E^s(\lambda, \mu)) = ^{\lambda,\mu} (E^s(\lambda, \mu)A_s) = A_s$$

for any symmetrical $(\lambda + 2\mu)$-dimensional matrix $A_s = (a_{\overline{l}_{(\mu)},\overline{s}_{(\lambda)},\overline{c}_{(\mu)}})$ of the order n [6].

The matrix $^{\lambda,\mu}A_s^{-1}$ is called (λ, μ)-inverse to the symmetrical $(\lambda + 2\mu)$-dimensional matrix $A_s = (a_{\overline{l}_{(\mu)},\overline{s}_{(\lambda)},\overline{c}_{(\mu)}})$ of the order n , if it satisfies the equalities:

$$^{\lambda,\mu}(A_s \;\; ^{\lambda,\mu}A_s^{-1}) = ^{\lambda,\mu} (^{\lambda,\mu}A_s^{-1}A_s) = E^s(\lambda, \mu). \tag{5}$$

The problem of inversion of the symmetrical multidimensional matrices is especially acute since the matrices associated with the symmetrical multidimensional matrices are singular. This means that the (usual) inverse matrix $^{\lambda,\mu}A_s^{-1}$ satisfying the equalities (5),

does not exist. The way out in this situation lies in the area of so-called pseudoinversion of matrices.

Definition. The generalized (λ,μ)-inverse Moore-Penrose matrix (MP (λ,μ)-inverse matrix) to the $(\lambda{+}2\mu)$-dimensional matrix $A=(a_{\bar{l}_{(\mu)},\bar{s}_{(\lambda)},\bar{c}_{(\mu)}})$ of the order n is the matrix ${}^{\lambda,\mu}A^{+}$ satisfying to the following equalities:

$$ {}^{\lambda,\mu}(A\ {}^{\lambda,\mu}({}^{\lambda,\mu}A^{+}\ A))=A, \tag{6}$$

$$ {}^{\lambda,\mu}({}^{\lambda,\mu}A^{+}\ {}^{\lambda,\mu}(A\ {}^{\lambda,\mu}A^{+}))={}^{\lambda,\mu}\ A^{+}, \tag{7}$$

$$ {}^{\lambda,\mu}(A\ {}^{\lambda,\mu}A^{+})\ \textit{and}\ {}^{\lambda,\mu}({}^{\lambda,\mu}A^{+}\ A)\ \textit{are symmetrical matrices.} \tag{8}$$

The following theorem shows that MP $(0,\mu)$-inverse matrix ${}^{0,\mu}A^{+}$ gives the solution to the least squares problem for the following multidimensional-matrix equation:

$$ {}^{0,\mu}(A_{(\mu,0,\mu)}X_{(\mu,0,0)})=B_{(\mu,0,0)}. \tag{9}$$

Theorem 2. Let the equation (9) with (2μ)-dimensional matrix $A=(a_{\bar{l}_{\mu},\bar{c}_{\mu}})=A_{(\mu,o,\mu)}$ is solved and ${}^{0,\mu}A^{+}={}^{0,\mu}\ A^{+}_{(\mu,o,\mu)}$ is the MP $(0,\mu)$-inverse matrix to the matrix $A_{(\mu,0,\mu)}$. The matrix $X={}^{0,\mu}\ ({}^{0,\mu}A^{+}B)$ provides the minimal value of the Euclidean norm $||{}^{0,\mu}(AX)-B||$ and has the minimal Euclidean norm $||X||$.

Theorem 3. The matrix ${}^{\lambda,\mu}A^{+}$ (6), (7), (8), MP (λ,μ)-inverse to the symmetrical $(\lambda{+}2\mu)$-dimensional matrix $A=(a_{\bar{l}_{(\mu)},\bar{s}_{(\lambda)},\bar{c}_{(\mu)}})$ of the order n is the (λ,μ)-inverse matrix ${}^{\lambda,\mu}A_{s}^{-1}$ (5).

The converse to theorem 3 statement is also true: the symmetrical (λ,μ)-inverse matrix ${}^{\lambda,\mu}A_{s}^{-1}$ (5) satisfies the conditions (6), (7), (8), i.e. it is the MP (λ,μ)-inverse matrix ${}^{\lambda,\mu}A^{+}$.

Thus, the symmetry of the matrix A is the necessary and sufficient condition for the equality ${}^{\lambda,\mu}A^{+}={}^{\lambda,\mu}\ A_{s}^{-1}$.

The algorithm for finding the MP (λ,μ)-inverse matrix is as follows: 1) the twodimensional matrix (λ,μ)-associated with the inverting $(\lambda+2\mu)$-dimensional matrix A is formed; 2) the MP inverse matrix to the associated matrix is found; 3) the inverse transformation is performed from the twodimensional MP inverse matrix to the MP (λ,μ)-inverse matrix.

Many programming systems have the programs for finding the MP inverse matrix. This is the function pinv in the Matlab programming system.

4 Example

The need to use the MP (λ,μ)-inverse matrix arises in the multidimensional-matrix polynomial regression analysis to estimate the coefficients of the multidimensional-matrix polynomials regression with the following mathematical model of the measurements [7]:

$$ y_{o,i}=\sum_{k=0}^{m}{}^{0,kq}(x_{i}^{k}C_{(kq,p)})+\epsilon_{i},\ \ m=0,1,\ldots,\ \ i=1,\ldots,n,$$

by the measurements $(x_i, y_{o,i})$, where the input variable x is the q-dimensional matrix, the output variable (response) y is the p-dimensional matrix, the coefficient $C_{(kq,p)}$ is the $(kq+p)$-dimensional matrix, x_i are the measurements of the input variable x , $y_{o,i}$ are the measurements of the response y_i with errors.

The estimations $\hat{C}_{(kq,p)}$ of the coefficients $C_{(kq,p)}$ are defined as the solution to the following system of the multidimensional-matrix linear equations relative the $C_{(kq,p)}$ [7]:

$$\sum_{k=0}^{m} {}^{0,kq}(s_{x^{\lambda+k}} C_{(kq,p)}) = s_{x^{\lambda} y}, \;\; \lambda = 0, 1, \ldots, m, \tag{10}$$

where

$$s_{x^{\lambda+k}} = \frac{1}{n} \sum_{i=1}^{n} x_i^{\lambda+k}, \;\; s_{x^{\lambda} y} = \frac{1}{n} \sum_{i=1}^{n} x_i^{\lambda} y_{o,i}.$$

The diagonal elements $s_{x^2}, s_{x^4}, s_{x^6}, \ldots$ of the cell of the system (10) are the symmetrical $(0 + 2kq)$-dimensional matrices, and associated matrices to the $s_{x^4}, s_{x^6}, \ldots$ are singular one. The need to inverse these matrices arises when solving the system by the elimination Gauss method.

References

1. E.H. Moore. On the reciprocal of the general algebraic matrix. Bulletin of the American Mathematical Society 26, 394–395 (1920). http://www.ams.org/bull/1920–26–09/S0002–9904–1920–03322–7/S0002–9904–1920–03322–7.pdf

2. R. Penrose. A generalized inverse for matrices. Proceedings of the Cambridge Philosophical Society 51, 406–413 (1955). DOI: https://doi.org/10.1017/S0305004100030401

3. Arthur Albert. Regression and the Moore–Penrose Pseudoinverse. Academic press. New-York and London, 1972, 180 p.

4. J.C.A. Barata, M. S. Hussein. The Moore–Penrose Pseudoinverse: A Tutorial Review of the Theory. Brazilian Journal of Physics. 31 October 2011. Pp. 1–23. https://www.semanticscholar.org/reader/154168e98cff79dc81c5de95f6b1f095fbff8b59

5. Sokolov N.P. Introduction to the theory of multidimensional matrices. Kiev, Naukova dumka, 1972. 176 p. (In Russian).

6. Mukha V.S. Analysis of the multidimensional data. Minsk, Technoprint, 2004, 368 p. In Russian.

7. Mukha V.S. Multidimensional–matrix polynomial regression analysis. Estimations of the parameters // Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series. No 1, 2007. Pp. 45–51. (In Russian).

ESTIMATION OF PARAMETERS OF NLFSR USING MARKOV CHAINS WITH PARTIAL CONNECTIONS

U.Yu. Palukha[1], A.A. Pardaev

[1]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]palukha@bsu.by

This paper investigates the estimation of parameters of nonlinear feedback shift registers (NLFSRs) using Markov chains with partial connections. A computational experiment demonstrates the effectiveness of the algorithm in recovering feedback functions.

Keywords: cryptographic generators, Markov chains, nonlinear feedback shift registers, statistical analysis, pseudorandom sequences

1 The Method of Estimation of Parameters of Nonlinear Feedback Shift Registers

1.1 Feedback Shift Registers

Feedback shift registers (FSRs) are effective tools for generating pseudorandom sequences, widely applied in cryptography, hardware testing, data compression, and other fields. FSRs are categorized into two main types. Linear feedback shift registers (LFSRs) determine their state through a linear function of previous states. They are simple to implement, but lack sufficient cryptographic strength. LFSRs are well studied with established theoretical foundations. Consequently, nonlinear feedback shift registers (NLFSRs), which generalize LFSRs by using nonlinear transformations of previous states, are of greater practical interest due to their enhanced cryptographic security.

An FSR consists of n binary storage cells, each holding one bit. Each cell i is associated with a state variable x_i, representing its current value, and a feedback function f, which determines the updated value of bit i. The state of an FSR can be represented as a vector of its state variables $(x_0, x_1, \ldots, x_{n-1})$. The period of an FSR is defined as the length of the longest repeating output sequence it can produce. The value of the cell number 0 determines the output of the FSR, while the input is defined by the value of the cell number $n-1$.

An FSR is classified as an LFSR if it employs only linear feedback functions (i.e., $f(x_0, x_1, \ldots, x_{n-1}) = c_0x_0 \oplus c_1x_1 \oplus \cdots \oplus c_{n-1}x_{n-1}$). Otherwise, it is an NLFSR. Nonlinear feedback functions have been extensively studied. For example, in [1] a list of n-bit NLFSRs with optimal periods of $2^n - 1$ is provided. The study considered three types of feedback functions with algebraic degree two:

- $f(x_0, x_1, ..., x_{n-1})) = x_0 \oplus x_a \oplus x_b \oplus (x_c \cdot x_d)$,

- $f(x_0, x_1, ..., x_{n-1}) = x_0 \oplus x_a \oplus (x_b \cdot x_c) \oplus (x_d \cdot x_e)$,
- $f(x_0, x_1, ..., x_{n-1}) = x_0 \oplus x_a \oplus x_b \oplus (x_c \cdot x_d) \oplus (x_e \cdot x_h)$,

where $0 \leq a, b, c, d, e, h < n$.

These functions were analyzed earlier in [2], although not all optimal-period feedback functions were listed. A new method to construct such functions is proposed in [3], completing the lists for the above types. Furthermore, in [4] a new type of NLFSR feedback function with optimal periods for $8 \leq n \leq 23$ is introduced:

$$f(x_0, x_1, ..., x_{n-1}) = x_0 \oplus x_a \oplus x_b \oplus x_c \oplus \oplus x_d \oplus x_e \oplus x_h \oplus (x_w \cdot x_z),$$

where $0 \leq a, b, c, d, e, h, w, z < n$. This function has an algebraic degree of two, as the largest product term involves two variables. The authors of [4] identified 639 new NLFSR feedback functions with optimal periods, summarized in Table 1.

Table 1: Number of new NLFSR feedback functions with optimal periods

n	Number of Functions	n	Number of Functions
8	12	16	84
9	12	17	70
10	34	18	44
11	26	19	23
12	64	20	24
13	64	21	17
14	76	22	12
15	70	23	7
Total		639	

This study focuses on performing methods from [2] to these new functions.

1.2 Algorithm of Estimation of parameters of Shift Registers Using Markov Chains

Analyzing the behavior of new type NLFSRs, characterized by complex feedback functions, requires methods that account for the generators determinism and hidden dependencies between state bits. A Markov chain model of order s with r partial connections (MC(s, r)) approximates the generators dynamics while minimizing the number of model parameters. The key assumption is that the probability of the current state depends on a subset of r significant states from a history of length s.

In [5] an algorithm for estimating MC(s, r) parameters is proposed. This algorithm constructs an estimate of the one-step transition probability matrix Q, which can be used to build the truth table of a Boolean function. The matrix Q has dimensions $2^r \times 2$, with rows containing probabilities of generating 0 or 1 based on history $x_{t-1}, \ldots, x_{t-s}$. The Boolean function is derived from the second column of Q, where, for an undistorted generator, the rows are $(0, 1)$ or $(1, 0)$.

To develop an algorithm for estimating NLFSRs based on the MC(s, r) model, the following parameters must be determined:

- the optimal Markov chain order s, representing the history length needed to describe the registers dynamics;
- the number of significant connections r, minimizing the number of bits required to reconstruct the registers behavior;
- the template M of significant bits, indicating which history bits have the most influence on transitions.

The algorithm proceeds as follows:

1. Data preparation. Generate sequences of length T using the feedback function.
2. Parameter Estimation. Apply the MC(s, r) parameter estimation algorithm to estimate the order s, the number of significant connections r, the template M and the transition matrix Q.
3. Truth Table Construction. List all possible 2^r states of the significant bits in the history based on M. Compute the output for each history to form the truth table.
4. Algebraic Normal Form (ANF) Recovery. Express the Boolean function f as $f(x_0, x_1, ..., x_r) = c_0 \oplus c_1 x_1 \oplus \cdots \oplus c_{12} x_1 x_2 \oplus \cdots$. Using the truth table, derive the ANF via the method of undetermined coefficients, which should match the original feedback function.

2 Computational Experiment

The computational experiment aims to verify the effectiveness of the proposed algorithm for recovering NLFSR feedback functions by analyzing generated sequences. The main tasks include:

- generating sequences of specified lengths for various NLFSRs;
- estimating the order s, the number of significant connections r, the template M and the transition matrix Q using the MC(s, r) parameter estimation algorithm;
- recovering the ANF of feedback functions;
- analyzing the results.

The experiment considered five NLFSRs, each defined by its feedback function f_i, register size s, number of connections r, and template M. The parameters are listed in Table 2. For each NLFSR, three sequences of length $T = 4000$ bits were generated. Initial register states were chosen randomly to ensure input diversity. Sequences were

Table 2: Parameters of shift registers and estimation algorithm

Register	s	r	$r_{\min}$	$r_{\max}$
f_1	17	6	4	7
f_2	24	7	5	8
f_3	17	9	7	10
f_4	8	7	5	8
f_5	23	8	6	9

generated using the feedback function, computing the next bit and shifting the register at each step.

The implemented algorithm for recovering NLFSR parameters consists of the following steps:

1. Generate a bit sequence of length T using the feedback function and initial state.
2. Extract $(s+1)$-grams representing state transitions.
3. Count the frequency of each $(s+1)$-gram to estimate transition probabilities.
4. Apply the A3 template reduction algorithm [5] to evaluate templates M for orders from $r_{\min}$ to $r_{\max}$, maximizing mutual information I_{r+1}. Start with the maximum order and iteratively reduce it to select the optimal template.
5. Compute the transition probability matrix Q based on the selected template.
6. Form the truth vector from Q and compute the ANF of the feedback function.

The experiment was conducted for all five NLFSRs. For each register and its three sequences, the A3 algorithm [5] correctly identified the order r, template M, and accurately recovered the ANF of the feedback function. Figure 1 illustrates the program output during the recovery of function f_1.

```
Register f3 (s=17, true r=9, true M=[0, 2, 3, 6, 9, 10, 12, 13, 14], r_minus=7, r_plus=10):
2025-05-28 13:59:47,552 - INFO - ▶ Start of analysis: register f2, sequence 1
2025-05-28 13:59:47,552 - INFO - ▶▶ A3: register f2, sequence 1
2025-05-28 13:59:47,555 - INFO - Evaluation of 11440 templates for r=10
2025-05-28 14:01:56,339 - INFO - A1 completed: r=10, I=0.6924, template: (0, 2, 3, 6, 9, 10, 12, 13, 14, 15), time: 128.78s
2025-05-28 14:01:56,348 - INFO - Estimating 9 reduced templates for r=9
2025-05-28 14:01:56,441 - INFO - Estimating 8 reduced templates for r=8
2025-05-28 14:01:56,505 - INFO - Estimating 7 reduced templates for r=7
2025-05-28 14:01:56,555 - INFO - A3 completed: best r=9, template: (0, 2, 3, 6, 9, 10, 12, 13, 14), time: 129.00s
 Sequence 1: ✅ Success
 Estimated r: 9, Estimated M: (0, 2, 3, 6, 9, 10, 12, 13, 14)
 Recovered ANF: x0 ^ x10 ^ x12 ^ x13 ^ x14 ^ x2 ^ x9 ^ (x3 & x6)
```

Figure 1: Recovery of the feedback function by the implemented program

Previous study [2] found that a sequence length of $T = 1000$ bits was sufficient for recovering simpler feedback functions described in [1]. However, for the new-type NLFSRs, a longer sequence of $T = 4000$ bits was necessary due to the exponential growth in the number of contexts 2^r. The number of observations per context is approximately $T/2^r$. For accurate probability estimation, this number must be sufficiently large. The

standard error for a binomial probability p is $\sqrt{p(1-p)/N}$, where N is the number of observations. For $p = 0.5$, the error is $0.5/\sqrt{N}$, requiring $N \geq 25$ for an error less than 0.1. Consider examples for two NLFSRs:

- For f_1 with $s = 17$, $r = 6$, at $T = 1000$, the number of $(s+1)$-grams is $T-s = 983$, with $2^r = 64$ contexts, yielding $983/64 \approx 15.36$ observations per context. This is insufficient for reliable estimation. At $T = 4000$, the number of $(s + 1)$-grams is 3983, yielding $3983/64 \approx 62.23$ observations, which, though below 100, proved sufficient for recovery.

- For f_5 with $s = 23$, $r = 8$, at $T = 4000$, the number of $(s+1)$-grams is 3977, with $2^r = 256$ contexts, yielding $3977/256 \approx 15.54$ observations, which is acceptable. Increasing T further improves accuracy.

For NLFSRs with large s and r, such as f_2, the number of possible templates grows significantly. For example, with $s = 24$, $r = 8$, the algorithm evaluates $\binom{23}{7} = 245157$ templates, requiring substantial data for accurate template selection based on mutual information.

References

1. Dubrova E. (2012). A List of Maximum Period NLFSRs. *Cryptology ePrint Archive, Report 2012/166.* `http://eprint.iacr.org/2012/166`.

2. Palukha V.Yu., Kharin Yu.S. (2015). Evaluation of Cryptographic Generators Based on High-Order Markov Chains. *Information Security Issues*. No. **1**, p. 12–14.

3. Almuhammadi S., Al-Hejri I., Bin Talib G., Gaamel A. (2018). NLFSR Functions with Optimal Periods. *Proceedings of the 18th International Conference, Melbourne, VIC, Australia, July 2–5, 2018*, p. 67–79.

4. Al-Hejri I., Al-Kharobi T. (2019). A New Type of NLFSR Functions with Maximum Periods. *Transactions on Networks and Communications*. Vol. **7**, No. **2**, p. 20–30.

5. Kharin Yu.S., Petlitskiy A.I. (2007). Markov Chain of Order s with r Partial Connections and Statistical Inferences about Its Parameters. *Discrete Mathematics*. Vol. **19**, No. **2**, p. 109–130.

IMPORT INTENSITY DYNAMICS IN CHINESE ECONOMIC SECTORS: TIME SERIES CLUSTERING OF INPUT-OUTPUT DATA (1981–2018)

U. Parkhimenka[1], A. Bykau[2]

[1]*Belarusian State University of Informatics and Radioelectronics*
[2]*Belarusian State Economic University*
Minsk, BELARUS
e-mail: [1]parkhimenko@yandex.ru, [2]aliaksei.bykau@yandex.ru

This paper analyzes the evolution of import intensity in Chinese domestic production across 18 economic sectors (1981–2018). Using Dynamic Time Warping (DTW) time-series clustering on annual input-output tables via the `dtwclust` R package, we identify four distinct temporal patterns of import dependence. As a pilot study using estimated import matrices derived via the proportionality assumption, these findings provide preliminary insights into China's sectoral import intensity dynamics.

Keywords: China's economic dynamics, import intensity, input-output model, structural change, time series clustering.

1 Introduction

China's extraordinary economic expansion over recent decades remains a major subject of economic inquiry (e.g., [1]). Among the analytical frameworks employed to examine this development, input-output analysis has been utilized by researchers (e.g., [2, 3]).

This paper examines the evolution of import dependency across Chinese economic sectors during this period, as revealed by input-output tables. Quantifying sectoral reliance on foreign inputs is crucial not only for understanding growth patterns but also for modeling the impact of global value chain disruptions (e.g., from natural disasters or geopolitical events) and assessing national exposure to foreign supply shocks.

2 Data

This paper utilizes the China Time-Series Input-Output Tables (1981–2018) [4], developed by researchers at Renmin University of China's School of Applied Economics. These annually compiled tables, valued at current producer prices, cover 18 sectors under a consistent classification: one agricultural, ten industrial, one construction, and six service sectors. The methodology aligns with China's National Bureau of Statistics (NBS) practices and adheres to the competitive import assumption.

A key limitation is the aggregation of interindustry flows without distinguishing domestic and imported components. While methodologically sound within the classical Leontief framework, this contrasts with practices in many other countries. Disaggregating imports is essential for advanced analyses, such as tracing imported goods within

the economy, calculating sector-specific import intensity of domestic output, modeling price transmission from imported intermediates, and assessing multiplier effect leakage abroad.

Theoretically, the tables follow Approach A in input-output methodology [5, pp.142–151], where import disaggregation requires indirect estimation or supplementary data. From Leontief's perspective, approximation methods are mathematical exercises rather than substitutes for empirical data. Nevertheless, in the absence of primary statistics, estimation remains a pragmatic tool for generating preliminary insights.

3 Method

The data analysis workflow comprised the following sequential procedures:

A. Data Preparation

1. Conversion of source data (*The China Time-Series Input-Output Tables*) from `.xlsx` to `.RData` format, including necessary preprocessing.

B. Input-Output Matrix Construction

2. Derivation of the 18×18 direct input coefficients matrix ($\mathbf{A}$) as $\mathbf{A} = \mathbf{Z}\mathbf{D}_{\mathbf{x}}^{-1}$, where $\mathbf{Z}$ is the transactions matrix and $\mathbf{D}_{\mathbf{x}}^{-1}$ is the diagonal matrix with the reciprocals of sectors' gross outputs ($x_j^{-1}, \quad j = 1, \ldots, 18$) along its diagonal.
3. Construction of the 18×18 direct import coefficients matrix ($\mathbf{A}_m$) using the proportionality assumption: $a_{m,ij} = a_{ij} \times (m_i / z_i)$, where m_i represents total imports and z_i denotes total sectoral transactions for commodity i.
4. Computation of the 18×18 domestic direct input coefficients matrix ($\mathbf{A}_d$) as $\mathbf{A}_d = \mathbf{A} - \mathbf{A}_m$.

C. Import Intensity Analysis

5. Calculation of the Leontief inverse matrix: $\mathbf{L} = (\mathbf{I} - \mathbf{A}_d)^{-1}$, where $\mathbf{I}$ is 18×18 identity matrix.
6. Determination of sectoral import intensity (total import content of domestic output) for each year (1981–2018) via row-wise summation of the matrix product $\mathbf{A}_m\mathbf{L}$.

D. Temporal Pattern Analysis

7. Time-series clustering via Dynamic Time Warping (DTW) using the `dtwclust` R package (v.5.5.3; `type = "partition"`, `distance = "sbd"`, `k = 4`) [6, 7, 8].
8. Graphical representation of clustering results: temporal trajectories of sectoral import intensity with cluster centroids.

4 Results

Time-series clustering identified four distinct import intensity trajectories (Figure 1):

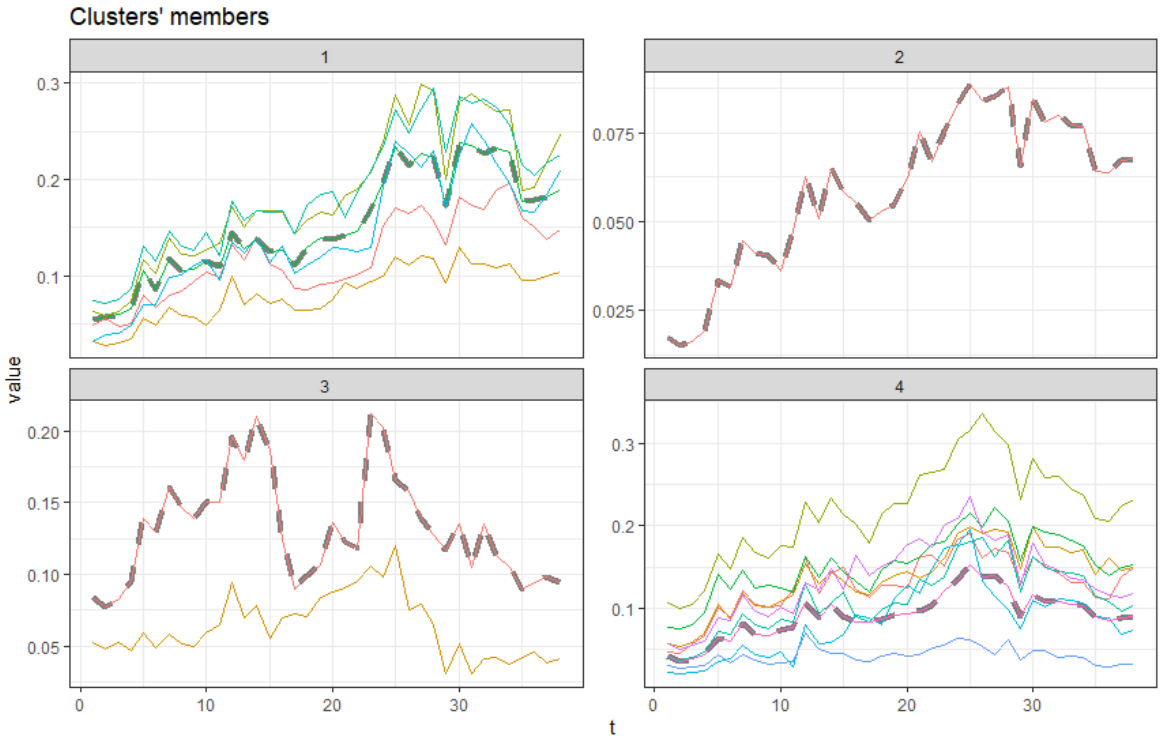


Figure 1: Time series clustering results with centroids

1. **Sustained growth**: Steady increase from 5% to 20–25% (avg.).
   Sectors: Mining; Food/tobacco; Petroleum/chemicals; Non-metallic minerals; Metals; Utilities.

2. **Moderate growth**: Progressive rise to 8% (avg.).
   Sector: Agriculture/forestry/fisheries.

3. **Volatile trend**: Alternating periods of increase and decrease.
   Sectors: Other manufacturing/repair; Wholesale/retail.

4. **Inverted U-shape**: Peak 2007–2008, followed by decline.
   Sectors: Textiles/apparel; Wood/paper/printing; Machinery/equipment; Construction; Transport/warehousing; IT services; Finance/real estate; R&D; Other services.

5 Discussion

This study presents a preliminary, data-driven exploration of import intensity dynamics. Given the estimated nature of the import matrices (derived via proportionality rather than direct observation), the identified clusters should be interpreted as indicative patterns, not definitive truths. Future research should rigorously test the sensitivity of results to alternative clustering parameters (e.g., distance metrics, number of

clusters) and, critically, seek validation using empirically observed import flow data where available.

References

1. Maddison A. (2007), Chinese Economic Performance in the Long Run, 960-2030 AD, Second Edition, Revised and Updated, Development Centre Studies, OECD Publishing, Paris, https://doi.org/10.1787/9789264037632-en

2. Teng J. (1996). Input-output analysis of economic growth and structural changes in China. *Journal of Applied Input-Output Analysis*, vol. 3. [Electronic resource] Mode of access: https://www.gakkai.ne.jp/papaios/en/img/jaia1996_3-2.pdf Date of access: 14.06.2025.

3. Bykov A.A., Tolkachev S.A., Parkhimenka U.A., Shablinskaya T.V. (2021). China's Economic Growth in 2010–2017: Analysis from the Perspective of the Input-Output Model and Modern Monetary Theory. *Finance: Theory and Practice*, no. 25(2), pp. 166–184.

4. Zhang H., Xia M., Su R., Lin C. Compilation of China's Time-Series Input-Output Tables: 1981–2018. *Statistical Research*, 2021, vol. 38, no. 11, pp. 3–23. [Electronic resource] Mode of access: http://ae.ruc.edu.cn/docs/2021-12/618767e3cf1843208c6d79ba8d10dd87.pdf Date of access: 14.06.2025.

5. Miller R.E., Blair P.D. (2022). Input-output analysis: foundations and extensions. Cambridge university press.

6. Package dtwclust. Reference manual [Electronic resource] Mode of access: https://cran.r-project.org/web/packages/dtwclust/dtwclust.pdf Date of access: 14.06.2025.

7. Paparrizos J., Gravano L. (2015). k-shape: Efficient and accurate clustering of time series. In: *ACM SIGMOD ICMD*, pp. 1855–1870.

8. Sard-Espinosa A. (2019). Time-Series Clustering in R Using the dtwclust Package. *The R Journal*, Vol. 11/1, pp. 22-43.

A TWO-SAMPLE TEST BASED ON MULTIVARIATE RANKS

N.V. Pastukhov[1]

[1]*Lomonosov Moscow State University*
Moscow, RUSSIA
e-mail: [1]nikita.pastukhov@math.msu.ru

The paper investigates a two-sample test based on multivariate ranks. Using the multivariate rank map, we transport the initial sample onto a polar grid. Then we find the polar ray with the maximal number of points of the first sample. Using it, we construct the test statistic. A limit theorem for this statistic is proved, and a Poisson approximation is established via Sevastyanov theorem. Finally, the proposed test is compared against existing multivariate two-sample tests.

Keywords: multivariate ranks, two-sample problem, optimal transport problem, Poisson approximation, Sevastyanov theorem

1 Introduction

Nonparametric two-sample tests are essential for assessing distributional homogeneity without parametric assumptions. In one-dimensional settings, a convenient nonparametric approach is to use ranks, but a direct extension to multivariate data is not straightforward. Chernozhukov [1] constructed a multivariate analog of ranks via optimal transport. In this work, we develop a multivariate two-sample test that is based on these ranks. The method assigns multivariate ranks by optimally transporting sample points to a structured reference grid composed of rays and concentric circles.

Let us formally define the optimal transport plan and multivariate ranks. Consider the sample points $\{x_i\}_{i=1}^n$, each carrying the unit mass, and denote by $A = \{y_j\}_{j=1}^n$ a fixed reference multiset in $\mathbb{R}^d$. We seek the transport plan $\Sigma = (\sigma_{i,j})_{i,j\leq n}$ that minimizes the total cost

$$T \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{i,j}\,\rho(x_i, y_j),$$

subject to the marginal constraints

$$\sum_{i=1}^{n} \sigma_{i,j} = 1, \quad \sum_{j=1}^{n} \sigma_{i,j} = 1, \quad \sigma_{i,j} \geq 0,$$

where $\sigma_{i,j}$ is the mass transported from x_i to y_j and $\rho(\cdot,\cdot)$ is the chosen distance metric. Under mild conditions on ρ and A, an optimal plan exists that induces a bijection between $\{x_i\}$ and $\{y_j\}$. The multivariate rank of x_i is then defined as its image under the transport map.

The reference grid A itself consists of evenly spaced rays half-lines emanating from the origin and circles with increasing radii.

To test the null hypothesis of the two-sample problem, we first pool the two independent samples into one combined dataset. Applying the optimal transport map, we

send each pooled point to a node on our reference grid. We then count the number of first-sample points assigned to each ray and use the largest of these values as our test statistic. It should be noted that this test is not effective for the general alternative, so we consider the particular alternative.

2 Model

Let $X_1, \ldots, X_n \sim F$ and $Y_1, \ldots, Y_m \sim G$ be independent samples with cumulative distribution functions F and G on $\mathbb{R}^d$. We test the null hypothesis

$$H_0 : \ F = G,$$

against the alternative hypothesis

$$H_1 : \ \mathbb{E}[X] \neq \mathbb{E}[Y].$$

Let Q be the radial grid with

$$a = \lfloor \sqrt{n+m} \rfloor \quad \text{rays}, \quad b = \left\lfloor \frac{n+m}{a} \right\rfloor \quad \text{points per ray.}$$

We construct a correspondence between points of the joint sample and points of Q via an optimal transport plan. We denote the number of observations from the first sample assigned to ray i by $X_{i,n}$. Consider the case $n = m$ for simplicity. Then under H_0 we have

$$X_{i,n} \sim \text{HyperGeom}(2n, n, \sqrt{n}),$$

and our test statistic is then given by

$$T_n \ = \ \max_{1 \leq i \leq \sqrt{n}} X_{i,n}.$$

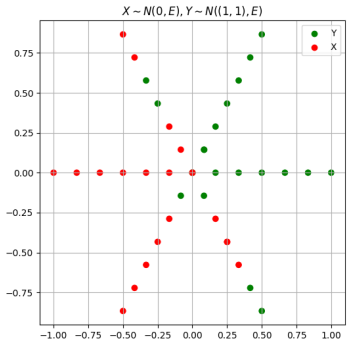


Figure 1: Result of transporting the combined sample from $\mathcal{N}((0,0), E)$ and $\mathcal{N}((1,1), E)$ (20 points each), where E denotes the 2×2 identity covariance matrix.

We choose as the test statistic the maximum number of sample points on any ray.

3 Main Results

Theorem 1. *Under the null hypothesis H_0 there exists the sequence $\{x_n\}$ such that*

$$\mathbf{P}\big(T_n \le x_n\big) \to e^{-\lambda}, \quad \sum_{i=1}^{\sqrt{n}} \mathbf{I}\{X_{i,n} > x_n\} \xrightarrow{d} \text{Poisson}(\lambda),\ n \to \infty,$$

where

$$\lambda = \frac{1}{s\sqrt{2\pi}}.$$

Here

$$x_n = \frac{\sqrt{n}}{2} + y_n \frac{n^{1/4}}{2},$$

where y_n is the unique solution of the equation

$$y_n^2 + 2\ln y_n = \ln n + 2\ln s. \tag{1}$$

So, for a given significance level α we can determine s, solve the equation (1) to find y_n and compute x_n. We then reject the null hypothesis of homogeneity if

$$T_n > x_n.$$

For small n the p-value can be approximated by simulating the hypergeometric counts under H_0.

In the report, we will present comparisons with well-known criteria and discuss the sensitivity of the proposed test.

References

1. Chernozhukov, V. [et. al.] (2015). *Monge-Kantorovich depth, quantiles, ranks and signs*. cemmap Working Paper No. CWP04/15. Centre for Microdata Methods and Practice (cemmap), London.

COMPARISON OF TWO APPROACHES FOR FINANCIAL TIME SERIES FORECASTING

Ya.D. Pleshakou[1], A.Yu. Kharin[2]
[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]pleshakouyauheni@gmail.com, [2]KharinAY@bsu.by

The problem of forecasting price movements in financial markets using historical data is a critical challenge in modern quantitative finance. This study focuses on comparing the effectiveness of machine learning (ML) methods [1], specifically XGBoost [2], with stochastic approaches based on Markov chains (MC) [3] and hidden Markov models (HMMs) [4] for predicting the direction of stock price changes in the S&P 500 index. Theoretical and empirical analyses are conducted, including data preprocessing, model implementation, and accuracy evaluation using classification metrics. The results provide insights into the strengths and limitations of each method, along with recommendations for future research. The research is partially supported by the National Science Foundation, Grant No. F023Uzb-080.

Keywords: financial time series, machine learning, Markov chain, hidden Markov model, XGBoost, stock price forecasting

1 Introduction

Modern financial markets are characterized by high volatility and complexity, necessitating advanced tools for analyzing and predicting price movements [5]. A key challenge in this domain is developing reliable models capable of forecasting price directions based on historical data. With the exponential growth of data and technological advancements, machine learning (ML) and stochastic methods [6] — such as Markov chains (MCs) and hidden Markov models (HMMs) — have gained prominence in financial time series forecasting.

The relevance of this study stems from the demand for robust predictive models to support investment decisions under uncertainty. While numerous methods exist for financial data analysis, hybrid and comparative approaches are increasingly adopted to evaluate the efficacy of different model classes. Specifically, the comparison between deterministic ML techniques (e.g., XGBoost) and probabilistic methods (e.g., MCs and HMMs) for short-term stock price direction forecasting remains an active research area.

The problem is well-studied in different scientific societies. ML methods, including XGBoost, have demonstrated high accuracy in classification and regression tasks, while MCs and HMMs remain widely used for time series analysis and stochastic process modeling. However, a comprehensive comparison of these approaches on a unified dataset and task is still lacking, underscoring the significance of this research.

The primary goal is to compare the performance of ML-based (XGBoost), MC, and HMM methods in predicting the direction of S&P 500 stock price movements.

2 Markov chain models

Let $(\Omega, \mathcal{F}, P)$ be a probability space, where $\{X_t\}_{t=1}^{T}$ is a sequence of random variables taking values in a finite state space $S = \{s_1, \ldots, s_k\}$. A **Markov chain** is a stochastic process satisfying the Markov property: $P(X_{t+1} = x_{t+1} \mid X_t = x_t, \ldots, X_1 = x_1) = P(X_{t+1} = x_{t+1} \mid X_t = x_t)$.

For financial applications, we typically consider: discrete-time processes with t representing trading days; finite state space $S = \{\text{down}, \text{neutral}, \text{up}\}$; homogeneous chains with time-invariant transition probabilities.

The behavior of a first-order Markov chain is characterized by its transition probability matrix $P = [p_{ij}]_{k \times k}$ where: $p_{ij} = P(X_{t+1} = s_j \mid X_t = s_i)$, $\sum_{j=1}^{k} p_{ij} = 1$.

For financial time series, states are typically defined via price return thresholds:

- s_1: Return $< -0.5\%$;
- s_2: $-0.5\% \leq$ Return $\leq 0.5\%$;
- s_3: Return $> 0.5\%$.

The maximum likelihood estimator for transition probabilities is $\hat{p}_{ij} = \frac{N_{ij}}{\sum_{m=1}^{k} N_{im}}$, where N_{ij} counts observed transitions $s_i \to s_j$.

A second-order Markov chain extends the dependency to two previous states:

$$P(X_{t+1} \mid X_t, X_{t-1}, \ldots, X_1) = P(X_{t+1} \mid X_t, X_{t-1}).$$

This requires a transition tensor $P^{(2)} = [p_{ijl}]$ where: $p_{ijl} = P(X_{t+1} = s_l \mid X_t = s_j, X_{t-1} = s_i)$.

The estimation becomes: $\hat{p}_{ijl} = \frac{N_{ijl}}{\sum_{m=1}^{k} N_{ijm}}$; N_{ijl} counts sequences $s_i \to s_j \to s_l$.

3 Machine learning methods

3.1 XGBoost framework

We consider the classification of financial time series using a machine learning approach based on gradient boosting. The goal is to predict the directional movement of asset returns using observed heterogeneous features such as lagged returns, volatility indicators, and technical metrics.

Let $x_1, x_2, \ldots$ denote the observed financial features, and suppose we model the output (e.g., direction of return) via an ensemble method. We employ the XGBoost algorithm (eXtreme Gradient Boosting), a regularized gradient boosting framework.

The XGBoost method optimizes a regularized objective function of the form:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{1}$$

where $l(y_i, \hat{y}_i)$ is a differentiable loss function (e.g., log-loss for classification), and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 + \alpha \|w\|_1$ is a regularization term controlling model complexity.

To build each tree f_t at iteration t, XGBoost uses a second-order Taylor approximation of the loss:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \tag{2}$$

where g_i and h_i are the first and second derivatives of the loss with respect to the previous prediction $\hat{y}_i^{(t-1)}$.

3.2 Hidden Markov models for regime-switching

Hidden Markov Models (HMMs) provide a probabilistic framework for modeling financial time series with latent regimes (e.g., bull or bear markets). They assume that the observed data is generated by an underlying unobserved Markov process.

The HMM is defined by parameters $\lambda = (S, A, B, \pi)$, where: $S = \{s_1, ..., s_N\}$ – hidden market regimes (states); $A = \{a_{ij}\}$ – transition probabilities between states; $B = \{b_j(o_t)\}$ – emission probabilities of observed returns/features; π – initial state distribution.

The model assumes the Markov property:

$$P(q_t \mid q_{t-1}, ..., q_1) = P(q_t \mid q_{t-1}), \quad P(o_t \mid q_t, o_{t-1}, ...) = P(o_t \mid q_t). \tag{3}$$

HMMs offer a powerful tool for identifying latent market regimes, capturing structural changes in time series, and improving interpretability in financial modeling. They are especially suitable for regime-switching trading strategies and volatility forecasting.

4 Results of comparative analysis

Let $\{P_t\}_{t=1}^{T}$ be a sequence of daily closing prices for S&P 500 constituent stocks from 2013-2018, where: $T = 1,508$ – trading days; $N = 505$ – unique stocks; returns are calculated as $r_t = (P_t - P_{t-1})/P_{t-1}$.

The target variable $y_t \in \{0, 1, 2\}$ represents:

$$y_t = \begin{cases} 0 & \text{if } r_t < -0.5\% \\ 1 & \text{if } -0.5\% \leq r_t \leq 0.5\% \\ 2 & \text{if } r_t > 0.5\% \end{cases} \tag{4}$$

XGBoost configuration:

$$\mathcal{L} = \sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)] + \gamma T + \frac{1}{2} \lambda \|w\|^2. \tag{5}$$

Markov models:

- First-order MC: $P = [p_{ij}]$, $p_{ij} = P(y_t = j | y_{t-1} = i)$;

- Second-order MC: $P = [p_{ijk}]$, $p_{ijk} = P(y_t = k | y_{t-1} = j, y_{t-2} = i)$;
- HMM: $\lambda = (A, B, \pi)$ with 3 hidden states.

The classification performance is evaluated by indicators called accuracy and F1-score. Results for comparison are presented in Table 1.

Table 1: Model Performance Comparison

Model	Accuracy	Balanced Accuracy	F1-score
XGBoost	0.730	0.740	0.740
1st-order MC	0.400	0.370	0.360
2nd-order MC	0.400	0.350	0.300
HMM	0.546	0.541	0.545

5 Conclusion

Summarizing the comparative analysis results, we have: XGBoost demonstrates superior predictive accuracy across all market regimes; HMM provides interpretable regime detection but with lower accuracy; traditional Markov chains showed in the experiments limited predictive power. Robustness analysis can be performed using the approach from [7].

References

1. Bishop C. M. (2020). *Pattern Recognition and Machine Learning*. Dialektika, Moscow. 960 pp.
2. Chen T., Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD*. ACM, New York. pp. 785-794.
3. Norris J. R. (1998). *Markov Chains*. 2nd ed. Cambridge University Press, Cambridge. 248 pp.
4. Blunsom P. (2004). *Hidden Markov Models*. Lecture Notes. 28 pp.
5. Tsay R. S. (2005). *Analysis of Financial Time Series*. 2nd ed. Wiley, Hoboken. 576 pp.
6. Murphy K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge. 1104 pp.
7. Kharin A. (2017). An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. sl Journal of Mathematical Sciences. Vol. **227** (2), pp. 196-203.

MONTE CARLO SSA FOR EXTRACTING WEAK SIGNALS

E. Poteshkin[1], N. Golyandina[2]

[1,2]*Saint Petersburg State University*
Saint Petersburg, RUSSIA
e-mail: [1]egor.poteshkin@yandex.ru, [2]n.golyandina@spbu.ru

The paper addresses the issue of extracting signals from noise using singular spectrum analysis (SSA). We propose an algorithm that automatically selects significant modulated harmonics without specifying their periods. This algorithm relies on the Monte Carlo SSA criterion to identify the significant frequencies, which are then extracted.

Keywords: time series, signal extraction, signal detection, singular spectrum analysis

1 Introduction

Consider the following model: $\mathsf{X} = \mathsf{S} + \mathsf{R}$, where $\mathsf{X} = (x_1, \ldots, x_N)$ is the observed time series, S is the signal and R is the noise, i.e., the realisation of some stationary process. This paper considers two problems: signal detection and signal extraction in cases where the signal is present.

To solve the first problem, we use the Monte Carlo SSA (MC-SSA) [2] method, which tests the hypothesis $H_0 : \mathsf{S} = 0$. The second problem is solved by applying singular spectrum analysis (SSA) [4, 7], which decomposes the series into elementary components. After grouping these components, we obtain a decomposition of the series into trend, periodic components, and noise. As one of the SSA steps involves visual analysis to identify the signal components, there is a need to automate this step. This problem has been addressed in works such as [1, 8, 3, 6], which are mainly devoted to trend extraction or smoothing. This paper aims to define an approach to the automatic extraction of weak oscillating signals detected by the MC-SSA criterion.

2 The autoMCSSA algorithm

Let us introduce some notation and assumptions.

The proposed algorithm uses a modified version of MC-SSA that corrects for multiple comparisons [5]. Sine waves with equidistant frequencies $\omega_k = k/(2L)$, $k = 1, \ldots, L$, were chosen as the projection vectors needed to construct the criterion statistics. With this choice of projection vectors, we can identify significant frequencies present in the signal.

For a series X of length N and $0 \leqslant \omega_1 \leqslant \omega_2 \leqslant 0.5$, we will use the same frequency measure used in [1] for trend extraction. Let the measure $T(\mathsf{X}; \omega_1, \omega_2)$ reflect the contribution of frequencies from the interval $[\omega_1, \omega_2)$ calculated from the periodogram of the series.

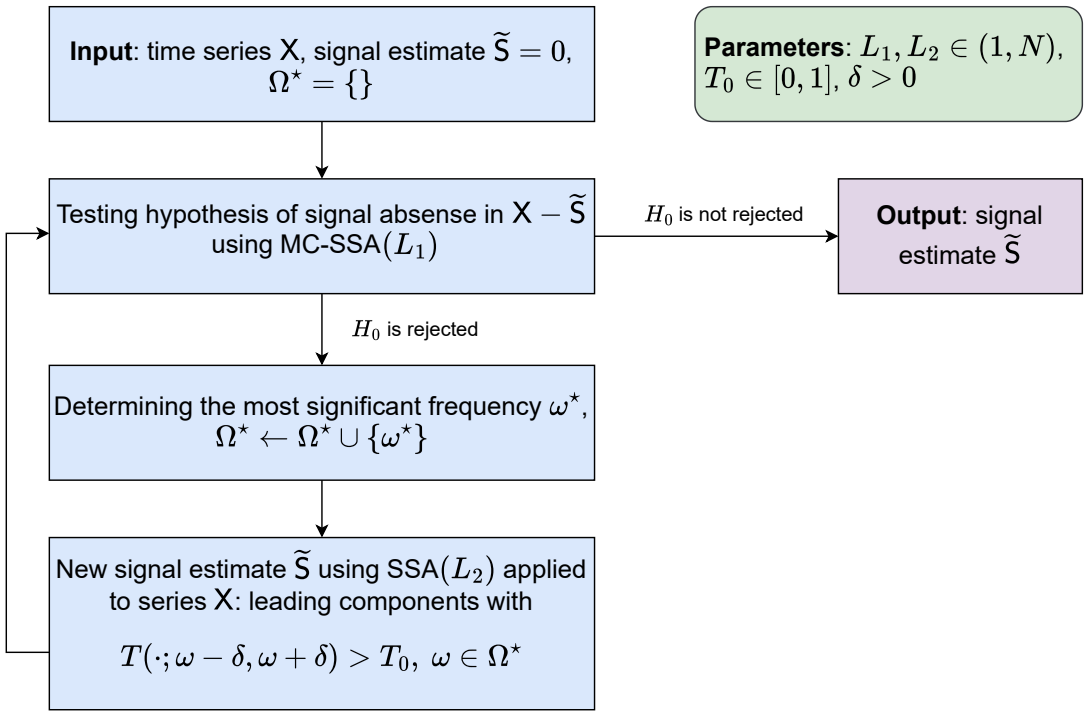


Figure 1: Flowchart of autoMCSSA

Figure 1 shows the flowchart of the autoMCSSA algorithm. It consists of applying the MC-SSA criterion sequentially until the hypothesis is no longer rejected. If the hypothesis is rejected at the next iteration of the algorithm, the most significant frequency $\omega^\star$ is determined and a new signal estimate is calculated using SSA applied to the original series: for each frequency ω from the set of significant frequencies $\Omega^\star$, the leading components with a measure T exceeding the threshold T_0 are selected. Once the null hypothesis is no longer rejected, the algorithm terminates, with the final signal estimate being given by $\widetilde{\mathsf{S}}$.

We will estimate the frequency $\omega^\star$ using a weighted average of the nearest significant frequencies, where the weights are determined by the frequencies' significance. This estimation method enables us to obtain a more accurate estimate of ω in cases when it does not fall on the $k/(2L)$ frequency grid.

Let us consider an example of the work of the proposed algorithm. Let $\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is the AR(1) model with parameters $\phi = 0.7$ and $\sigma^2 = 1$, $N = 200$, $\mathsf{S} = (s_1, \ldots, s_N)$,

$$s_n = 0.075\, e^{0.02n} \cos(2\pi n/8) + 2\cos(2\pi n/4) + 0.2 \cdot (-1)^n.$$

Figure 2 shows the first 15 elementary components reconstructed using SSA. The signal corresponds to components with indices 1, 2, 5, 6, and 13. Without knowing the formula by which this signal is generated, it is difficult to say with certainty which components are non-random, since the components (3, 4) and (11, 12) look like pairs

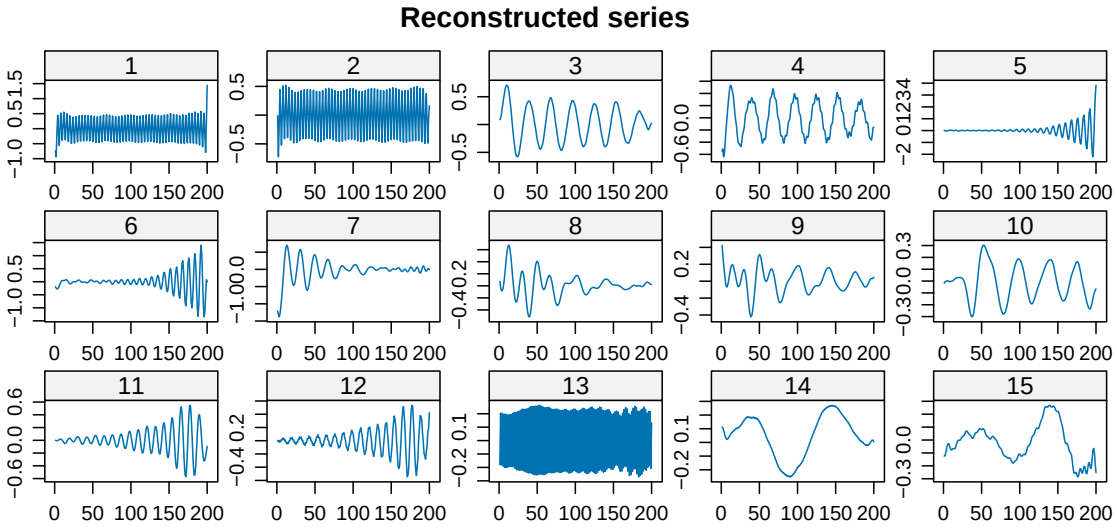


Figure 2: Elementary series components

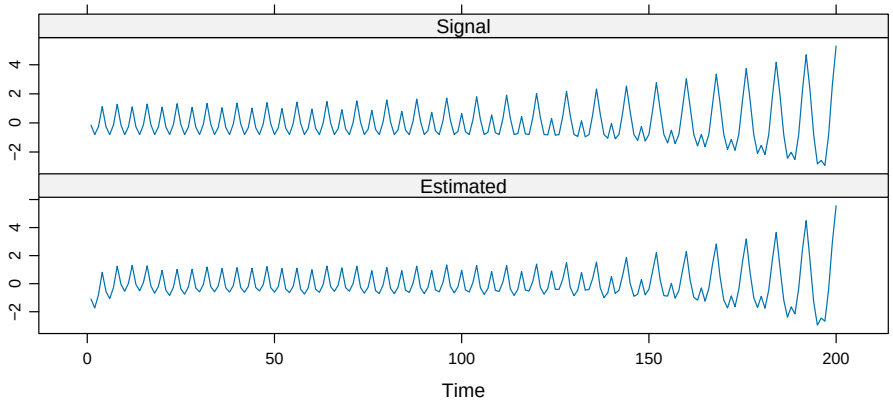


Figure 3: Estimated signal using autoMCSSA ($L_1 = 50$, $L_2 = 100$, $\delta = 1/80$, $T_0 = 0.5$)

of harmonics. We applied the autoMCSSA algorithm to this series and found that the developed method correctly identified the components corresponding to the signal. Figure 3 shows the signal S and its estimate by the autoMCSSA method.

3 Conclusion

The paper proposes an algorithm capable of extracting only the significant time series components without specifying the period of a periodic component. Multiple modulated periodic components may be present, each with different amplitude modulation.

References

1. Alexandrov Th. (2009). A method of trend extraction using singular spectrum analysis. *RevStat*. Vol. **7**, Num. **1**, pp. 1-22.

2. Allen M., Smith L. (1996). Monte Carlo SSA: detecting irregular oscillations in the presence of coloured noise. *Journal of Climate*. Vol. **9**, pp. 3373-3404.

3. Bogalo J., Poncela P., Senra E. (2021). Circulant singular spectrum analysis: A new automated procedure for signal extraction. *Signal Processing*. Vol. **179**, 107824.

4. Broomhead D., King G. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*. Vol. **20**, Num. **2-3**, pp. 217-236.

5. Golyandina N. (2023). Detection of signals by Monte Carlo singular spectrum analysis: multiple testing. *Statistics and Its Interface*. Vol. **16**, Num. **1**, pp. 147-157.

6. Golyandina N., Dudnik P., Shlemov A. (2023). Intelligent identification of trend components in singular spectrum analysis. *Algorithms*. Vol. **16**, Num. **7**, 353.

7. Golyandina N., Nekrutkin V., Zhigljavsky A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman and Hall/CRC: New York.

8. Kalantari M., Hassani. H. (2019). Automatic grouping in singular spectrum analysis. *Forecasting*. Vol. **1**, Num. **1**, pp. 189-204.

ASSESSING CONTRACTOR CONNECTIONS BASED ON BIG DATA SETS

D.M. Rahel[1]

[1]*Belarusian State University of Informatics and Radioelectronics*
Minsk, BELARUS
e-mail: [1]ragel@bsuir.by

When assessing the risk of transactions, companies or financial and credit organizations have a question about the specifics of this procedure in relation to interconnected companies that are part of a conglomerate or have any obligations to other market participants. We considering a problem of assessing contractor connections using big data sets.

Keywords: big data, contractor connections, risk of transaction

In the case of interconnected companies, it makes sense to start with the general rating of the group of participants to which it relates in accordance with the preliminary assessment. It is necessary to determine the significance of such an entity in the conglomerate, that is, this task can be solved by assigning it to one of the specified categories:

1. A **managing or significant company** that owns the largest amount of assets of the community in question or the largest revenue among its members;

2. A **dependent company**, i.e. a company of the group in question that does not have a significant amount of assets, whose creditworthiness depends on the entire group, at least the balance of payments of the latter has a significant impact on this type of enterprise.

The links between companies are hidden and the nature of the relationship is not advertised, but with sufficient data, we can characterize, if there are large data sets, the activities of the counterparties in question over several periods of time. In this case, we used a data set that characterized the activities of 100 similar counterparties for 8 months of 2023. We had assumptions about the connection of some of them, and using standard procedures, we confirmed the connection of their business activities, which allowed us to adjust the scoring assessments and affected the possibilities of issuing them credit funds in the future.

To operate this information within the framework of analytical procedures and make initial conclusions characterizing the dynamics of the industry or market under consideration, it is necessary to conduct a classification for the purpose of further studying groups of such entities with common quantitative characteristics. As the simplest and most obvious solution at the first stage, an algorithm for constructing an interval variation series was used using the **Sturges formula** as a tool for classifying the obtained values of the dynamics of economic activity.

Based on the table data, $X_{\min} = 0.54$ million rubles and the maximum, as the upper limit - $X_{\max} = 6.55$ million rubles.

The variation range in this case will be $R = 6.01$. According to the calculations, the following optimal interval duration is obtained: $H = 0.7863$. This means that our values will be divided into **8 groups** for further analysis of frequencies and distribution.

Based on the data obtained, the image of the objective picture in the sample under consideration can be displayed as a polygon of the distribution of values by the enterprises of the studied market (Figure 1). In this case, the asymmetry of the tails of the polygon under consideration is obvious, as well as the shifts in the central part of the graph, such a situation requires a more detailed further interpretation.

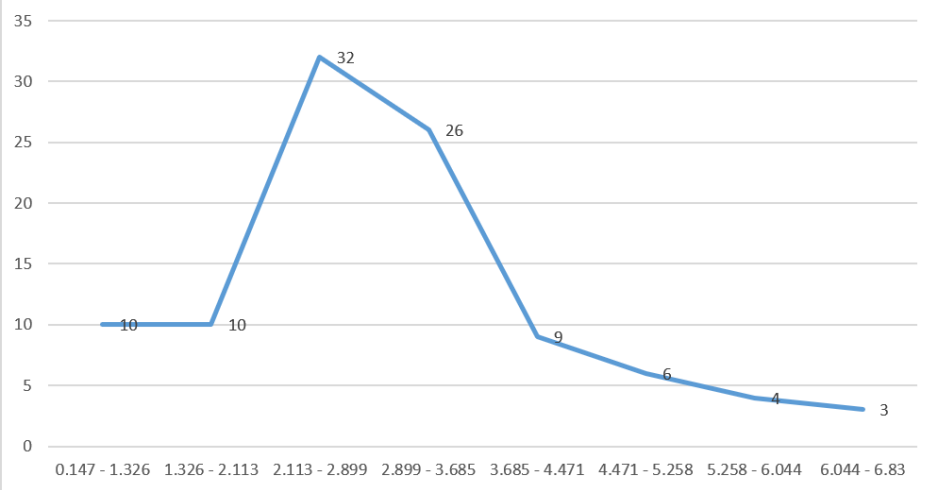


Figure 1: Distribution polygon of enterprise values in the studied market

The obtained results indicate that the largest number of participants belong to the average range of values, i.e. 32 legal entities are in the range from RUB 2.113 million to RUB 2.899 million and 26—from RUB 2.899 million to RUB 3.685 million. The maximum range included 3 business entities, the minimum—10. At the same time, 48 companies achieved a value above the average—RUB 2.93 million. More specific results can be obtained during further stages of the analysis.

It is necessary to pay attention to the nature of the data distribution in our sample; it is characterized by asymmetry with a shift to minimum values. Thus, we are talking about the need to search for a factor that influenced business entities. The analysis should be continued in relation to the cause that caused this situation in the data set.

At the second stage, it is necessary to clarify the significance of the mutual influence of the companies under consideration on each other. This can be done based on the application of clustering algorithms. In this case, we will compare the two most common types and compare the reliability of the results that data clustering gives based on the calculation of the Euclidean distance between the data in the considered set and the results that can be obtained using the agglomerative clustering algorithm, which is calculated based on the Mahalanobis distance formula. The difference between the algorithms used is in considering the determination between the elements of the considered data array in the case of agglomerative clustering and the lack of consideration of dependencies in the case of calculating clusters based on the Euclidean distance.

In the event of a significant difference in the division of data into clusters, we will be able to conclude that there are unaccounted factors that affect the elements of the array, as well as make initial conclusions about the mutual influence of the elements under consideration and based on this, specify further analysis for the connectivity of economic activities.

When performing clustering of the data under consideration based on the Euclidean distance, several assumptions and prerequisites necessary for further classification of the economic data under consideration were used. The following characteristics were used to configure the algorithm: 2 clusters (based on empirical assessment), iterations stopped obtaining equivalent values of distances between groups (Figure 2).

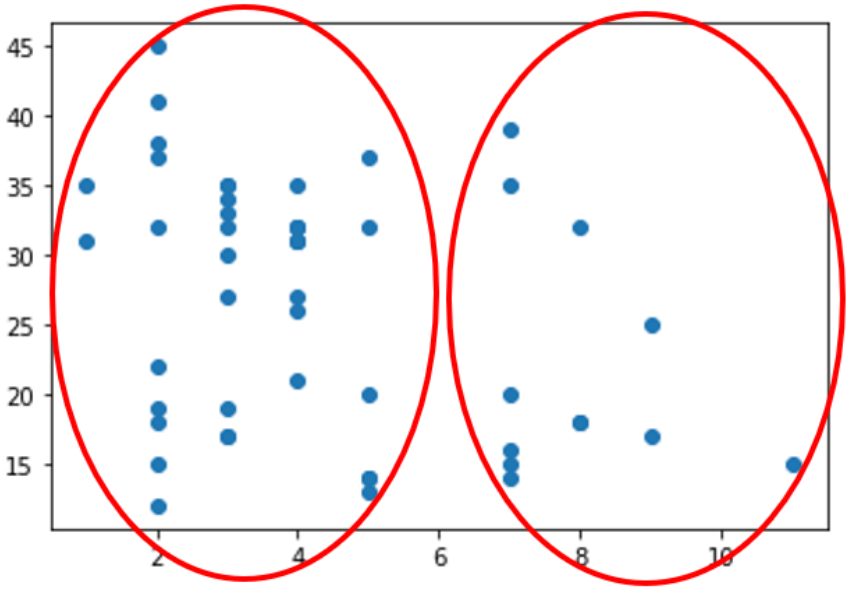


Figure 2: Results of data clustering based on the calculation of the Euclidean distance using the nearest neighbor method

Figure 2 shows the clusters formed based on the algorithm under consideration; because of the implementation, no clear cluster groups were formed, and it is impossible to draw conclusions about the essential features and interrelationships of the counterparties considered in the analysis.

When using the agglomerative clustering algorithm, which implies taking into account the correlation between the data under consideration and, due to this, is invariant to the scale of the volume of data under consideration, we obtained a picture that changes based on the fact that when calculating the distance between elements in this case, the covariance between the elements is taken into account and it is this type of assessment that will allow for further analysis on the interrelationships of the elements being assessed (Figure 3).

In this case, we see the division of groups considering the interconnectedness of their activities and the consistency in the dynamics of the data provided. We cannot clearly speak about interdependence based only on the implementation of this algorithm for clustering the obtained data, but at the same time we determine the direction of further

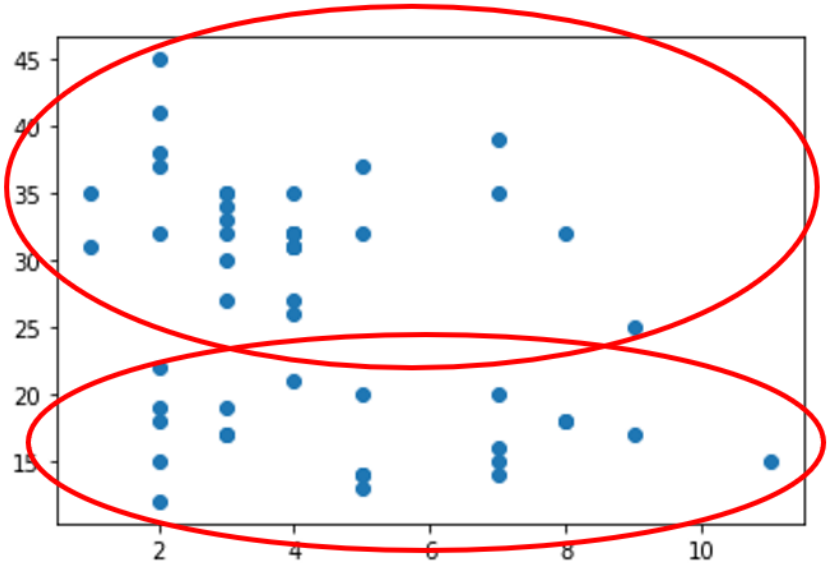


Figure 3: Results of data clustering based on the calculation of the Mahalanobis distance using the nearest neighbor method

analysis. It can be assumed that at the initial stage, companies similar in activity, or united based on some currently undetected features, ended up in one cluster, which allows us to make initial conclusions about their interconnectedness and specify further analytical procedures.

References

1. Nicholson W.L. (2012). Exploring Data Analysis. Nobel Press: Oakland.

RELATIVITY PRINCIPLE AND MEASURE THEORY

V.M. Romanchak[1], P.M. Lappo[2]

[1]*Belarusian National Technical University*
[2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]romanchak@bntu.by, [2]lappopm@gmail.com

We discuss connections between relativity principle and measure theory.

Keywords: relativity principle, affine space, measure, ratio scale, pairwise comparisons

1 Introduction

Measure in mathematics generalizes and formalizes intuitive notions of size: length, area, volume, as well as mass and probability. These seemingly distinct concepts are unified through a common mathematical approach and play a key role in probability theory and integration. The mathematical construct of measure unifies heterogeneous intuitive concepts (length, probability) via axioms of finite or countable additivity. However, classical definitions of measure conflict with the principle of relativity in physics, which denies absolute reference frames as formulated by Galileo and Einstein. In measure theory, $\mu(\emptyset) = 0$ is a technical condition ensuring additivity. Fixing $\mu(\emptyset) = 0$ resembles choosing an "absolute zero" for measurements. In physics, this contradicts the relativity principle, where measurements depend on the reference frame [1,2]. Example: attempting to define a measure $\mu(A) = \mu(A) + c$ violates the axiom $\mu(\emptyset) = c \neq 0$.

2 Interconnection of Vector Spaces and Set Algebras

Let $V = \mathbb{R}$ be a vector space over the field $\mathbb{R}$ and F a set algebra. Define a mapping $v : F \to V$.

Additivity

The mapping v is additive if for disjoint sets $A, B \in F$:

$$v(A \cup B) = v(A) + v(B).$$

By definition, $v(\emptyset) = 0$.

Affine Interpretation

The vector space V can be interpreted as an affine space with a fixed reference point O. Unlike vector spaces, affine spaces lack a distinguished origin [3]. The difference $v(A) - v(B)$ corresponds to relative displacement between points, aligning with the relativity principle. If $B \subseteq A$, then $v(A \setminus B) = v(A) - v(B)$, defining a relative "measure of difference."

Example of Affine Measure Application

Let $A_1, A_2, \ldots, A_n$ be objects (e.g., players) in pairwise comparisons. Matrix $K = \|K_{ij}\|$ contains K_{ij} victories of A_i over A_j. Total matches between A_i and A_j: $K_{ij}+K_{ji}$. The pairwise comparison matrix $M = \|M_{ij}\|$ is defined as:

$$M_{ij} = \frac{K_{ij} - K_{ji}}{K_{ij} + K_{ji}}, \quad i \neq j.$$

Diagonal elements $M_{ii} = 0$. M is skew-symmetric ($M_{ij} = -M_{ji}$). Assume M_{ij} relates to parameters via $M_{ij} = a_i - a_j$. This is an affine measure model (differences of measures), where parameters are points in an affine space (object ranks). Matrix elements M are vectors from point a_i to a_j. Parameter estimation involves matrix transformation and structural analysis.

3 Parameter Estimation Algorithm

1. **Matrix Transformation**: Transform M into M'. For each row i:

   $$M'_{ij} = M_{ij} - M_{i1}.$$

   This removes dependence on the "reference point" (first object), analogous to changing reference frames in physics.

2. **Transformed Matrix Analysis**: If $M_{ij} = a_i - a_j$ holds, rows of M' will be identical. For real data, estimate parameters a_j as column averages of M':

   $$a_j = \frac{1}{n} \sum_{i=1}^{n} M'_{ij}.$$

   Here, $a_1 = 0$ (reference point).

3. **Model Adequacy Check**: Compute Pearson correlation coefficients between rows i and k of M'.

4 Example

Rating statistical journals based on citations (dataset `citations` in R package `BradleyTerry2`). Transforming M into M' (subtracting the first row element) resulted in nearly linearly dependent rows [4].

5 Result Interpretation

1. High row correlations: All coefficients near 1 indicate strong linear dependence.
2. Highest correlation $r_{13} \approx 0.999$ signals near-ideal linearity.
3. Lowest correlation remains very high, confirming overall similarity. This validates the feasibility of affine data analysis models.

References

1. Whitehead, A. N. (2005). *Principle of Relativity*. Barnes & Noble Publishing.
2. Tao, T. (2011). An Introduction to Measure Theor. Vol. 126, American Mathematical Soc.
3. Barzilai, J. (2005). Measurement and preference function modelling. International Transactions in Operational Research.
4. Turner, H., Firth, D. (2012). Bradley-Terry models in R. *Journal of Statistical Software*

COMPARATIVE ANALYSIS OF MACHINE AND DEEP LEARNING ALGORITHMS IN NETWORK TRAFFIC ANOMALY DETECTION

T.T. Safiullin[1]

[1]*Belarusian State University*
[1]*Department of Mathematical Modeling and Data Analysis*
Minsk, BELARUS
e-mail: [1]tuleubay.safiullin@mail.ru

The paper presents the results of research on the use of machine learning algorithms and neural networks for detecting anomalies in corporate network traffic. A representative set of test data is used to study the effectiveness of classification algorithms and their ensembles, as well as algorithms for detecting abnormal observations in the presence and absence of a training sample, respectively. It has been established that the use of ensembles of machine learning algorithms, as well as neural network algorithms, allows achieving high efficiency. anomaly detection. The results obtained are of practical importance for the development of systems for detecting and preventing cyberattacks in corporate networks, and also open up prospects for further improvement of network traffic protection technologies.

Keywords: cybersecurity, anomaly detection, network traffic

1 Introduction

The corporate network is crucial for modern organizations, and anomalies due to cyber attacks can lead to significant economic and reputational damage. Early detection of threats is key to resilience strategies. Constant monitoring of network traffic helps identify potential threats, and machine learning algorithms can detect and block attacks promptly, reducing losses and speeding up recovery.

In Belarus, cybercrimes accounted for over 25% of all registered crimes in 2024 [1]. Similarly, in Russia, financial fraud losses reached 250–300 billion rubles in 2024, with countermeasures expected to reduce this figure in 2025 [2].

The purpose of this paper is to investigate the effectiveness of using the most commonly used in various applications machine learning and artificial intelligence algorithms based on neural networks to solve the problems of anomaly detection in corporate network traffic. The choice of the most effective algorithms for different types of cyberattacks is a key aspect in ensuring cyber security and stable operation of information systems, which determines the relevance of this study.

2 Algorithms and indicators of their efficiency

The study employs two main algorithm categories: (1) machine learning (classification, anomaly detection, and ensemble methods) and (2) neural networks for network traffic analysis.

Classification algorithms (LR, SVM, KNN) require labeled data with two classes [3], [4]: anomalies/illegitimate traffic and normal traffic. For unlabeled or imbalanced data, anomaly detection methods are used [5]: One-Class SVM, IF, LOF, and EE.

Ensemble methods include RF, LightGBM, and stacking/blending/bagging [6]. Neural networks comprise FNN, RNN [5], and LSTM [7] architectures.

Optimal hyperparameters for each algorithm were determined through grid search - an exhaustive parameter tuning method that evaluates all predefined value combinations. These configurations, set prior to model training, directly impacted both the learning process and final performance.

The most accurate ML and neural network algorithms were combined into classifier ensembles. Their performance was evaluated using four key metrics:

- Precision: Ratio of correctly predicted "traffic anomaly" cases among all instances classified as anomalies
- Recall: Ratio of correctly identified anomalies among all actual anomaly cases
- F1-score: Harmonic mean of precision and recall (balancing both metrics)
- AUC: Model's class discrimination ability (1 = perfect classification)

3 Dataset and Methodology

The study employs the CICIDS2017 dataset - a comprehensive benchmark for intrusion detection containing 2.8 million network records (83% normal traffic, 17% attacks including DoS and scanning) collected over 5 days. Data preprocessing was performed using Python 3.11 with pandas and scikit-learn, focusing on feature scaling [8], categorical encoding, and handling class imbalance through robust metrics.

The methodology follows four key stages:

- Utilizing pre-collected network packet data (originally captured via Wireshark [9])
- Data normalization and feature engineering [10]
- Model training with optimized ML classifiers [11], anomaly detection algorithms [12], and neural networks [13]
- Comprehensive evaluation using precision, recall, F1-score and AUC metrics

4 Experimental Study of Algorithm Performance

This chapter presents a comprehensive evaluation of anomaly detection methods in corporate network traffic, identifying the most promising approaches for practical implementation. Visual results are summarized in Figure 1 for direct comparison.

Classification Algorithms

- SVM (RBF kernel) demonstrated the highest effectiveness (Precision: 0.8807, Recall: 0.9430, F1: 0.9107) but required significant computational resources.
- LR and KNN showed lower performance compared to SVM.

Anomaly Detection Algorithms

- One-Class SVM outperformed other methods (Precision: 0.90, Recall: 0.90), correctly identifying 90% of anomalies with minimal false positives.
- IF, LOF, and EE were trained on a subset with only 5% anomalies, yielding lower metrics.

Neural Networks

- FNN and LSTM achieved the best results (Recall: 0.943), balancing precision and recall effectively.
- RNN underperformed due to difficulties with long-term dependencies.

Ensemble Methods

- Blending (SVM + LSTM + RF) achieved the highest performance, with only 3.1% missed anomalies and 2.9% false positives.
- Stacking and LightGBM produced identical results, slightly below blending.

Blending ensembles deliver the best overall results, minimizing both missed detections and false alarms.

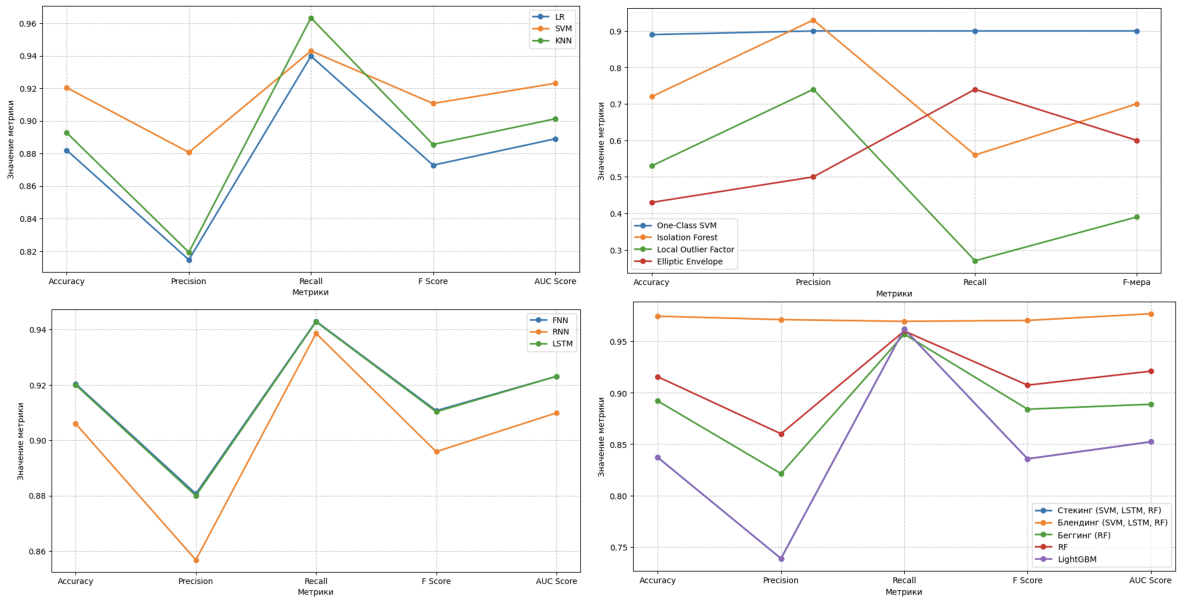


Figure 1: Performance results

References

1. Belta.by [Electronic resource] Mode of access: `https://shorturl.at/mmvzf` Date of access: 14.05.2025

2. Ria.ru [Electronic resource] Mode of access: `https://ria.ru/20250111/moshenniki-1993226383.html` Date of access: 18.04.2025

3. Kharin Yu.S., Malyugin V.I., Abramovich M.S. (2008) *Mathematical and Computer Foundations of Statistical Modeling and Data Analysis*. Minsk: BSU, 455 p.

4. Aurlien G. (2020) *Applied Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools*. Dialektika, pp. 163-170.

5. Bishop C. (2020) *Pattern Recognition and Machine Learning*. Williams, pp. 331-337.

6. Raschka S. (2017). *Python and Machine Learning*. DMK Press, pp. 202-221.

7. Goodfellow I., Bengio Y., Courville A. (2018). *Deep Learning*. DMK Press, p. 35.

8. Kamalov F. [et. al.] (2024) Forward Feature Selection: Empirical Analysis. *Journal of Intelligent Systems and Internet of Things*. Vol. **11**, no. **1**, pp. 44-54.

9. Setiawan K., Wibowo A. (2023). Data Mining Implementation for Detection of Anomalies in Network Traffic Packets Using Outlier Detection Approach. *Jurnal Informatika dan Komputer*. Vol. **6**, no. **2**, pp. 79-87.

10. Dai S., Zhao Y., Huang J. (2023). Online Network Traffic Anomaly Detection Method Combining OS-ELM and SADE. *IEEE Access*. Vol. **11**, pp. 3645-3658.

11. Vikram A., Mohana. (2020). Anomaly Detection in Network Traffic Using Unsupervised Machine Learning Approach. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 476-479.

12. Yu B. [et. al.] (2023). A Network Traffic Anomaly Detection Method Based on Gaussian Mixture Model. *Electronics*. Vol. **12**, no. **6**, pp. 1-15.

13. Laxman S., Hiwarkar T. (2022). A Result Analysis of Supervised Machine Learning Approach to Detect Anomaly from Network Traffic. *International Journal of Computer Science and Mobile Computing*. Vol. **11**, no. **6**, pp. 152-168.

ON THE STOCHASTIC MODEL OF A CLOUD COMPUTING SYSTEM

D.A. Salnikov[1], T.V. Rusilko[2]

[1,2] *Yanka Kupala State University of Grodno*
Grodno, BELARUS
e-mail: [1]dima.saln.gr@gmail.com, [2]tatiana.rusilko@gmail.com

The paper presents a stochastic model of a cloud computing system in the form of a G-network with rewards. The purpose of the study is to analyze and calculate the accumulated reward of a cloud computing system taking into account the costs associated with damaged, obsolete, or canceled requests. The reward sequence is formed as a result of the transmission of requests between network nodes. The results of mathematical modeling allow us to predict the total expected reward of the queueing model as a function of the remaining time for a given initial state.

Keywords: queueing model, cloud computing system, G-network, network with rewards, asymptotic analysis method

1 Introduction

Cloud computing systems represent one of the most promising and actively developing paradigms in the field of distributed computing. Cloud platforms, most famously Amazon Web Services, Microsoft Azure, and Google Cloud, connect physically distributed data centers to create a single virtual computing space.

Cluster systems are typically located in a single physical or corporate center (e.g., a single data center or campus) and are owned by a single organization. In contrast to cluster systems, cloud computing platforms and their resources are combined into huge pools that can be distributed across multiple data centers and even regions. Cloud providers manage these resources centrally, but users receive an abstraction of the physical infrastructure, allowing them to use the resources without being tied to a specific location.

A key element of such systems is virtualization, which allows you to create and scale virtual machines, containers, or serverless functions. Users are provided with a high-level virtualized environment that hides the physical infrastructure and allows dynamic allocation and scaling of resources on demand. Services can automatically scale according to current needs, allowing them to operate efficiently both during peak loads and periods of low activity.

At the physical level, the architecture of cloud computing systems includes servers, data storage systems, and a network that connects computing nodes. Let us define a model of a cloud computing system using a closed structure queueing network. First, it is necessary to establish a correspondence between the architecture of the cloud computing system and the queuing network. We will assign queueing systems to cloud servers or computing nodes. Each server receives incoming requests and places them in a queue for execution. All requests, after being processed by the server, are either

considered fully executed and leave the system, or are distributed to another server for further processing. The execution of requests and their transmission between the nodes of the queueing model are specified using the transition probability matrix. Routing between servers, clusters, and data centers determines the structure of the queueing network and the existing connections between its nodes. In order to take into account requests of different types, including corrupted or canceled, it naturally makes sense to use a generalized queueing network with requests of several types, namely a G-network. For executing requests, the cloud system receives some reward, usually payment according to a set tariff, so the model needs to establish a mechanism for accounting for the reward earned. For this reason, the G-network with rewards is used as a model in this paper.

2 Model

As a model of a cloud computing system, we consider a closed exponential G-network composed of a finite set of nodes S_0, S_1, ..., S_n. Let K denote the total number of requests circulating within the network, where each request represents a task being processed by the system. The node S_0 serves as an IS-node, consisting of K identical exponential servers, and S_0 functions as an abstract finit source of requests with capacity of K requests. The node S_0 generates requests only at the moment of request arriving at its input; the requests it generates load the network of nodes S_1, S_2, ..., S_n. Suppose that node S_0 generates a Poisson flow of arrivals at a rate of $\lambda_0 k_0$, where λ_0 is the flow parameter and k_0 is the number of customers present at S_0. This incoming flow is divided into two classes of requests. Regular operational tasks are categorized as positive requests, while corrupted, outdated, or canceled tasks fall under the negative class. In a small time interval Δt, the probability of a positive request arriving is given by $\lambda_0 k_0 p_{0i}^{+}\Delta t + o(\Delta t)$ and similarly, the probability of a negative arrival is $\lambda_0 k_0 p_{0i}^{-}\Delta t + o(\Delta t)$, $i = \overline{1,n}$, and $\sum\limits_{i=1}^{n}\left(p_{0i}^{+} + p_{0i}^{-}\right) = 1$.

The cloud computing system is modeled by the G-network of nodes S_1, S_2, ..., S_n. Each node S_i operates as a queueing system with identical m_i servers and an unlimited buffer for positive requests. In any small time interval Δt, the probability that node S_i completes the service of a positive request is $\mu_i \min(m_i, k_i)\Delta t + o(\Delta t)$, where k_i is the number of requests currently at node S_i, and $o(\Delta t)$ represents the probability of two or more requests being completed simultaneously. Moreover, the events corresponding to the completion of the service at different nodes during time Δt are mutually independent. The queueing algorithm is FIFO. A serviced request is immediately transmitted from node S_i to node S_j: as a positive request with probability p_{ij}^{+}, or as a negative request with probability p_{ij}^{-}. If the request is fully executed, it leaves the cloud system and is transferred to S_0 with probability $p_{i0}^{+} = 1 - \sum\limits_{j=1}^{n}\left(p_{ij}^{+} + p_{ij}^{-}\right)$, where $i \neq j$, $i, j = \overline{1,n}$. Negative requests are not processed by the node servers; they act as signals. In particular, when a negative request arrives at node S_i, $i = \overline{1,n}$, it immediately cancels one positive request present at this node. The incoming negative

request and the canceled positive request are immediately routed to the IS-node S_0 as positive requests. Alternative strategies for handling negative requests can certainly be explored.

The state of the network model under study at time t is represented by a continuous-time Markov process in the finite state space

$$\mathbf{k}(t) = (k_1(t), k_2(t), \ldots, k_n(t)),$$

where $k_i(t)$ denotes the number of requests at node S_i at time t. The number of requests at the IS-node S_0 is given by $k_0(t) = K - \sum_{i=1}^{n} k_i(t)$.

Suppose that the system earns R_{ij}^{+} conventional units (c. u.) when a positive request transitions from node S_i to node S_j, and the system earns R_{ij}^{-} c. u. when a negative request makes the same transition, $i \neq j$, $i, j = \overline{0, n}$. We call R_{ij}^{+} and R_{ij}^{-} the "rewards" associated with the transitions of positive and negative requests, respectively, from S_i to S_j [1]. Let us assume that the model receives a reward at rate of $R(\mathbf{k}, t)$ c. u. per unit time it occupies the state $\mathbf{k}$. Let $V(\mathbf{k}, t)$ denote the expected total model reward that the G-network will earn in time t if it starts in the state $\mathbf{k}$. The central question is: what are the expected total earnings $V(\mathbf{k}, t)$ of the model in time t, if the current state of the network is $\mathbf{k}$.

The main purpose of asymptotic methods in queueing theory is to study the servicing processes of queueing networks by finding suitable approximations for them under the specific limit assumption [2]. Cloud computing systems undoubtedly handle a large number of customer requests. In connection with this, consider the important asymptotic case of a large number of requests K. The asymptotic technique used is described in [3, 4]. We use the passage to the limit from a Markov chain $\mathbf{k}(t)$ to a continuous-state Markov process $\boldsymbol{\xi}(t) = \left(\frac{k_1(t)}{K}, \frac{k_2(t)}{K}, \ldots, \frac{k_n(t)}{K}\right)$ as K tends to be large [5]. In the context of the asymptotic approximation, we use the notation $v(\mathbf{x}, t)$ to represent the reward density, where $\mathbf{x}$ is the initial state and t is the remaining time. As a result, we obtain a generalized multidimensional Kolmogorov backward equation with an added earning component $q(\mathbf{x}, t)$:

$$\frac{\partial v(\mathbf{x}, t)}{\partial t} = -\sum_{i=1}^{n} A_i(\mathbf{x}, t) \frac{\partial v(\mathbf{x}, t)}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{n} B_{ij}(\mathbf{x}, t) \frac{\partial^2 v(\mathbf{x}, t)}{\partial x_i \partial x_j} + q(\mathbf{x}, t), \tag{1}$$

$$A_i(\mathbf{x}, t) = \sum_{j=1}^{n} \mu_j \min(m_j/K, x_j)(p_{ji}^{-} - p_{ji}^{+} + \delta_{ji}) +$$

$$+\mu_i \min(m_i/K, x_i) \sum_{j=1}^{n} p_{ij}^{-}(1 - \theta(x_j)) - \lambda_0 \left(1 - \sum_{i=1}^{n} x_i\right) \left(p_{0i}^{+} - p_{0i}^{-}\right),$$

$$B_{ii}(\mathbf{x}, t) = \frac{1}{K} \sum_{j=1}^{n} \mu_j \min(m_j/K, x_j)(p_{ji}^{+} + p_{ji}^{-} + \delta_{ji}) +$$

$$+\mu_i \min(m_i/K, x_i) \sum_{j=1}^{n} p_{ij}^{-}(1 - \theta(x_j)) + \lambda_0 \left(1 - \sum_{i=1}^{n} x_i\right) \left(p_{0i}^{+} - p_{0i}^{-}\right),$$

$$B_{ij}(\mathbf{x},t) = \frac{1}{K}\mu_i \min(m_i/K, x_i)\left(p_{ij}^- - p_{ij}^+\right), i \neq j,$$

where δ_{ji} is the Kronecker delta, $\theta(x)$ is the Heaviside step function.

Equation (1) is not explicitly solvable, so we have to apply a further approximation. Notice that the diffusion coefficients $B_{ij}(\mathbf{x},t)$ of equation (1) are of order ε. Therefore, up to terms of order $\mathrm{O}(\varepsilon^2)$, the reward density satisfies the equation:

$$\frac{\partial v(\mathbf{x},t)}{\partial t} = -\sum_{i=1}^{n} A_i(\mathbf{x},t)\frac{\partial v(\mathbf{x},t)}{\partial x_i} + q(\mathbf{x},t). \tag{2}$$

By integrating the equation (2) within a n-dimensional region D we obtain a first order ordinary linear differential equation for the expected reward $V_D(t)$:

$$\frac{d}{dt}V_D(t) = \sum_{i=1}^{n} \frac{\partial A_i(\mathbf{x},t)}{\partial x_i} \cdot V_D(t) + \int\limits_D \int \ldots \int q(\mathbf{x},t)d\mathbf{x}, \tag{3}$$

where $V_D(t)$ is the expected total reward that the G-network will earn in time t if it starts in state $\mathbf{x}$, $\mathbf{x} \in D$.

Equation (3) serves as a mathematical model of the expected total reward that the cloud computing system will earn in time t, provided that the start state of the system belongs to the set D. With the initial condition $V_D(0)$ specified, the linear differential equation (3) completely defines $V_D(t)$.

References

1. Howard R.A. (1960). *Dynamic programming and Markov processes*. Cambridge: Technology Press of Massachusetts Institute of Technology.

2. Nazarov A.A., Moiseeva S.P. (2006). *Method of asymptotic analysis in theory of queuing systems*. Tomsk: NTL.

3. Rusilko T.V. (2023). The G-network as a stochastic data network model. *Journal of the Belarusian State University. Mathematics and Informatics*. Num. **2**, pp. 45–54.

4. Rusilko T.V., Salnikov D.A. (2024). Asymptotic analysis of a closed G-network with rewards. *Tomsk State University Journal of Control and Computer Science*. Num. **68**, pp. 38–47.

5. Tikhonov V.A., Mironov M.A. (1977). *Markov processes*. Moscow: Sovetskoe radio.

THE LIMIT JOINT DISTRIBUTIONS OF STATISTICS OF THE NIST TESTS AND THEIR GENERALIZATIONS

M.P. Savelov[1]

[1]*Lomonosov Moscow State University*
Moscow, RUSSIA
e-mail: [1]savelovmp@gmail.com

Consider the problem of testing the hypothesis H_0 against the alternative H_1, where the hypothesis H_0 is that the tested sequence consists of independent random variables with a given polynomial distribution, and the alternative hypothesis H_1 corresponds to a scheme of trials in which the distribution of the tested sequence approaches its distribution under H_0. To solve this problem, we consider four types of statistics, that are generalizations of statistics of tests of the NIST package and other packages. We present the limiting joint distribution of statistics of these 4 types and the estimate of the rate of convergence to this limiting distribution (analogue of Berry-Esseen inequality). We also present necessary and sufficient conditions for asymptotic independence of the statistics under consideration.

Keywords: joint distribution of statistical tests, NIST STS, asymptotically independent statistics, Berry-Esseen-type estimates, limit distributions

1 Introduction

One of the most famous tools used to test random and pseudorandom number generators is the NIST statistical test suite [1]. In recent years, many papers have been published on this topic (e.g., [2]– [6]).

In this paper we consider four types of statistics, that are generalizations of statistics of some tests of the NIST package and other packages.

2 Main results

Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ be random variables, taking values in the set $\{0, 1, \ldots, R-1\}$. The hypothesis H_0 is that the tested sequence consists of independent random variables with a known polynomial distribution, and the alternative hypothesis H_1 corresponds to a scheme of runs in which the distribution of the tested sequence converges to its distribution under H_0.

Let N_{lb} and L_{sb} be natural numbers (they are the parameters by which the statistics T_{lb} and T_{sb} will be constructed). We consider the case when N_{lb} and L_{sb} are fixed and n tends to infinity. Put

$$L_{lb} = \left[\frac{n}{N_{lb}}\right], \quad N_{sb} = \left[\frac{n}{L_{sb}}\right].$$

If the last $n - N_{lb}L_{lb}$ elements are discarded from the sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, then the remaining elements may be divided into N_{lb} "long" non-intersecting blocks of length L_{lb}: the first block is $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{L_{lb}})$, the second one is $(\varepsilon_{L_{lb}+1}, \varepsilon_{L_{lb}+2}, \ldots, \varepsilon_{2L_{lb}})$, etc. Similarly, we can discard the last $n - L_{sb}N_{sb}$ elements and split the remaining ones into N_{sb} "short" blocks of length L_{sb}.

Next, fix natural numbers m_{sum}, m_{lb}, K_{sb} and three functions $f_{sum} : \{0, 1, \ldots, R-1\}^{m_{sum}} \to \mathbb{R}$, $f_{lb} : \{0, 1, \ldots, R-1\}^{m_{lb}} \to \mathbb{R}$ and $f_{sb} : \{0, 1, \ldots, R-1\}^{L_{sb}} \to \mathbb{R}$. Divide the set of values of the function f_{sb} into $K_{sb} + 1$ non-empty disjoint subsets: $f_{sb}(\{0, 1, \ldots, R-1\}^{L_{sb}}) = \bigsqcup_{i=0}^{K_{sb}} \alpha_{sb}(i)$. We will use the following notation. The quantity $\mathbf{E}_{H_0} f_{sum}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{sum}})$ is the expectation of $f_{sum}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{sum}})$, calculated under the assumption that the hypothesis H_0 is true. The variance $\mathbf{D}_{H_0} f_{sum}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{sum}})$, etc., are interpreted similarly. We will be interested in the case when the numbers $R, p_0, \ldots, p_{R-1}, m_{sum}, m_{lb}, N_{lb}, L_{sb}, K_{sb}$, the functions f_{sum}, f_{lb}, f_{sb} and the sets $\alpha_{sb}(0), \ldots, \alpha_{sb}(K_{sb})$ are fixed, do not depend on n, and n is a changing parameter.

Put

$$E_{sum} = \mathbf{E}_{H_0} f_{sum}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{sum}}),$$
$$\sigma_{sum}^2 = \mathbf{D}_{H_0} f_{sum}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{sum}}) +$$
$$+2 \sum_{i=2}^{m_{sum}} \operatorname{cov}_{H_0}\big(f_{sum}(\varepsilon_i, \varepsilon_{i+1}, \ldots, \varepsilon_{i+m_{sum}-1}), f_{sum}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{sum}})\big),$$

where it's supposed that $\sigma_{sum} \geq 0$.

Definition 1. If $\sigma_{sum} > 0$, then a statistic of the form

$$T_{sum} = \frac{\sum_{i=1}^{n-m_{sum}+1} \big(f_{sum}(\varepsilon_i, \varepsilon_{i+1}, \ldots, \varepsilon_{i+m_{sum}-1}) - E_{sum}\big)}{\sigma_{sum}\sqrt{n - m_{sum} + 1}} \tag{1}$$

is called a **summing** statistic.

Note that in [4], under the assumption that the tested sequence is binary, statistics similar to summary statistics are considered.

Put

$$W_k = \sum_{j=L_{lb}(k-1)+1}^{L_{lb}k-m_{lb}+1} f_{lb}(\varepsilon_j, \varepsilon_{j+1}, \ldots, \varepsilon_{j+m_{lb}-1}), \quad 1 \leq k \leq N_{lb},$$
$$E_{lb} = \mathbf{E}_{H_0} f_{lb}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{lb}}),$$
$$\sigma_{lb}^2 = \mathbf{D}_{H_0} f_{lb}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{lb}}) + 2 \sum_{i=2}^{m_{lb}} \operatorname{cov}_{H_0}\big(f_{lb}(\varepsilon_i, \varepsilon_{i+1}, \ldots, \varepsilon_{i+m_{lb}-1}), f_{lb}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_{lb}})\big),$$

where it is assumed that $\sigma_{lb} \geq 0$. Note that $\mathbf{E}_{H_0} W_k = (L_{lb} - m_{lb} + 1)E_{lb}$, $1 \leq k \leq N_{lb}$.

Definition 2. If $\sigma_{lb} > 0$, then a statistic of the form

$$T_{lb} = \frac{\sum_{k=1}^{N_{lb}} \big(W_k - (L_{lb} - m_{lb} + 1)E_{lb}\big)^2}{L_{lb}\sigma_{lb}^2}$$

is called a **long-block** statistic with N_{lb} blocks.

For $0 \le j \le K_{sb}$ put

$$w(j) = \sum_{i=1}^{N_{sb}} I_{f_{sb}(\varepsilon_{L_{sb}(i-1)+1}, \varepsilon_{L_{sb}(i-1)+2}, \ldots, \varepsilon_{L_{sb}i}) \in \alpha_{sb}(j)},$$
$$E_{sb}(j) = \mathbf{P}_{H_0}\big(f_{sb}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{L_{sb}}) \in \alpha_{sb}(j)\big).$$

Note that $\mathbf{E}_{H_0} w(j) = N_{sb} E_{sb}(j),\ 0 \le j \le K_{sb}$.

Definition 3. If $E_{sb}(j) > 0$ for all $0 \le j \le K_{sb}$, then a statistic of the form

$$T_{sb} = \sum_{j=0}^{K_{sb}} \frac{(w(j) - N_{sb}E_{sb}(j))^2}{N_{sb}E_{sb}(j)}$$

is called a **short-block** statistic with blocks of length L_{sb}.

Next, for each $1 \le q \le Q_{quad}$ fix a natural number $m_{sum}^{[q]}$ and a function $f_{sum}^{[q]} : \{0, 1, \ldots, R-1\}^{m_{sum}^{[q]}} \to \mathbb{R}$ and put

$$T_{sum}^{[q]} = \frac{\sum_{i=1}^{n-m_{sum}^{[q]}+1} \left(f_{sum}^{[q]}\big(\varepsilon_i, \varepsilon_{i+1}, \ldots, \varepsilon_{i+m_{sum}^{[q]}-1}\big) - E_{sum}^{[q]} \right)}{\sigma_{sum}^{[q]} \sqrt{n - m_{sum}^{[q]} + 1}}, \tag{2}$$

where the quantities $E_{sum}^{[q]}$ and $\sigma_{sum}^{[q]}$ are defined in the same way as the quantities E_{sum} and σ_{sum} in (1). The statistics $T_{sum}^{[1]}, \ldots, T_{sum}^{[Q_{quad}]}$ are summming statistics.

Definition 4. Consider a positive integer τ_{quad} and real numbers $d_{quad}(i, q)$ $(1 \le i \le \tau_{quad}, 1 \le q \le Q_{quad})$. A statistic of the form

$$T_{quad} = \sum_{i=1}^{\tau_{quad}} \left(\sum_{q=1}^{Q_{quad}} d_{quad}(i, q) T_{sum}^{[q]} \right)^2 \tag{3}$$

is called a **quadratic** statistic constructed from the statistics $T_{sum}^{[1]}, \ldots, T_{sum}^{[Q_{quad}]}$.

The relationship between the summing, long-block, short-block statistics, quadratic statistics and the statistics from NIST STS [1] and TestU01 was discussed in [5, 6]. In particular, the following was proved with respect to the NIST STS in [5]. Let $R = 2$ and $p_0 = p_1 = \frac{1}{2}$. The statistics T_{fr}, T_{templ} of "Frequency Test within a Block" and "Non-overlapping Template Matching Test" are long-block statistics, the statistics $T_{longrun}, T_{matrix}, T_{lincompl}$ of "Test for the Longest Run of Ones in a Block", "Binary Matrix Rank Test", "Linear Complexity Test" are short-block statistics (taking into account the caveat from Example 7 [5]), the statistics T_{mon} of "Monobit Test" is a summing one, the statistics T_{runs} of "Runs Test" coincides (under certain additional conditions) with the summing statistics with an accuracy of up to $o_P(1)$, the statistics $T_{serial1}, T_{serial2}$ of the "Serial Test" and the statistic T_{entr} of "Approximate Entropy

Test" are such that each of them in a wide class of cases coincides up to $o_P(1)$ with some quadratic statistic.

The report will present the limiting joint distribution of statistics of 4 types (summing, long-block, short-block and quadratic) and necessary and sufficient conditions for asymptotic independence of these statistics. In addition, for summing, long-block and short-block statistics we obtain an analogue of the Berry-Esseen inequality. The report consists of results from [5, 6] and new results.

References

1. Rukhin A., Soto J., Nechvatal J., Smid M., Barker E., Leigh S., Levenson M., Vangel M., Banks D., Heckert A., Dray J., Vo S. (2010). A statistical test suite for the validation of random number generators and pseudo random number generators for cryptographic applications. NIST Special Publication 800-22 Revision 1a, ed. L. E. Bassham III, NIST.

2. Voloshko V. A., Kharin Yu. S., Trubey A. I. (2022). On power comparison for some tests on pure randomness under Markov high-order dependencies. *Computer Data Analysis and Modeling: Stochastics and Data Science: Proc. of the XIII Intern. Conf.*, pp. 211-217.

3. Zubkov A. M., Serov A. A. (2023). Experimental study of NIST Statistical Test Suite ability to detect long repetitions in binary sequences. *Mat. Vopr. Kriptogr.* Vol. **14**, Num. **2**, pp. 137-145.

4. Voloshko V. A. (2023). On the asymptotic properties of the family of χ^2-tests of pure randomness of a binary sequence. *Theoretical and Applied Cryptography, Proc. of the II Intern. Sci. Conf.*, pp. 15-43.

5. Savelov M. P. (2024). The limit joint distributions of statistics of tests of the NIST package and their generalizations. *Diskr. Mat.* Vol. **36**, Num. **2**, pp. 71-116.

6. Savelov M. P. (2025). Asymptotic independence of statistics of tests of the NIST package and their generalizations. *Diskr. Mat.* Vol. **37**, Num. **1**, pp. 76-111.

ON WEB-TEXTS CLASSIFICATION WITH METHODS OF COMPUTER DATA ANALYSIS

V.S. Selezneva[1]
[1]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]kbKz222@yandex.ru

The relevance of this research is due to the analysis of web-texts by means of automatic sentiment analysis. The Apify software product was chosen for the research and reviews web-texts were used as the studied material. In the course of the research the positive, neutral and negative opinions were revealed, which in the future will allow users to identify the best product, and the owners of the product to increase their credibility.

Keywords: electronic text (e-text), web-text, sentiment analysis, reviews classification, Apify software

1 Introduction

With the widespread use of computer technologies and melting of the Internet with the everyday real life, a new form of existence and interpretation of text in digital media space has appeared – electronic text (hereinafter –*e-text*) that acquires new features and characteristics, presented in new or modified genres and formats.

Nowadays for computer data analysis it is quite natural to consider e-text as a certain amount of data subject to computer processing. "Text in digital form is essentially data to be processed" [1], p. 124. E-text is considered as digital material encoded as a binary alphabet. "Electronic text can be defined by taking the point of departure in the digital format in which everything is represented in the binary alphabet" [2], p. 1. Another important feature of e-text is the presence of hyperlinks, which are in any kind of electronic text. As "hypertext can be defined as a coded relation between anchor, link and destination" [2], p. 4, any coded link and connected nodes are identified as an essential part of the e-text.

For this study we detail that one of the varieties of e-text is web-texts, which are typical for online media formed on the Internet, namely: online news papers and magazines, web sites of news agencies, information portals, news feeds, blogs, etc. User comments are also can be interpreted as network texts. Based on the above, it should be supposed that classical methods of text analysis will not be relevant for the evaluation of text data in the digital environment, because firstly the data are characterized by large volumes and complex structure (the presence of hyperlinks), requiring time for evaluation. Therefore, new approaches are appearing for analyzing e-text through computer technology, involving automation of the processing.

2 Results of data analysis for reviews sentiment classification

Manual text tagging nowadays is being replaced by automatic evaluation of sentiment analysis. Automation is performed with the help of machine learning algorithms [3]. A training corpus of texts with tagged sentiment is selected in advance, and then a model for classifying texts by sentiment is tested. The electronic texts are tagged with affective labels (*A-labels*) for positive, neutral and negative sentiment.

Positive sentiment refers to the expression of positive emotions such as joy, pleasure, and approval. An example of positive sentiment is a review of a product or service in which the user expresses his or her positive evaluation.

Neutral sentiment refers to the absence of obvious emotional evaluations. For example, informational news or documentary content is often characterized by neutral sentiment.

Negative sentiment is associated with the expression of negative emotions, including anger, sadness, and disappointment. A negative review of a restaurant or service may be an example of negative sentiment.

Currently, there is an increased interest in the possibilities of sentiment analysis on the part of commercial projects that use this technology to study public opinion about their products and services, namely user comments.

In this study, the Apify software was chosen to analyze web-texts. Sentiment Analysis Online Tool can analyze the sentiment of any e-text you provide. This tool can classify positive, neutral or negative sentiment of the text and offers a confidence score to indicate the certain classification. The software interface is designed in a convenient way. Using this sentiment analysis tool work the user need to input e-text in the appropriated space. E-text has to be not longer than approximately 250 characters (if the text is longer, it will be trimmed), the software processes English e-text. An artificial intelligence model processes it and gives classification of positive, neutral, or negative sentiment. There is a confidence score for the presented classification below the analyzed e-text.

This study examines examples of web-texts with positive, neutral and negative sentiments to widely demonstrate the performance of the software product. The texts are comments made by web site users about traveling experience, based on which other users can find most appropriate options for traveling to different places. The results of the research in the software product are presented below:

"index": 0,

"inputText": "We had an excellent day out at Milford as the night before we visited it had rained & there were literally scores of waterfalls all through the fiord ... very magical.",

"finalClassification": "positive",

"finalScore": 0.9839617013931274,

"negativeScore": 0.0030287420377135277,

"neutralScore": 0.013009591959416866,

"positiveScore": 0.9839617013931274.

With the finalClassification score of 0.984, this comment has a positive sentiment.

The neutral indicator was 0.013 and the negative indicator was 0.003.

Next, the following comment is considered:

"index": 0,
"inputText": "NZ$ 225 per pax includes coach return transfer from Queenstown. 100 min cruise on triple deck ship. Complimentary coffee/tea/hot water. Organised by Real Journeys.",
"finalClassification": "neutral",
"finalScore": 0.7315853238105774,
"negativeScore": 0.005336072761565447,
"neutralScore": 0.7315853238105774,
"positiveScore": 0.2630786597728729.

Focusing on finalClassification score of 0.732 we conclude that the text has a neutral sentiment and has only informative character for users. The positive and negative indicator scores are respectively 0.263 and 0.005.

Consider the following comment with a negative sentiment below:

"index": 0,
"inputText": "Viator went to the wrong hotel to pick me up and then told me to take a taxi to outrun the coach to try to take the boat. What's worse they charged me for it",
"finalClassification": "negative",
"finalScore": 0.9004036784172058,
"negativeScore": 0.9004036784172058,
"neutralScore": 0.09445048123598099,
"positiveScore": 0.005145765375345945.

This comment based on the finalClassification 0.900 has a negative sentiment. The negative connotation prevails over the neutral indicator score of 0.095 and the positive score of 0.005.

3 Conclusion

Based on the results of Apify software product, comments with positive, neutral and negative sentiment were found. This methodology of sentiment analysis allows to get information about emotions and opinions of people on certain topics on the basis of their comments in social networks, helps to identify problems and defects of the product and highlight areas for its improvement, as well as to conduct competitive analysis for the formation of business strategy.

References

1. Buzzetti D. (2018) Digital text representation expression and content. *FORD, D. Contexts: Proceedings of ANPA*. Vol. **31**, pp. 124-145.

2. Finnemann N. (2018) E-text. *The Oxford Research Encyclopedia of Literature. Ed. by P. Rabinowitz*. New York: Oxford University Press, pp. 1-47. (DOI: 10.1093/acrefore/9780190201098.013.272)

3. Selezneva V.S. (2023) Expert system as a toolkit for metalinguistic communication. *Russian Linguistic Bulletin*. No. 8 (44): 14. (DOI: 10.18454/RULB.2023.44.14)

RELIABILITY OF TWO-LEVEL TESTING APPROACH OF THE NIST TEST SUITE

A.A. Serov[1]

[1]*Steklov Mathematical Institute of Russian Academy of Sciences*
Moscow, RUSSIA
e-mail: [1]serov@mi-ras.ru

The two-level approach for testing RNGs involving the well known NIST SP 800-22 test suite, i.e., counting the sequences passing a basic test and checking the p-values distribution with a chi-square test, was considered. It is shown that for AES-based sequences two-level testing approach is not reliable. For a reliable second-level test, systematic error in the computing of the p-values should be smaller, or at least, approximately equal to $\sigma/N = \frac{1}{k}\sqrt{\frac{k-1}{N}}$, where $\sigma = \sqrt{\frac{1}{k}\left(1-\frac{1}{k}\right)N}$. Such heuristic assumptions and carried out experiments suggest that for example in the second-level test of the Frequency test of NIST SP 800-22 test suite with $n = 2^{20}$ the number of tested sequences N should not exceed 26184.

Keywords: random sequences, pseudorandom sequences, statistical testing, reliability of statistical test, two-sided estimates

1 Introduction

Random sequences are used in a large variety of areas, such as quantum mechanics, game theory, statistics, cryptography and so on. Random Number Generators (RNGs) represent a fundamental component in many applications, they are essential for cryptographic systems (see, for example, [2]). For any type of RNG statistical hypothesis tests have been widely employed to assess the quality of the RNG, which evaluate whether the output sequences conform with the given null hypothesis $\mathcal{H}_0$ (e. g., the elements of the sequence independent and uniformly distributed) or not.

The quality checking of binary sequences usually is based on some well-known batteries of tests, each of which is composed of a serial of tests, include Diehard [1] proposed by Marsaglia, SP 800-22 [5] standardized by US National Institute of Standard and Technology (NIST) or a software library TestU01 [6].

From a mathematical point of view a test may be considered as a function of a sequence of n elements (e.g., a sequence of n bits) with output value in $[0, 1]$, called a p-value. In null-hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.

2 NIST SP 800-22 test suite

The most commonly used statistical test suite, the SP 800-22 test suite from US National Institute of Standard and Technology (NIST) [5], is considered. This statistical

test suite is build for analyzing the randomness properties of sequences and generators, is composed of 15 tests. The purpose of the research is to consider the testing strategy proposed in Section 4 of the NIST publication [5] and discuss under which assumptions this strategy increases the reliability and when, on the other hand, produces incorrect results, i. e. the empirical significance level α does not correspond to the theoretical one. In that context, a *reliable test* should be understood as a test such that the probability of a false positive (Type I error) is agreed with the expected one.

3 Second-level testing approach of the NIST SP800-22

In [5] NIST recommends using the second-level testing approach (it was found to increase the testing capability [3]); a long binary sequence is partitioned into N subsequences, each with n bits. A standard test is applied for each sequence, and the distribution of the N obtained p-values is compared with a uniform distribution $F_p^{(0)}$. To check this NIST proposes a chi-square goodness-of-fit test, this test is again a statistical test and gives another (a second-level) p-value p_{II}.

In every statistical test some approximations are adopted, introducing errors in the p-value computation and so in the p-value distribution. It was observed [3, 4] that for extremely large values of N, the level-two approach always fail, it ends with $p_{II} \simeq 0$. In this case we can say that the test is *not reliable*.

4 Experiments

All 15 tests from the SP 800-22 test suite were applied to pseudorandom sequences generated by AES block cipher. The detailed description of the design of AES-based RNG may be found in [7, 8]. The two-level test was performed for all 188 statistics values computed by NIST test suite, but only 15 of them are presented in the Table 1. These results confirm that for extreme values of N (e.g., $N = 2^{20}$) the two-level testing approach was carried out on the sequences obtained by AES algorithm (it is known that such sequences are practically indistinguishable from random one, see [7, 8]) are failed too, i.e. it ends with $p_{II} \simeq 0$.

In order to identify the problem, let's take a look on the simple Frequency test.

5 Frequency (Monobit) Test

The purpose of the Frequency test is to determine whether the number of ones and zeros in a sequence are approximately the same as would be expected for a truly random sequence.

Assuming the bound ϵ of the error in the computation of a p-value from Berry and Esseen theorem, we can bound also the maximal error Δ in the number of N p-values

Table 1: Results of the χ^2-based two-level randomness test by N sequences for the AES-based RNG (small P-values $p_{II} \leq 0.01$ are in bold)

#	**Test Name**	$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 2^{20}$
1	*Frequency*	0.616305	0.290806	0.588411	**0.000803**
2	*Block Frequency*	0.187581	0.773212	0.374097	**0.000125**
3	*Cumulative Sums*	0.401199	0.124765	0.959543	**0.000009**
4	*Runs*	0.150340	0.885418	0.910568	0.107966
5	*Longest Run*	0.610070	0.239883	**0.000355**	**0.000000**
6	*Binary Matrix Rank*	0.878618	0.341017	**0.000000**	**0.000000**
7	*Discrete Fourier Transform*	0.371941	0.014836	**0.000000**	**0.000000**
8	*Overlapping Templ. Match.*	0.071177	0.202268	**0.000000**	**0.000000**
9	*Universal statistical test*	0.574903	0.108534	**0.000000**	**0.000000**
10	*Approximate Entropy*	0.246750	0.078038	0.219501	**0.000000**
11	*Serial*	0.942198	0.174057	0.213964	0.572679
12	*Linear Complexity*	0.839507	0.279152	0.299852	0.117305
13	*Non-overlap. Templ. Match.*	0.092041	0.372782	0.121382	0.275416
14	*Random Excursion*	0.914727	0.663838	0.346173	**0.000028**
15	*Random Excursion Variant*	0.238264	0.133576	**0.000080**	**0.000000**

in k sub-interval and obtain a very simple reliability condition for Frequency Test

$$\Delta < \sqrt{N(k-1)}/k\,.$$

In the case $\Delta = 2N\epsilon$, $\epsilon = 9,3 \cdot 10^{-4}$, $k = 10$ and $n = 2^{20}$, we get

$$N \leq \frac{1}{4\epsilon^2 k}\left(1 - \frac{1}{k}\right) \simeq 26163.7.$$

6 Conclusion

The two-level approach for testing RNGs involving the well known NIST SP 800-22 test suite was considered. Such approach may increase the reliability of the test. However it is sensitive to the approximation error introduced by the computing of p-values. Systematic error in the computing of the p-values is dependent only on the accuracy of approximation of the exact distribution of statistic by its theoretical counterpart and the number of bits in the analyzed sequences n.

References

1. Marsaglia G. (1996). DIEHARD: a battery of tests of randomness. `http://stat.fsu.edu/ geo/diehard.html`

2. Meneze A.J., van Oorschot P.C., Vanstone S.A. (1996). *Handbook of Applied Cryptography*. CRC Press.

3. Pareschi F., Rovatti R., Setti G. (2007). Second-level NIST randomness test for improving test reliability. *ISCAS 2007*. New Orleans (USA), May 27–30, pp. 1437-1440.

4. Pareschi F., Rovatti R., Setti G. (2012). On statistical tests for randomness included in the NIST SP800-22 test suite and based on the binomial distribution. *IEEE Trans. Inf. For. Sec.* Vol. **7**, Num. **2**, pp. 491-505.

5. Rukhin A., Soto J., Nechvatal J., Smid M., Barker E., Leigh S., Levenson M., Vangel M., Banks D., Heckert A., Dray J., Vo S. (2010). *A statistical test suite for the validation of random number generators and pseudo random number generators for cryptographic applications*. NIST Special Publication 800-22 Revision 1a.

6. LEcuyer P., Simard R. (2013). *TestU01*. Dept. dInform. Rech. Oper. Univ. Montreal. P. 214. `http://simul.iro.umontreal.ca/testu01/guideshorttestu01.pdf`

7. Zubkov A.M., Serov A.A. (2019). Testing the NIST Statistical Test Suite on artificial pseudorandom sequences. *Matematicheskie Voprosy Kriptografii*. Vol. **10**, Num. **2**, pp. 89-96.

8. Zubkov A.M., Serov A.A. (2021). A natural approach to the experimental study of dependence between statistical tests. *Matematicheskie Voprosy Kriptografii*. Vol. **12**, Num. **1**, pp. 131-142.

DETECTING SAMPLE RATIO MISMATCH WITH SEQUENTIAL TESTING

M.A. Shevtsova[1], V.V. Kharlamov[2], G.V. Zasko[3]

[1,2,3] *T-Bank, Applied Statistics Laboratory*
Moscow, RUSSIA
e-mail: [1]shevtsova1ma@gmail.com, [2]vi.v.kharlamov@gmail.com,
[3]zasko.gr@bk.ru

Online controlled experiments, or A/B tests, are the most reliable method for evaluating the impact of product changes and making data-driven business decisions. In A/B tests, unintended deviations from the designed group allocation ratio can occur. This phenomenon is called a sample ratio mismatch (SRM). The presence of SRM indicates an issue with the experiment and suggests that the results may be biased. Early detection of SRM is crucial because it allows biased experiments to be stopped quickly. In this report, we study the sequential methods for detecting SRM both theoretically and numerically. Specifically, we focus on group sequential methods based on the Pearson chi-squared statistic, as well as sequential methods with a Bayesian alternative. We compare these methods through numerical experiments, using both real and synthetic data.

Keywords: A/B-testing, sample ratio mismatch, sequential testing

1 Motivation

Online controlled experiments, or A/B tests, play a key role in data-driven decision-making in industrial applications [2]. One of the fundamental assumptions in A/B testing is the randomized allocation of units to variations according to a given ratio. When this assumption does not hold, the observed and planned ratios may differ, with the difference being statistically significant. This phenomenon is called a sample ratio mismatch (SRM) [1]. In online experiments, SRM usually happens due to network effects, and issues with the randomization process and data filtering.

Note that SRM acts as an indicator that the experiment is biased, which means that valid statistical inferences cannot be drawn from the results. Therefore, early detection of SRM is crucial to prevent wrong decisions in the future and minimize resource losses.

2 Methods

Consider a sequence of independent observations, each belonging to one of the $d \geq 2$ variations. Suppose the probability of an observation falling into the j-th variation is $\theta_j > 0$, and the vector of true probabilities is $\theta = (\theta_1, \ldots, \theta_d)$, and θ is the same for all observations. For $t \geq 1$ and $j \in [1, d]$, define the indicator variable

$$I_j(t) = \begin{cases} 1, & \text{if the } t\text{-th observation belongs to the } j\text{-th variation ,} \\ 0, & \text{otherwise.} \end{cases}$$

Let $\theta^0 = (\theta_1^0, \ldots, \theta_d^0)$ be the target allocation ratio specified in the experimental design. The problem of detecting sample ratio mismatch (SRM) can be formulated as a hypothesis testing problem:

$$H_0 : \theta = \theta^0.$$

Let $X(n)$ be the Pearson chi-squared statistic computed from the first n observations:

$$X(n) = \sum_{j=1}^{d} \frac{1}{n\theta_j^0} \left(\sum_{t=1}^{n} I_j(t) - n\theta_j^0 \right)^2.$$

To test H_0 against the alternative $H_1 : \theta \neq \theta^0$, one may apply the chi-squared goodness-of-fit test after the experiment is finished. This approach controls the Type I error rate and typically has high power, but does not allow SRM to be detected before the end of the experiment. In industrial applications, the detection of SRM is desirable as soon as possible. Therefore, we focus on sequential testing procedures, which also control the Type I error rate but allow to stop the experiment when a sufficient evidence against H_0 accumulates.

Among sequential testing methods, we focus on two approaches. The first approach groups observations by day, which aligns with how data typically accumulate in many A/B testing scenarios. This approach is called a group sequential testing.

Let n_k be the total number of observations collected during the first k days, and let $X(n_k)$ be the corresponding Pearson statistic. Chi-squared tests apply to $X(n_k)$ with a decreasing significance level α_k, controlling Type I errors by ensuring $\sum_{k=1}^{\infty} \alpha_k \leq \alpha$. Since the number of experiment days is not known in advance, this approach typically leads to conservative error control.

Note that the total number of days in the experiment is not known in advance, so this method typically provides a conservative control of the overall Type I error. Alternatively, a more powerful method can be implemented based on the limit theorem proved in [4]. Let n be the total number of observations. As $n \to \infty$, the sequence of stochastic processes $X(\lfloor nt \rfloor)$ converges in finite-dimensional distributions to the process $\mathrm{Bes}_{d-1}^2(t)/t$, where $\mathrm{Bes}_{d-1}(t)$ is the $(d-1)$-dimensional Bessel process. Each day, the finite-dimensional distribution of the Bessel process is used to estimate the distribution of $X(n_k)$ under condition that the Pearson statistic assumes specific values on preceding days.

The second approach follows the framework introduced in [3], where the hypothesis testing problem is formulated as

$$H_0 : \theta = \theta^0, \qquad H_1 : \theta \sim \mathrm{Dirichlet}(\beta),$$

where $\beta \in \mathbb{R}_+^d$. This formulation is called a Bayesian alternative. The method proposed in [3] is an iterative algorithm that updates the posterior odds $O(n)$ of H_1 over H_0 as observations accumulate, where n is the number of observations. Then the value of $O(n)$ is compared to a given threshold. A key result enabling the theoretical justification of the method is that, under H_0, $O(n)$ is a nonnegative supermartingale. This property enables one to prove the following inequality under H_0

$$\mathsf{P}(\exists n_* \in \mathbb{N} : O(n_*) \geq 1/\alpha) \leq \alpha$$

where α is a Type I error rate. While the method is straightforward to implement, it is easy to see that it overestimates the Type I error rate by design.

3 Conclusion

In this study, we provide a theoretical comparison of the two approaches described above and evaluate their empirical performance on both real-world and synthetic datasets under various target allocation ratios and alternatives. We further discuss the advantages and limitations of each method and draw conclusions regarding their applicability in the context of online controlled experiments.

References

1. Fabijan A., Dmitriev P., Holmstrm Olsson H., Bosch J., Vermeer L. , Lewis D. (2019) Three Key Checklists and Remedies for Trustworthy Analysis of Online Controlled Experiments at Scale. *IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. pp. 1-10.

2. Kohavi R., Tang D., Xu Y. (2020). *Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press.

3. Lindon M., Malek A. (2022). Anytime-valid inference for multinomial count data. *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*. Num. **204**, pp. 2817-2831.

4. Zubkov A.M., Savelov M.P. (2016). Convergence of the Sequence of Pearson Statistic Values. *Discrete Mathematics*. Vol. **28**, Num. **3**, pp. 49-57.

EXPERT KNOWLEDGE PECULIARITIES OF "FUZZY ASSESSMENTS" AND "FUZZY MEASUREMENTS" FOR MODELING THE STATE OF COMPLEX AGRICULTURAL OBJECTS

A.V. Spesivtsev[1], I.T. Kimyaev [2], V.A. Spesivtsev [3], A.V. Inyutin[4]

[1,3] *Saint Petersburg Institute for Informatics and Automation of Russian Academy of Sciences Saint Petersburg, RUSSIA*
[2] *LLC IC "Sibintek" Moscow, Saint Petersburg, RUSSIA*
[4] *United Institute of Informatics Problems of National Academy of Sciences of Belarus Minsk, BELARUS*
e-mail: [1]sav2050@gmail.com, [2]igor95a@mail.ru, [3]ryukuro@yandex.ru, [4]avin@newman.bas-net.by

In the context of digitalization of agriculture, new approaches to scientific research are needed in constructing analytical expressions as the most conducive to computerization of a wide range of specific applied problems. Significant uncertainty of agricultural information forces decisions to be made mainly by humans. This situation leads to the use of expert knowledge as a material for constructing mathematical models. For this purpose, the most convenient is the fuzzy-possibility approach, which is capable of formalizing verbal expert information with an analytical expression. In this paper, it is used to construct a fuzzy-possibility model of the state of environmental safety of a cattle farm. At each stage of model construction, the concepts of "fuzzy estimates", "fuzzy measurements" and NON-factors were identified and interpreted as inevitable attributes of the study. The resulting adequate model allowed us to conclude that the degree of nitrogen preservation during disposal and use is a universal quantitative indicator of the farm's environmental sustainability.

Keywords: Expert knowledge, Fuzzy-possibility approach, Environmental safety of cattle farm, Fuzzy measurements and fuzzy evaluation, Non-factors

1 Introduction

An analysis of domestic and foreign scientific and technical literature on agricultural topics [1-13] showed that recently, methods of mathematical modeling of both individual local phenomena (for example, increasing milk yields) and entire technologies (FPM technologies for the production of feed from grasses) have been increasingly used. This is due to the requirements of digitalization as one of the concepts of Industry 4.0, which involves the transition to automated agricultural production controlled by intelligent systems in real time.

Each time, when approaching a new specific task, the manager (researcher, decision

maker, expert) is always in a situation of uncertainty: what method to use to solve it, whether there are measuring instruments, whether the team has enough experience to solve it on time and with a certain accuracy, and a host of other management difficulties. An important feature here is the talent for identifying and overcoming many other uncertainties: unknown, incomplete, unreliable, imprecise, underdetermination, incorrectness, consistency, etc. Thanks to the brilliant research of our compatriot A.S. Narinyani [14,15], such nouns with the prefix "non" are united under the common name of NON-factors, which are fundamental components of informatics and play a decisive role in increasing the efficiency of modeling processes.

In agriculture, all technological processes and even their individual stages should be considered from a mathematical point of view as complex objects (CO) [16]. This approach allows the use of a powerful developed and tested apparatus of the model-algorithmic approach to solving practical problems of agricultural production [17].

Figure 1 schematically shows the main fuzzy linguistic variables of the multifactor space, systematically characterizing the ecological state of the cattle farm (CF), allowing the synthesis of multifactor mathematical models on their basis [18]. When analyzing Figure 1, uncertainty is manifested by several NON-factors: underdetermination of the factor space (many different factor spaces can be constructed), the presence of NON-quantitative (qualitative, verbal) fuzzy linguistic variables and their incompleteness, and possible inaccuracy in determining the boundaries of change of variables.

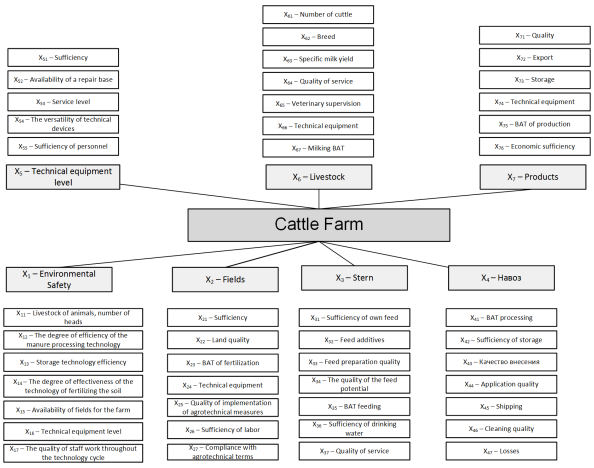


Figure 1: Factor space characterizing the environmental sustainability of a cattle farm

NON-factors are one of the features of "fuzzy computing" introduced by the creator of fuzzy mathematics Lotfi Zadeh when studying and predicting the state of the CL of any subject area [19]. The NON-factor of data inaccuracy manifests itself not only in the measurement of physical quantities, but also in errors in their recording. It should be taken into account that any measured value of any physical quantity is indistinguishable in a certain range of inaccuracy. The widespread use of implicit NON-factors in everyday and practical activities is manifested regardless of the subject area [14,15]. Thus, despite the declarative nature (in other words, vagueness, imprecision) of

the work statement by the manager, employees refract it to their front of responsibilities due to professional experience and the innate ability of a person to recognize the vagueness of an image, verbal or real. This, in fact, is what the manager's confidence in the team's overcoming of the vagueness and ambiguity of the essence of the upcoming (declared) work is based on. This study provides an opportunity to get acquainted with some important features of the use of "fuzzy measurements" and "fuzzy assessments" based on expert knowledge in the context of NON-factors.

2 Materials and methods

Currently, the principles of digitalization give rise to very special requirements for materials and methods in scientific publications on modeling, which differ significantly from those accepted in journals on agricultural topics. In this study, one of the main features is that the expert knowledge (EK) serve as the materials, and the ones used **methods** use specific methodological model-algorithmic approaches that contribute to **representation and formalization of EK by analytical expressions**. Methods for assessing the state of the CO in agriculture have the peculiarity that in practice most decisions are made by a person. This situation is also typical in other subject areas (CO), for example, even in astronautics from 50 to 70 per of decisions are made collegially. In such conditions, it is quite natural to use the fuzzy-possibility approach (FPA), based on the use of explicit and implicit EP [16]. At the same time, costruction fuzzy-possibility models (FPM) quantitative assessment of the state of the CO occurs under conditions of uncertainty, which includes imprecision, fuzzyness and other nouns with the prefix "non" with the general name of NON-factors [14,15], Method of constructing FPM includes the use of the ideology of three fundamental theories: the theory of fuzzy sets, the theory of experimental design and the theory of NON-factors. Below we consider the specific features of the application of each of the listed theories in relation to the modeling of technical and technological processes and technologies of agricultural production.

Features of "fuzzy measurements" and "fuzzy assessments" based on expert knowledge Using the example of mathematical assessment of the state of environmental sustainability of cattle farms, the main elements of the selection and theoretical justification of the methodology for constructing the FPM are shown, taking into account the ecological rehabilitation and features of the application of "fuzzy measurements" and "fuzzy assessments". The scientific and methodological apparatus of the formal description of the state of the CO based on explicit and implicit EK includes, as shown in Figure 2, three main stages: extraction (operator g_1), presentation (operator g_2) and formalization (operator g_3). General mapping of EK the operator μ in set-theoretical form can be represented as:

$$\mu = g_3{}^{\circ}g_2{}^{\circ}g_1 : T \times U \times X \to Y/\Xi, \tag{1}$$

where T is the set of moments of time t at which the object is observed; U,Y are the sets of input U and output Y effects, respectively; X is the set of states of the object, characterized at each moment of time t∈T is a set of fuzzy linguistic variables of the

factor space; Y/Ξ, is a factor set of states of the SO, to one of which the calculated value of Y according to the constructed model must be assigned.

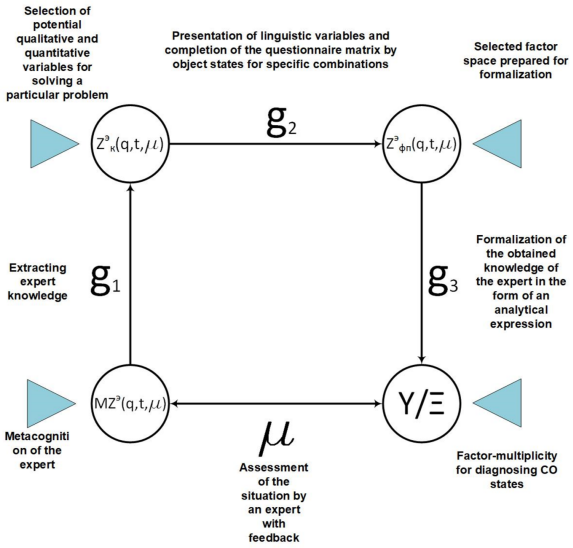


Figure 2: Commutative diagram of the processes of extraction, representation and formalization of explicit and implicit expert metaknowledge

Feature extraction of the EK (g_1, Figure 2) at the first stages of working with an expert is the need to include in the study as many variables as possible that are inherent in the functioning of the studied CO. Usually, the variables are depicted in fuzzy linguistic form (Figure 3) to enable the reflection of "fuzzy measurements", on the basis of which the expert makes a "fuzzy assessment". The abscissa axis of Figure 3 actually contains three transition scales: at the top is a verbal scale for ease of use by the expert (the expert thinks in words, not numbers); at the bottom are the quantitative values of the variable (0.3, ..., 0.7) and the third is a standardized scale for applying the methods of the theory of experimental design ("-1", ..., "+1"). The membership function is located along the ordinate axis, in which only the modes of verbal assessments correspond to the value "+1". It should be especially noted that the view shown in Figure 3 fuzzy linguistic variable is fuzzy model of the element of the EK [16-18], which translates the expert's verbal "fuzzy assessments" into numerical values as "fuzzy measurements". In this case, each mode on the verbal abscissa axis (Low, ..., Average, ..., High) represents fuzzy number, for example, the "Above Average" mode is based on the interval of the set of numbers [0.5–0.7] and makes it possible to express the measure fuzzy expert "fuzzy assessment" in the vicinity of a given mode. The same technique is especially important at the stage of determining the degree of adequacy of the constructed FPM based on statistical data, but already as a "fuzzy measurement" for each fuzzy linguistic variable.

Knowledge representation expert (g_2, Figure 2), in contrast to the representation of a knowledge element (Figure 3), consists of selecting from the entire preliminary set of such variables that, in the expert's opinion, will be included in the factor space for

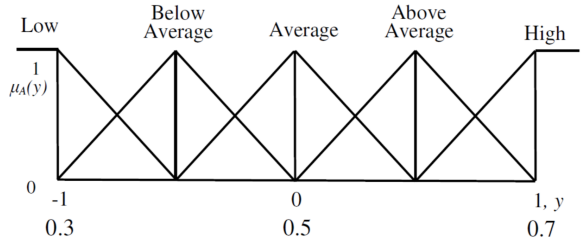


Figure 3: General view of a linguistic variable

solving a specific problem. By choosing the correct factor space, the expert tries to ensure the systematicity fuzzy linguistic variables in the description of various aspects of the functioning of the CO. The selected factor space fully reflects the knowledge and experience of the expert in solving the assigned task. The expert pays special attention to the selection and justification of the dependent variable Y, which should reflect in a generalized form the changes in the state of the CO in various situations. Thus, when studying the problem of processing and using manure, the expert chose the environmental sustainability of the cattle farm as Y. Since such an indicator does not exist in quantitative form in agricultural practice, it was decided to choose a dimensionless fuzzy oppositional scale in the interval [0.3, ..., 0.7], as shown in Figure 3. In this case, a verbal-numerical table must be developed, where the verbal characteristics for the transition to the corresponding numerical intervals are indicated, as shown in Table 1.

Table 1: Translation of verbal characteristics into numerical intervals

Intervals	Modes of intervals and their meanings	Descriptive characteristics
0.4 and below	Low (L) 0.3	An environmentally unsustainable cattle farm with a risk of causing environmental damage of more than 70%
	Below average (BA) 0.4	An environmentally unsustainable cattle farm with a risk of causing environmental damage greater than 50%
0.4 – 0.6	Average (A) 0.5	A cattle farm of medium ecological sustainability with a risk of occurrence of localized cases of damage to the environment of less than 50%
	Above average (AA) 0.6	Sustainable farm with minimal risk of local environmental damage
0.6 and above	High (H) 0.7	An environmentally sustainable farm with no significant risks of harming the environment

Even white analysis of table 1 reveals uncertainty verbal information. fuzziness descriptive characteristics for “fuzzy measurements” and numerical interval values for

"fuzzy estimates". However, even in conditions of such a significant uncertainties. The constructed FPM quite effectively represents the experience and knowledge of an expert on a given specific issue. In the present study, the factor space on which the FPM is constructed includes seven (NON-factor incompleteness) fuzzy linguistic variables [18]: X1 is number of animals on the farm, 400-2500 heads; X2 is degree of efficiency of manure processing technology; X3 is degree of efficiency of storage technology; X4 is degree of efficiency of fertilizer application technology; X5 is availability of fields for the farm; X6 is level of technical equipment; X7 is level of organization and control of technological processes. The linguistic form of the variable Y on the abscissa axis contains only two scales (Figure 3), since there is no need to translate it into a standardized scale. Here it is necessary to clarify the difference in the nature of the manifestation of "fuzzy assessments" and "fuzzy measurements". As follows from the analysis of the factor space, all fuzzy linguistic variables. When constructing NVMs, they are used to determine the "fuzzy assessments" of the corresponding factors by the expert when filling out the survey matrix, and in the process of assessing the degree of adequacy of the model, the same linguistic form of their presentation is used for "fuzzy measurements". Further, according to the methodology based on the methods of the theory of experimental design, an expert survey matrix is constructed (Table 2) as a semi-replica of the full factorial experiment 27-1, where each row of the matrix represents fuzz zy production rule of the implicative type "if, ..., then". The expert fills in each line (situation) with a verbal assessment of YEV taking into account the verbal-numerical table 1 and after translating his assessments into quantitative values YEN a polynomial model Y is constructed using a special software product [20].

Table 2: Expert survey matrix

	x1	x2	x3	x4	x5	x6	x7	YEV	YEN	YM
1	-1	-1	-1	-1	-1	-1	1	L	0.30	0.30
2	1	-1	-1	-1	-1	-1	-1	L-BA	0.35	0.35
3	-1	1	-1	-1	-1	-1	-1	BA	0.40	0.42
4	1	1	-1	-1	-1	-1	1	BA-A	0.45	0.49
...	...	...	...	...	...	...	...	...	...	...
61	-1	-1	1	1	1	1	1	AA	0.60	0.62
62	1	-1	1	1	1	1	-1	AA-H	0.65	0.64
63	-1	1	1	1	1	1	-1	AA	0.60	0.61
64	1	1	1	1	1	1	1	H	0.70	0.70

Processing of expert information according to Table 2 in quantitative terms led to the model [18]:

$$Y = 0.52734 + 0.02891x_1 + 0.05078x_2 + 0.05391x_3 + 0.03047x_4 + 0.02422x_5 + 0.02891x_6 + 0.01172x_7 - 0.02578x_2x_3 - 0.01172x_5x_6 + 0.00859x_5x_7 + 0.00703x_1x_4x_7 - 0.00859x_1x_5x_7 - 0.01328x_2x_3x_5 - 0.01016x_3x_4x_6,$$

where only terms with significant coefficients are presented, and variables are in a

standardized scale according to the formulas:

$$x_i = \frac{X_i - \overline{X_l}}{\Delta X_i}, \quad \bar{X}_i = \frac{X_{\max} + X_{\min}}{2}, \quad \Delta X_i = \frac{X_{\max} - X_{\min}}{2}, \tag{2}$$

$i = 7$ is number of variables.

3 Results and discussion

The very first feature of this study is the understanding that NON-factors have been used, are used and will be used by almost all managers, decision makers, researchers, but only implicitly. This study provides an opportunity to "feel" (reveal) the effect of as many NON-factors as possible and be able to overcome them. As follows from the above, the choice of factor space is fully manifested by the NON-factor under determination, because another the expert can choose another factor space from a set of potentially possible ones fuzzy linguistic variables. And the linguistic variables themselves, by definition and by properties, are unclear. Fuzzy is also manifested in the choice of variable names, as in our example: Y is a generalized indicator of the ecological state of a cattle farm. Officially, such an indicator does not exist, and we introduced it as a characteristic reflecting the degree of multifactorial ecological impact of a specific farm on the environment. If we talk about fuzzy information used, the main role in it is played by inaccuracy. Here the issue is in the culture of handling quantitative information. For example, fresh cattle bedding manure contains 0.40 percent total nitrogen with a 5 percent analysis error, the zone of indistinguishability is 0.38 – 0.42 percent. Then, in economic or production calculations in this zone of nitrogen values, the difference in results is statistically unprovable, it can only be assessed in fact at the end of the production cycle, for example, by crop yield or income received in rubles. The specifics of the expert's professionalism in choosing the factor space, its "fuzzy measurements" and "fuzzy assessments" are revealed only at the final stages of constructing the FPM when checking the degree of adequacy of the model calculations to the actual data, as shown in Table 3 for three farms in the Leningrad Region [18]. In the absence of statistics on such specific data, a study was conducted

Table 3: Results of assessing the degree of adequacy of calculations according to the model to the actual values of the state of environmental sustainability of cattle farms on three farms

Farms	x_1	x_2	x_3	x_4	x_5	x_6	x_7	Y_F	Y_M	Y_{MV}
1	-1	-0.5	-0.25	-0.75	-0.5	1	1	BA	0.463	L-BA
2	1	0.5	0.5	0	0.5	1	1	AA	0.647	A-AA
3	0	1	1	1	-0.5	1	1	AA-H	0.663	AA-H

using the developed methodology situationally in three farms in the Leningrad Region. Analysis of the results in Table 3 allows us to conclude that the calculations are highly similar YM according to the model (1) actual state environmental sustainability of

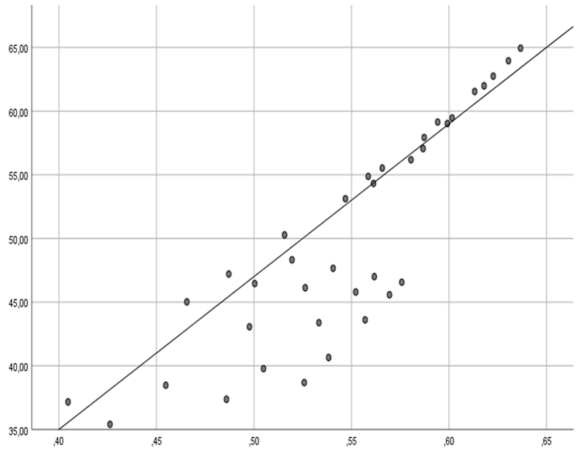


Figure 4: The relationship between nitrogen conservation and the environmental sustainability index of a cattle farm, correlation coefficient R= 0.864

cattle farms, conducted by independent experts YF, YMV - YM assessment in verbal form according to Table 1. The mathematical model (1), as the analysis shows, can be used to assess the condition of farms by this indicator and conduct more in-depth studies. Thus, Figure 4 shows the correlation between the nitrogen content after the introduction of organic fertilizers into the soil with incorporation and the environmental sustainability of a cattle farm.

As a result of the demonstrated dependence, an important conclusion can be made: the content of retained nitrogen introduced into the farm's land can serve as a universal quantitative indicator of its environmental sustainability.

4 Conclusion

The importance and peculiarity of "fuzzy assessments" and "fuzzy measurements" are demonstrated in detail using a specific example of constructing a fuzzy-possibility model for assessing the state of environmental safety of a cattle farm. Their significance lies in the fact that only such an approach to using the knowledge and experience of an expert provides the opportunity to construct analytical fuzzy-possibility models of the state of complex objects and technologies of agricultural production. At the same time, the awareness of the uncertainties of any information as an inevitable feature of scientific research in any subject area leads to the need to identify and overcome the effect of many NON-factors. In such a situation, the peculiarity of publications on agricultural topics in the "Materials and Methods" section should take into account that the material is expert knowledge with the involvement of methods for formalizing it in an analytical form. This approach, using the example of constructing a FPM for environmental safety of a cattle farm, made it possible to explain not only the use of "fuzzy assessments" and "fuzzy measurements", but also to identify the effect of specific uncertainties at various stages of the study. Thus, the conducted study showed that on the basis of fuzzy mathematics in conditions of fuzzy initial information in

a fuzzy space of fuzzy linguistic variables, it is possible to obtain clear results in the management of agricultural production.

5 Acknowledgements

The study was funded by a grant from the Russian Science Foundation No. 24-19-00823, https://rscf.ru/project/24-19-00823.

References

1. Shalavina, E.V., Uvarov, R.A., Vasiliev, E.V.2022Methodology for calculating the distribution of total nitrogen and total phosphorus between the fractions of pig manureEngineering technologies and systems32154–70

2. Bryukhanov, A.Yu. [et. al.] (2019). Methods for solving environmental problems in animal husbandry and poultry farming. *Agricultural machinery and technology*. Vol. **13**, No. **4**. P. 32–37.

3. Sukhoparov, A., Spesivtsev, A. (2021). Evaluation of the efficiency of perennial grass cultivation on the basis of a fuzzy-possibility model. *Proc. XX Int. Sc. Conf. ENGINEERING FOR RURAL DEVELOPMENT. May 26–28, 2021, Jelgava*. Vol. **19**. P. 1768–1773.

4. Mikhailov, V.V. [et. al.] (2021). Multimodel Evaluation of Phytomass Dynamics of Tundra Plant Communities Based on Satellite Images. *Izv. Atmos. Ocean. Phys.*. Vol. **57**. P. 1198–1210.

5. Medennikov, V.I., Raikov, A.N. (2020). Analysis of the experience of digital transformation in the world for Russian agriculture. Trends in the development of the Internet and the digital economy. *Proc. III All-Russian scientific and practical conference with international participation. Simferopol, 2020*. P. 57–62.

6. Ereshko, F.I., Kulba, V.V., Medennikov, V.I. (2019). End-to-end technologies in the agroindustrial complex based on digital standards. *Fuzzy measurements and calculations*. Vol. **10**, No. **23**. P. 29–36.

7. Akimov, S.S., Bolodurina, I.P. (2021). Construction of DSS based on the ontology of dairy production. *Ontology of design*. Vol. **11**, No. **1(29)**. P. 64–75.

8. Raikov, A.N., Abrosimov, V.K. (2022). *Intelligent agricultural robots*. Career Press: Moscow.

9. Bulakh, G.V. (2016). Livestock enterprises: environmental problems and basic environmental requirements. *NovaInfo*. Vol. **44**. P. 1–6.

10. Wankhade, B.D. [et. al.] (2024). Smart Agriculture in Southeast State of Brazil: An Overview of Technology and Adoption. *Smart Innovation, Systems and Technologies*. Vol. **397**. P. 23–34.

11. Pivoto, D. [et. al.] (2019). Factors influencing the adoption of smart farming by Brazilian grain farmers. *Int Food Agribus Manag*. Vol. **22**, No. **4**. P. 571–588.

12. Jin-Jhu Su [et. al.] (2021). A design of a solar fermentation system on chicken manure by fuzzy logic temperature control. *Applied Sciences*. Vol. **11**, No. **22**. Art. 10703.

13. Rath, P. [et. al.] (2024). Ag-IoT: Empowering Sustainable and Economic Organic Agriculture. *Smart Innovation, Systems and Technologies*. Vol. **397**. P. 477–490.

14. Narinyani, A.S. (2004). NON-factors 2004. *Proc. IX National Conference on Artificial Intelligence (CII-2004). Tver, 2004*. Vol. **1**. P. 420–432.

15. Narinyani, A.S. (2019). Introduction to Underdetermination. *Problems of Informatics*. Vol. **1**, No. **42**. P. 61–82.

16. Ignatiev, M.B. [et. al.] (2018). *Modeling of weakly formalized systems based on explicit and implicit expert knowledge*. POLYTECH-PRESS: Saint Petersburg.

17. Spesivtsev, A.V., Spesivtsev, V.A. (2024). Fuzzy-possibility approach as a tool for assessing quantitative indicators of environmental sustainability of agricultural production XIV. *Proc. All-Russian Conference on Management Problems (VSPU-2024). Moscow, June 17-20, 2024*. P. 4417–4421.

18. Vasiliev, E.V., Shalavina, E.V., Spesivtsev, V.A. (2021). Model for describing the environmental sustainability of a cattle farm. *Equipment and technologies in animal husbandry*. Vol. **4**, No. **44**. P. 93–102.

19. Zadeh, L.A. (1994). Fuzzy Logic, Neural Network and Fuzzy Computing. *Communications of the ACM*. Vol. **37**, No. **3**. P. 77–84.

20. Spesivtsev, A.V., Spesivtsev, V.A. (2014). Creation of logical-linguistic models based on implicit expert knowledge. *Certificate of state registration of computer programs*. No. **2014610613**.

LIMIT THEOREM FOR SUBMISSION PROCESS IN ONLINE CONTEST

I.N. Terekhov[1]

[1]*Lomonosov Moscow State University*
Moscow, RUSSIA
e-mail: [1]terivan14@gmail.com

We propose a sequential test for detecting inhomogeneities among problem versions in timed competitions. Modeling participant arrivals and solve times as stochastic processes, we show that under the homogeneity null hypothesis the sequential divergence statistic converges weakly to a chi-square stochastic process. This enables real-time fairness monitoring. The method provides a criterion for withdrawing problematic versions during ongoing contests.

Keywords: online contest, empirical process, Gaussian process, chi-square process, weak convergence

1 Introduction

Imagine a competition with several versions for problem lists (presumably equally complex), where users join the contest at random moments of a certain interval, receive problems, and solve a problem for some time.

The main task is to check the homogeneity of different versions during the contest. We want to remove inhomogeneous version as soon as possible. Thus, we construct a sequential test to control the homogeneity of problem list versions directly during the contest. To construct a sequential test it is necessary to study the behavior of the stochastic process corresponding to the incoming data. In this article we consider a limit theorem for such process.

2 Model

Suppose we want to check the fairness of one particular problem, this problem has m versions. We assume that the versions are distributed independently and equiprobably for each participant, each participant solves only one version, so we assume that the total number of people solving each possible version does not depend on the version and is equal to N.

Let S_i^j be the time when the i-th person from the j-th version started solving his problem, T_i^j be the total time this person spent solving the problem. We assume that for all i, j that S_i^j are independent identically distributed random variables, T_i^j are independent and, for fixed j, identically distributed random variables. Then it is natural to assume that the complexity of version is strongly correlated with the solution time. Therefore, all versions are equally complex if the solution time for all versions has the same distribution.

Let us introduce the processes that are studied in this article. Let

$$\xi_i^j(t) = I\{S_i^j + T_i^j \le t\}$$

be the indicator that the participant has solved the problem at time t,

$$\eta_i^j(t) = (T_i^j + S_i^j) \cdot I\{S_i^j + T_i^j \le t\} + t \cdot I\{S_i^j + T_i^j > t\}$$

– the time that this participant spent on the solution till time t. Then the number of participants who have solved the j-th version of the problem at time t is equal to

$$K_j(t) = \sum_{i=1}^{N} \xi_i^j(t),$$

and the total time spent by participants on solving the j-th version of the problem at time t is

$$A_j(t) = \sum_{i=1}^{N} \eta_i^j(t).$$

Consider the random fields

$$\gamma_i(t) = (\xi_i^1(t), \eta_i^1(t), \dots, \xi_i^m(t), \eta_i^m(t))^T,$$

$$\hat{X}_N(t) = (K_1(t), A_1(t), \dots, K_m(t), A_m(t))^T = \sum_{i=1}^{N} \gamma_i(t),$$

$$X(t) = (\mathbf{E}\xi_i^1(t), \mathbf{E}\eta_i^1(t), \dots, \mathbf{E}\xi_i^m(t), \mathbf{E}\eta_i^m(t))^T.$$

We consider $h(\hat{X}_N(t))$, where

$$h(x_1^1, x_2^1, \dots, x_1^m, x_2^m) = \sum_{j=1}^{m} x_1^j \ln\left(\frac{x_1^j}{x_1^1 + \dots + x_1^m} \cdot \frac{x_2^1 + \dots + x_2^m}{x_2^j}\right).$$

We prove that under the null hypothesis the random process $h(\hat{X}_N(t))$ converges in distribution to a quadratic form of a Gaussian random field in the Skorokhod space $D[0, \infty)$ as $N \to \infty$:

$$h(\hat{X}_N(t)) \xrightarrow{D} \frac{1}{2} Z(t)^T \operatorname{Hess}(h)(X(t)) Z(t),$$

where $Z(t)$ is some Gaussian random field, $\operatorname{Hess}(h)$ is the matrix of second derivatives of the function h at the point $X(t)$.

This theorem allows us to construct sequential tests for the process $h(\hat{X}_N(t))$, using the distribution of the limit process.

References

1. Billingsley P. (2013). *Convergence of probability measures*. Wiley & Sons: New York.

2. Fernholz L. T. (1983). *Von Mises calculus for statistical functionals*. Springer-Verlag: New York.

VOLATILITY PREDICTION FOR THE GARCH MODEL

N.N. Troush[1], V.P. Tsybulka[2]

[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]TroushNN@bsu.by, [2]tsybulkavladislav@gmail.com

In this paper, the focus is on analyzing the returns of financial assets using the GARCH(1,1) model and various distributions: stable, Student's t-distribution, and skewed Student's t-distribution. The work includes a theoretical analysis of the model, as well as practical application to the return data of Apple Inc, Gazprom PJSC, Severstal PJSC, Microsoft, and Nike. The results show that stable distribution models provide more accurate volatility forecasts in conditions of high uncertainty. The choice of model and distribution proves to be critical for the precision of financial analysis, emphasizing the need to use more complex distributions to forecast volatility in financial markets.

Keywords: GARCH, volatility, financial assets, distribution, forecasting

1 Introduction

Models of autoregressive conditional heteroskedasticity (ARCH) were introduced by Engle (1982), and their extension, GARCH (Generalized ARCH), belongs to Bollerslev (1986). In these models, the key concept is conditional variance, which means that the variance depends on past values. In classic GARCH models, the conditional variance is expressed as a linear function of the squares of previous values in the series. This specification allows capturing the main stylized facts characterizing financial time series. At the same time, it is simple enough to ensure a complete study of the solutions.

2 Model

Definition 1. (GARCH(p,q) Process) The process ϵ_t is called a GARCH(p,q) process if its first two conditional moments exist and satisfy:

1. $E(\epsilon_t \mid \epsilon_u, u < t) = 0, \quad t \in \mathbb{Z}$.
2. There exist constants ω, α_i for $i = 1, \ldots, q$ and β_j for $j = 1, \ldots, p$, such that

$$\sigma_t^2 = \mathrm{Var}(\epsilon_t \mid u, u < t) = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \quad t \in \mathbb{Z}. \tag{1}$$

The last equation can be written more compactly as:

$$\sigma_t^2 = \omega + \alpha(B)t^2 + \beta(B)\sigma_t^2, \quad t \in \mathbb{Z}, \tag{2}$$

where B is standard backshift operator ($B^i\epsilon_t^2 = \epsilon_{t-i}^2$ and $B^i\sigma_t^2 = \sigma_{t-i}^2$ for any integer i), and α and β are polynomials of degrees q and p, respectively:

$$\alpha(B) = \sum_{i=1}^{q} \alpha_i B^i, \beta(B) = \sum_{j=1}^{p} \beta_j B^j.$$

Definition 2. (Strong GARCH(p,q) process) Let (η_t) be a sequence of independent and identically distributed random variables with distribution η. The process (ϵ_t) is called a strong GARCH(p,q) process (with respect to the sequence (η_t)) if

$$\begin{cases} \epsilon_t = \sigma_t \eta_t, \\ \sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \end{cases} \tag{3}$$

where α_i and β_j are nonnegative constants, and ω is a (strictly) positive constant.

When $p = q = 1$, the model (3) takes the form:

$$\begin{cases} \epsilon_t = \sigma_t \eta_t, \\ \sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \end{cases} \tag{4}$$

with $\omega \geq 0, \alpha \geq 0, \beta \geq 0$.

3 Preliminary analysis

To construct descriptive statistics of the company returns, I selected the following companies: Apple Inc (AAPL), Gazprom PJSC (GAZP), Severstal PJSC (CHMF), Microsoft (MSFT) and Nike (NKE). Data were obtained from the website ru.investing.com for the period from January 2015 to May 2025 (daily). The following descriptive statistics were chosen for the table: minimum, maximum, median, first quartile, third quartile, mean, and variance. In R, these can be calculated using the functions `max()`, `min()`, `median()`, `quantile(...,0.25)`, `quantile(...,0.75)`, `mean()` and `var()`. I used the "knitr" and "kableExtra" libraries to construct the table. The following results were obtained:

Table 1: Descriptive statistics

Company	Min	Max	Median	Q1	Q3	Mean	Variance
AAPL	-75.07	300.12	0.090	-0.750	1.0175	0.1820208	40.224610
GAZP	-30.46	24.95	-0.035	-0.880	0.8925	0.0240686	4.530015
CHMF	-22.11	11.44	0.050	-0.960	1.0800	0.4246404	3.912391
MSFT	-14.74	12.42	0.085	-0.6875	0.9700	0.1009575	2.984786
NKE	-19.99	15.53	0.030	-0.620	0.8000	0.0630000	3.699889

Based on the descriptive statistics obtained, the following conclusions can be made: **AAPL:** Positive returns, high stability, and predictability make it attractive for investors. **GAZP:** High volatility and risk, along with a near-zero average change, make

it less appealing to conservative investors. **CHMF:** Moderate returns and volatility place this company between AAPL and GAZP. **MSFT:** Balanced returns and moderate volatility make it suitable for investors seeking stable assets. **NKE:** Moderate returns with an acceptable level of risk make it interesting for those who prefer assets with growth potential.

4 Modeling GARCH(1,1) Processes. Estimation of the Parameter Vector

We will use the "rugarch" library for modeling. To check the accuracy of the modeling, we will use the Lewis-Lee test to check for autocorrelation in the residuals. Additionally, we will estimate the parameter vector of the constructed GARCH(1,1) model with the considered distributions.

Table 2: GARCH(1,1) model parameters and Lewis-Lee test results

Parameter	**Stable**	**Student's t**	**Skewed Student's t**
mu	1.661366e-02	-6.540943e-02	-6.406370e-02
omega	1.055867e-03	1.032074e-03	1.037009e-03
alpha1	1.345808e-09	1.168474e-09	3.746238e-10
beta1	9.98999e-01	9.989999e-01	9.990000e-01
skew	1.029812e+00	-	1.03287e+00
shape	5.999953e+01	9.983776e+01	5.999935e+01
p-value	0.153	0.07808	0.1527

5 Volatility forecast

We will build GARCH(1,1) models based on known returns (daily from January 2015 to May 2025) for the companies Apple Inc (AAPL), Gazprom PJSC (GAZP), Severstal PJSC (CHMF), Microsoft (MSFT), and Nike (NKE) with the following distributions: stable, Student's t-distribution, and skewed Student's t-distribution. We will also estimate volatility and forecast it for 6 months ahead. To do this, we will use the following algorithm:

1. Define the specification of the GARCH(1,1) model depending on the distribution considered: ugarchspec().

2. Fit the model to the data using the function ugarchfit().

3. Estimate volatility using the function sigma(). Plot the volatility.

4. Evaluate residuals using the function residuals(), then plot them to analyze their behavior and visually assess the estimated volatility.

5. Forecast volatility for the specified horizon of 6 months using the function ugarchforecast(). Plot the forecasted volatility for future periods.

6 Conclusion

The model with the Student's t-distribution demonstrated higher robustness to outliers, making it preferable for analyzing financial time series where such outliers frequently occur.

The skewed Student's t-distribution, in turn, provides additional flexibility by allowing for the modeling of asymmetry in the data. This is particularly important for financial time series, where both positive and negative outliers can significantly impact volatility. Models with skewed distributions show more accurate results in conditions where the data exhibit pronounced asymmetry.

The stable distribution, unlike the others, possesses greater flexibility due to its parameters. This model can adapt to various conditions and can be used to describe a wide range of data, including those with asymmetric distributions and infinite moments.

The comparison of models showed that the choice of distribution significantly impacts the estimation of volatility. Based on the conducted analysis, it is recommended to use GARCH(1,1) models with Student's t and its skewed variants or stable distributions for forecasting volatility in financial markets, especially under conditions of high uncertainty. These models allow for a more accurate consideration of data characteristics and provide better quality forecasts.

Thus, the analysis indicated that the choice of model and distribution is critical for the accuracy of volatility forecasting. Given the dynamics of financial markets and the presence of outliers, models with more complex distributions generally yield more reliable and informative results for analysis and decision-making.

References

1. ARCH, GARCH, EGARCH. How to measure volatility in equities. [Electronic resource]. Access mode: `https://medium.com/@NNGCap/arch-garch-egarch-92dd7277a966`. Access date: 15.04.2025.

2. Francq C. GARCH models Structure, Statistical Inference and Financial Applications // C. Francq, J. Zakoian. England: John Wiley and Sons, 2010.

3. How can ARCH/GARCH models forecast volatility? [Electronic resource]. Access mode: `https://www.linkedin.com/advice/1/how-can-archgarch-models-forecast-volatility-skills-economics-ye7tc`. Access date: 20.05.2025.

REDUCED PROCESSES IN NON-FAVORABLE RANDOM ENVIRONMENT

V.A. Vatutin[1], E.E. Dyakonova[2]

[1,2]*Steklov Mathematical Institute of Russian Academy of Sciences*
Moscow, RUSSIA
e-mail: [1]vatutin@mi-ras.ru, [2]elena@mi-ras.ru

Let $\{Z_n, n = 0, 1, \dots\}$ be a critical branching process in i.i.d. random environment, $Z_{r,n}$ be the number of particles in the process at moment $0 \le r \le n-1$ that have a positive number of descendants in generation n, and $\{S_n, n = 0, 1, \dots\}$ be the associated random walk of $\{Z_n, n = 0, 1, \dots\}$. It is known that if $\mathbf{E}S_1 = 0$ and $\sigma^2 = \mathbf{E}S_1^2 \in (0, \infty)$, then, for any $t \in [0, 1]$

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\log Z_{[nt],n}}{\sigma\sqrt{n}} \le x \Big| Z_n > 0\right) = \mathbf{P}\left(\min_{t\le s\le 1} B_s^+ \le x\right),\ x \in [0, \infty),$$

where $\left\{B_t^+, 0 \le t \le 1\right\}$ is the Brownian meander.
We supplement this result by description of the distribution of the properly scaled random variable $\log Z_{r,n}$ under the condition $\left\{S_n \le t\sqrt{k}, Z_n > 0\right\}$, where $t > 0$ and $r, k \to \infty$ in such a way that $k = o(n)$ as $n \to \infty$.

Keywords: reduced branching process, random environment, limit theorem

1 Introduction

We consider critical branching processes evolving in an non-favorable random environment. Let $\mathfrak{F} = \{\mathfrak{f}\}$ be the space of all probabilistic generating functions on the set $\{0, 1, \dots\}$. Let

$$F(s) := \sum_{k=0}^{\infty} F(\{k\})s^k, \quad s \in [0, 1],$$

be a random variable taking values in $\mathfrak{F}$, and

$$F_n(s) := \sum_{k=0}^{\infty} F_n(\{k\})s^k, \quad s \in [0, 1], \quad n \ge 0,$$

be a sequence of independent probabilistic copies of F. The infinite sequence $\mathcal{E} = \{F_n, n \ge 0\}$ is called a random environment.

A sequence of non-negative integer-valued random variables $\mathcal{Z} = \{Z_n,\ n \ge 0\}$, given on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, is called a branching process in a random environment (BPRE), if Z_0 is independent of $\mathcal{E}$ and, given $\mathcal{E}$ the process $\mathcal{Z}$ is a Markov chain in which

$$\mathcal{L}\left(Z_n | Z_{n-1} = z_{n-1}, \mathcal{E} = (f_1, f_2, \dots)\right) = \mathcal{L}(\xi_{n1} + \dots + \xi_{nz_{n-1}})$$

for all $n \ge 1$, $z_{n-1} \ge 0$ and $f_1, f_2, \dots \in \mathfrak{F}$, where $\xi_{n1}, \xi_{n2}, \dots$ is a sequence of independent identically distributed random variables with a distribution given by the generating function f_n.

A sequence

$$S_0 = 0, \quad S_n = X_1 + \cdots + X_n, \ n \geq 1,$$

where $X_i = \log F_i'(1), i = 1, 2, \ldots,$ is called the associated random walk for the process $\mathcal{Z}$.

The growth rate of the population size of the BPRE significantly depends on the properties of the associated random walk $\mathcal{S} = \{S_n, n \geq 0\}$. It is this phenomenon we investigate in the present paper.

2 Main results

We assume that

$$\mathbf{E}X_1 = 0, \quad \sigma^2 = \mathbf{E}X_1^2 \in (0, \infty). \tag{1}$$

Hence it follows that, as $n \to \infty$

$$\left\{ \frac{S_{[nt]}}{\sigma\sqrt{n}}, t \geq 0 \right\} \Longrightarrow \mathcal{B} = \{B_t, t \geq 0\},$$

where $\mathcal{B}$ is the standard Brownian motion and the symbol $\Longrightarrow$ denotes convergence in distribution in the space of functions that are continuous on the right and have limits on the left, equipped with a Skorokhod topology.

We now formulate the first condition that we impose on the analyzed BPRE.

Condition B1. Random variables $X_n, n \geq 1$, are independent copies of a random variable X satisfying condition (1) and having absolutely continuous distribution. Besides, there is an $n \geq 1$ such that the density $\mathbf{P}(S_n \in dx)/dx$ of the random variable S_n is bounded.

Since the associated random walk oscillate, it follows that we consider a critical BPRE (see [1] and [2]).

Our next condition on the properties of the random environment concerns the reproduction law of particles. Set

$$\eta := \frac{\sum_{i=1}^{\infty} i^2 F(\{i\})}{\left(\sum_{i=0}^{\infty} iF(\{i\})\right)^2}.$$

Condition B2. There is a number $\varkappa > 0$ such that

$$\mathbf{E}[\log^{2+\varkappa} \max(\eta, 1)] \ < \ \infty.$$

This condition excludes from consideration BPRE's with extremely productive particles.

Let $Z_{r,n}$ be the number of particles at moment $r \in [0, n-1]$, having positive number of descendants at time n, and let $Z_{n,n} = Z_n$. Given n the process

$$\mathcal{Z}_{red} := \{Z_{r,n}, r = 0, 1, \ldots, n\}$$

is called a reduced process on the interval $[0, n]$ or simply a reduced process.

It is known [3] that if condition (1) is valid and some additional technical conditions fulfilled then for any $t > 0$ and $s \in [0, 1]$

$$\begin{aligned} \mathbf{P}\left(\frac{1}{\sigma\sqrt{n}}\log Z_{[sn],n} \le t \Big| Z_n > 0\right) &\sim \mathbf{P}\left(\frac{1}{\sigma\sqrt{n}}\min_{[sn]\le m\le n} S_m \le t \Big| Z_n > 0\right) \\ &\sim \mathbf{P}\left(\inf_{s\le q\le 1} B_q^+ \le t\right) \end{aligned}$$

as $n \to \infty$, where $\left\{B_q^+, 0 \le q \le 1\right\}$ is a Brownian meander, i.e. a Brownian motion, considered to be nonnegative on the interval $[0, 1]$.

We study the distribution of the random variable $Z_{r,n}$ in cases when $\min(r, n-r) \to \infty$ as $n \to \infty$, and the random variable S_n is bounded from above by some function depending on n, the growth order of which is smaller than $\sqrt{n}$.

Theorem 1. *Let conditions $B1$ and $B2$ be valid. If $n \gg k \gg m = n - r \to \infty$, then, for any $z \in (-\infty, +\infty)$ and $t > 0$*

$$\lim_{n\to\infty} \mathbf{P}\left(\log Z_{r,n} - S_r \le \sigma z\sqrt{m} | S_n \le \sigma t\sqrt{k}, Z_n > 0\right) = \mathbf{P}\left(\min_{0\le s\le 1} B_s \le z\right).$$

Theorem 2. *Let conditions $B1$ and $B2$ be valid. If there is $\theta > 0$ such that $k \sim \theta m = \theta(n - r) \to \infty$ as $n \to \infty$ and $m = o(n)$, then, for all $y \ge 0$ and $t > 0$*

$$\lim_{n\gg k\to\infty} \mathbf{P}\left(\log Z_{r,n} \le \sigma y\sqrt{m} | S_n \le \sigma t\sqrt{k}, Z_n > 0\right) = A(\sqrt{\theta}t, \sqrt{\theta}\,(y \wedge t)),$$

where

$$A(T, y) = \frac{2}{T^2}\int_0^\infty w\mathbf{P}\left(-w \le \min_{0\le s\le 1} B_s \le y - w; B_1 \le T - w\right) dw.$$

References

1. Afanasyev V.I., Geiger J., Kersting G., Vatutin V.A. (2005). Criticality for branching processes in random environment. *Annals of Probability* Vol. **33**, Num. **2**. pp. 645-673.

2. Kersting G., Vatutin V. (2017). *Discrete Time Branching Processes in Random Environment*. Wiley: New Jersey.

3. Vatutin V.A. (2003) Reduced branching processes in random environment: the critical case. *Theory of Probability and Its Applications*. Vol. **47**, Num.**1**, pp. 99-113.

LOCAL INFORMATION GEOMETRY FOR HIGH-ORDER BINARY MARKOV CHAINS AND ITS APPLICATIONS

V.A. Voloshko[1]

[1]*Research Institute for Applied Problems of Mathematics and Informatics*
[1]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]valoshka@bsu.by

We present some new results on asymptotic properties of statistics derived from purely random (uniformly distributed) binary sequence of length $n \to +\infty$. These results are obtained by methods of information geometry applied to manifolds of Markov probability distributions on the set of infinite binary sequences. We briefly describe underlying information-geometric theory and technics of proofs and show how finding probabilistic and statistical properties comes down to geometric and combinatorial computations.

Keywords: information geometry, tangent space, asymptotic distribution, binary sequence, exponential family of Markov probability distributions

1 New results

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{V} ::= \{0, 1\}$ be "purely random" (uniformly distributed) binary sequence of length $n \in \mathbb{N}$: $\mathsf{P}\{\mathbf{x}_1^n = q\} \equiv 2^{-n}$, $q \in \mathbf{V}^n$, $\mathbf{x}_1^n ::= (\mathbf{x}_i)_{i=1}^n$. For convenience, we consider index i of $\mathbf{x}_i$ as discrete time and use corresponding "temporal" terminology. Below we give some examples of our new results on asymptotic properties of statistics derived from $\mathbf{x}_1^n$ as $n \to +\infty$. These results develop the earlier obtained ones [1, 2].

Introduce some notations: "$a|b$" means "a divides b"; $\mathcal{L}\{\cdot\}$ is the probability distribution of a random variable; $\mathsf{cov}\{\cdot,\cdot\}$, $\mathsf{cor}\{\cdot,\cdot\}$ are respectively the covariance and the correlation coefficient for a pair of random variables; $\mathbb{Z}_m = \{0, \ldots, m-1\}$ is the ring of integers modulo $m \in \mathbb{N}$ standardly associated with a nonnegative integer interval; $a \overset{m}{\equiv} b$ and $a \overset{m}{+} b$ are respectively equality and summation modulo m; $\mathrm{id}_m \in \mathbb{R}^{m\times m}$, $0_m \in \mathbb{R}^m$ are respectively the identity matrix of order $m \in \mathbb{N}$ and zero m-vector; $\mathtt{span}\{\cdot\}$ is a linear span of subset of linear space.

1.1 Asymptotic covariances of L-grams frequencies

Let us fix some finite pattern of indices $J = \{j_1, \ldots, j_L\} \subset \mathbb{N}$, $1 = j_1 < j_2 < \cdots < j_L$, $1 \le L = |J| < +\infty$, and denote by $\mathbf{x}_J ::= (\mathbf{x}_j)_{j\in J} \in \mathbf{V}^L$ the binary L-gram corresponding to the indices from J. Also let us call moving J-window any shifted pattern $i + J = \{i + j_1, \ldots, i + j_L\} \subset \mathbb{Z}$, $i \in \mathbb{Z}$. Consider the following frequencies of L-grams within moving J-windows:

$$f^{\{J\}} ::= (f^{\{J\}}(q))_{q\in\mathbf{V}^L} \in \mathbb{R}^{2^L}, \quad f^{\{J\}}(q) ::= \sum_{i=0}^{n-j_L} \mathbb{1}\{\mathbf{x}_{i+J} = q\}, \quad q \in \mathbf{V}^L. \qquad (1)$$

Here and in further notations we omit sequence length n for brevity. Covariance matrix of the frequencies vector (1) has the following asymptotics:

$$\mathbf{cov}\left\{f^{\{J\}}, f^{\{J\}}\right\} \underset{n\to+\infty}{=} n \cdot C^{\{J\}} + \mathcal{O}(1) \in \mathbb{R}^{2^L \times 2^L}. \tag{2}$$

This covariance matrix appears in many applications, mainly for standard "solid" patterns $J = \{1, \ldots, L\}$ without "holes". The case of arbitrary "sparse" patterns J concerns binary parsimonious high-order Markov chains with partial connections [3], whose sufficient statistics have the form (1). The main term $C^{\{J\}} \in \mathbb{R}^{2^L \times 2^L}$ of (2) has highly complicated irregular structure of entries (pairwise asymptotic covariances of frequencies (1), see [4] for instance), which makes its analysis quite a hard problem. However, the spectrum of eigenvalues of $C^{\{J\}}$ can be completely obtained.

Let us call two nonempty finite subsets $J', J'' \subset \mathbb{N}$ shift equivalent ($J' \sim J''$), if $J'' = i + J'$ for some $i \in \mathbb{Z}$. Denote by $2^J_+ ::= \{J' \subset J : J' \neq \varnothing\}$ the set of all nonempty subsets of pattern J.

Theorem 1. *Let $J_1, \ldots, J_K \subset 2^J_+$, $\sqcup_{k=1}^K J_k = 2^J_+$, be classes of shift equivalence of nonempty subsets of pattern J. Matrix $C^{\{J\}}$ has exactly K positive eigenvalues (taking into account multiplicity):*

$$\lambda_k(C^{\{J\}}) = 2^{-L}|J_k|, \quad k = 1, \ldots, K. \tag{3}$$

The corresponding eigenvectors for eigenvalues (3) can also be obtained in explicit form. Let us illustrate Theorem 1 by a brief example. Take $J = \{1, 2, 3, 5\}$, $L = |J| = 4$. The set 2^J_+ of 15 nonempty subsets $J' \subset J$ breaks into the following $K = 10$ classes of shift equivalence (in nondecreasing order of classes powers):

$$J_1 : \{1\},\ \{2\},\ \{3\},\ \{5\}; \quad \lambda_1 = \tfrac{4}{16};$$
$$J_2 : \{1,2\},\ \{2,3\}; \quad \lambda_2 = \tfrac{2}{16}; \quad J_3 : \{1,3\},\ \{3,5\}; \quad \lambda_3 = \tfrac{2}{16};$$
$$J_4 : \{2,5\}; \quad \lambda_4 = \tfrac{1}{16}; \quad J_5 : \{1,5\}; \quad \lambda_5 = \tfrac{1}{16}; \quad J_6 : \{1,2,3\}; \quad \lambda_6 = \tfrac{1}{16};$$
$$J_7 : \{1,2,5\}; \quad \lambda_7 = \tfrac{1}{16}; \quad J_8 : \{1,3,5\}; \quad \lambda_8 = \tfrac{1}{16}; \quad J_9 : \{2,3,5\}; \quad \lambda_9 = \tfrac{1}{16};$$
$$J_{10} : \{1,2,3,5\}; \quad \lambda_{10} = \tfrac{1}{16}.$$

The eigenvalues 1/16, 2/16 and 4/16 have multiplicities 7, 2 and 1 respectively (the remaining 6 eigenvalues of $C^{\{J\}} \in \mathbb{R}^{16\times 16}$ are zeroes).

For $L = 1$ it is always $K = 1$, $\lambda_1 = 1/2$. For $L = 2$: $K = 2$, $\lambda_1 = 2/4$, $\lambda_2 = 1/4$. For $L = 3$ there are two cases. The first case: $j_2 - j_1 = j_3 - j_2$ (scaled solid pattern), $K = 4$, $\lambda_1 = 3/8$, $\lambda_2 = 2/8$, $\lambda_3 = \lambda_4 = 1/8$. The second case: $j_2 - j_1 \neq j_3 - j_2$, $K = 5$, $\lambda_1 = 3/8$, $\lambda_1 = \cdots = \lambda_5 = 1/8$.

It follows from Theorem 1 that the trace

$$\mathtt{tr}(C^{\{J\}}) = \sum_{k=1}^{K} \lambda_k = 2^{-L}|2^J_+| = 1 - 2^{-L} \tag{4}$$

does not depend on the shape of pattern J, but only on its size $L = |J|$. In [4] the expression (4) was obtained for the solid pattern $J = \{1, \ldots, L\}$. The extremal eigenvalues (among the positive ones) also do not depend on the shape of pattern J:

$$\lambda_{\max}(C^{\{J\}}) = 2^{-L}|J_1| = L \cdot 2^{-L}, \quad \lambda^+_{\min}(C^{\{J\}}) = 2^{-L}|J_K| = 2^{-L}.$$

Except J_1 consisting of L one-element subsets $\{j_i\}$, for other classes obviously $|J_k| < L$, whence $\lambda_{\max}(C^{\{J\}})$ is always a simple eigenvalue.

1.2 Asymptotic dependency of chi-square statistics

Let us fix some step $\Delta \in \mathbb{N}$, gram length $L \in \mathbb{N}$ and consider analogue of frequencies (1):

$$f^{(L,\Delta)}(q) = \sum_{i=0}^{N^{(L,\Delta)}-1} \mathbb{1}\left\{\mathbf{x}_{i\Delta+\{1,\ldots,L\}} = q\right\}, \quad q \in \mathbf{V}^L, \quad N^{(L,\Delta)} = \left\lfloor \frac{n-L}{\Delta} \right\rfloor + 1. \tag{5}$$

For $\Delta = 1$ and solid pattern $J = \{1, \ldots, L\}$ (1) and (5) are equivalent: $f^{(L,1)}(q) \equiv f^{\{1,\ldots,L\}}(q)$, $q \in \mathbf{V}^L$. Denote standard chi-square statistics based on frequencies (5):

$$\gamma^{(L,\Delta)} = \sum_{q \in \mathbf{V}^L} \frac{(f^{(L,\Delta)}(q) - 2^{-L}N^{(L,\Delta)})^2}{2^{-L}N^{(L,\Delta)}}. \tag{6}$$

For $\Delta \geq L$ the L-grams in sum (5) do not overlap, whence $\gamma^{(L,\Delta)}$ has asymptotic chi-square distribution. For $\Delta < L$ this asymptotic distribution is generalized chi-square one. Consider two types of chi-square statistics based on (6):

$$\mathbf{M}_1^{(s,\Delta)} = \gamma^{((s+1)\Delta,\Delta)} - \gamma^{(s\Delta,\Delta)}, \qquad D_1^{(s,\Delta)} = 2^{s\Delta}(2^\Delta - 1), \qquad s \geq 0, \tag{7}$$

$$\mathbf{M}_2^{(s,\Delta)} = \gamma^{((s+1)\Delta,\Delta)} - 2\gamma^{(s\Delta,\Delta)} + \gamma^{((s-1)\Delta,\Delta)}, \quad D_2^{(s,\Delta)} = 2^{(s-1)\Delta}(2^\Delta - 1)^2, \quad s \geq 1, \tag{8}$$

where $\gamma^{(0,\Delta)} \equiv 0$. Statistics $\mathbf{M}_i^{(s,\Delta)}$ of both types $i \in \{1, 2\}$ have asymptotic chi-square distributions with $D_i^{(s,\Delta)}$ degrees of freedom. Statistics of the second type $\mathbf{M}_2^{(s,\Delta)}$ are asymptotically mutually independent for different $s \geq 1$ under any fixed step $\Delta \in \mathbb{N}$.

Statistic of the first type (7) for any fixed $s \geq 0$ is related to the family $\mathrm{MC}(s, \Delta)$ of stationary fully connected Markov chains of order s formed by non-overlapping Δ-blocks $\mathbf{x}_{i\Delta+\{1,\ldots,\Delta\}} \in \mathbf{V}^\Delta$, $i \in \mathbb{N}_0$. Number of degrees of freedom equals dimensionality (number of parameters) of this family: $\dim(\mathrm{MC}(s, \Delta)) = D_1^{(s,\Delta)}$. Similarly, statistic of the second type (8) for any fixed $s \geq 1$ is related to the subfamily $\mathrm{UMC}(s, \Delta) \subset \mathrm{MC}(s, \Delta)$ of Markov chains with uniform stationary distribution of s-grams of Δ-blocks, i.e., $(s \cdot \Delta)$-blocks $\mathbf{x}_{\{1,\ldots,s\Delta\}} \in \mathbf{V}^{s\Delta}$ of binary sequence.

Chi-square statistical test based on $\mathbf{M}_1^{(0,\Delta)}$ is the standard test of uniform distribution of non-overlapping Δ-blocks $\mathbf{x}_{i\Delta+\{1,\ldots,\Delta\}} \in \mathbf{V}^\Delta$, $i \in \mathbb{N}_0$ (FIPS Poker test [5] for $\Delta = 4$). Tests based on $\mathbf{M}_1^{(s,1)}$ and $\mathbf{M}_2^{(s,1)}$ are the NIST Serial tests [6] of types I and II respectively (with parameter $m = s + 1$). Test based on $\mathbf{M}_1^{(0,1)}$ is the Monobit test. Test based on $\mathbf{M}_2^{(1,1)}$ is asymptotically equivalent to the Runs test.

Introduce piecewise linear functions (see Figure 1):

$$\boldsymbol{\rho}_1(x) = \max\{0, 1 - \max\{0, x\}\}, \quad \boldsymbol{\rho}_2(x) = \max\{0, 1 - |x|\}, \quad x \in \mathbb{R}. \tag{9}$$

Difference relation holds: $\boldsymbol{\rho}_1(x) - \boldsymbol{\rho}_1(x+1) \equiv \boldsymbol{\rho}_2(x)$. The following result describes asymptotic dependencies between statistics (7), (8).

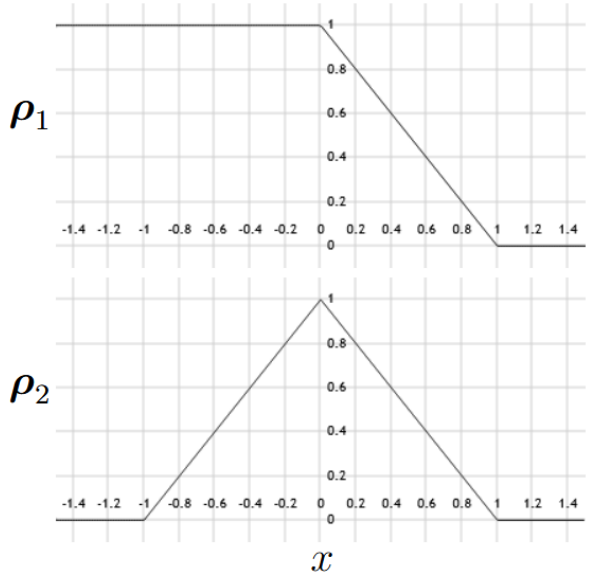


Figure 1: Functions (9)

Theorem 2. *Let* $(i,j) \in \{(2,2),(1,2),(1,1)\}$. *Asymptotic covariances of statistics* (7), (8) *are as follows:*

$$\lim \mathbf{cov}\left\{\mathbf{M}_i^{(s,\Delta)}, \mathbf{M}_j^{(s',\Delta')}\right\} = C + 2\frac{(2^\kappa-1)^2}{2^\kappa} \times \sum_{m_-\le m\le m_+} \boldsymbol{\rho}_i\left(\frac{m}{a}-s\right)\boldsymbol{\rho}_j\left(\frac{m}{a'}-s'\right)2^{m\kappa}, \qquad (10)$$

where $\kappa = \gcd(\Delta,\Delta')$ *(greatest common divisor),* $a = \Delta/\kappa$, $a' = \Delta'/\kappa$, $C = 2(2^\kappa-1)$ *for* $(i,j) = (1,1)$ *and* $C = 0$ *otherwise,* $m_+ = \min\{(s+1)a, (s'+1)a'\} - 1$,

$$m_- = \begin{cases} \max\{(s-1)a, (s'-1)a'\}+1, & (i,j)=(2,2),\\ (s'-1)a'+1, & (i,j)=(1,2),\\ 1, & (i,j)=(1,1). \end{cases}$$

Thus, asymptotic covariances of statistics (7), (8) are expressed as some weighted dot products over integer lattice of shifts and stretchings of functions (9). For $i=j$, $(s,\Delta) = (s',\Delta')$ a simple calculation allows to verify that (10) gives variance of chi-square distribution with $D_i^{(s,\Delta)}$ degrees of freedom:

$$\lim \mathbf{cov}\left\{\mathbf{M}_i^{(s,\Delta)}, \mathbf{M}_j^{(s',\Delta')}\right\} = \lim \mathbf{var}\left\{\mathbf{M}_i^{(s,\Delta)}\right\} = 2D_i^{(s,\Delta)}.$$

1.3 Asymptotic shift sensitivity of block-based chi-square statistics

Consider distortion of the binary sequence $\{\mathbf{x}_i\}$ when few first symbols $\mathbf{x}_1,\dots,\mathbf{x}_\delta$ are missed, $\delta \in \mathbb{N}_0$, and δ-shortened sequence $\mathbf{x}_{1+\delta}^n$ goes to the input of statistical test

instead of full sequence $\mathbf{x}_1^n$ (let us call it δ-distortion, $\delta = 0$ means no distortion). It is quite obvious that Monobit and Runs tests statistics are distorted insignificantly and stay asymptotically equivalent to the undistorted ones (δ is fixed as $n \to +\infty$). But the picture is completely different for the "block-based" tests that use non-overlapping Δ-blocks $\mathbf{x}_{i\Delta+\{1,\ldots,\Delta\}} \in \mathbf{V}^\Delta$, $i \in \mathbb{N}_0$, for computation. Say, the test of uniform distribution of non-overlapping 10-blocks (statistics $\mathbf{M}_1^{(0,10)}$) uses the very different sets of 10-grams under 6-distortion and without it:

$$\begin{aligned} &\text{undistorted case:} \quad (\mathbf{x}_1, \ldots, \mathbf{x}_{10}), \quad (\mathbf{x}_{11}, \ldots, \mathbf{x}_{20}), \quad (\mathbf{x}_{21}, \ldots, \mathbf{x}_{30}), \ \ldots; \\ &\text{6-distorted case:} \quad (\mathbf{x}_7, \ldots, \mathbf{x}_{16}), \quad (\mathbf{x}_{17}, \ldots, \mathbf{x}_{26}), \quad (\mathbf{x}_{27}, \ldots, \mathbf{x}_{36}), \ \ldots. \end{aligned}$$

Let us call two identically chi-square distributed random variables $\xi_1, \xi_2 \sim \chi^2_D$ semi-independent, if there exist three mutually independent chi-square distributed random variables $\eta_0 \sim \chi^2_d$, $\eta_1, \eta_2 \sim \chi^2_{D-d}$, $0 \le d \le D$, such that $\xi_i = \eta_i + \eta_0$, $i \in \{1, 2\}$. For $d = 0$ semi-independence is equivalent to independence, while $d = D$ means $\xi_1 = \xi_2$. Correlation coefficient of semi-independent chi-square distributed random variables have the following simple form:

$$\mathsf{cor}\,\{\xi_1, \xi_2\} = \frac{\mathsf{cov}\,\{\xi_1, \xi_2\}}{\sqrt{\mathsf{var}\,\{\xi_1\}\,\mathsf{var}\,\{\xi_2\}}} = \frac{d}{D}.$$

The following result describes asymptotic sensitivity of statistics (7), (8) to δ-distortion (shift sensitivity).

Theorem 3. *For $0 \le \delta \le \Delta$ undistorted and δ-distorted statistics (7), (8) are asymptotically semi-independent with the following asymptotic correlation coefficients:*

$$\lim \mathsf{cor}\left\{\mathbf{M}_1^{(s,\Delta)}[\mathbf{x}_1^n], \mathbf{M}_1^{(s,\Delta)}[\mathbf{x}_{1+\delta}^n]\right\} = \frac{2^\delta + 2^{\delta'} - 2}{2^\Delta - 1}, \tag{11}$$

$$\lim \mathsf{cor}\left\{\mathbf{M}_2^{(s,\Delta)}[\mathbf{x}_1^n], \mathbf{M}_2^{(s,\Delta)}[\mathbf{x}_{1+\delta}^n]\right\} = \frac{(2^\delta + 2^{\delta'})(2^\Delta + 1) - 2^{2+\Delta}}{(2^\Delta - 1)^2}, \tag{12}$$

where $\delta' = \Delta - \delta$.

Note that the values (11), (12) do not depend on parameter s. Their plots for $\Delta = 7$ are presented in Figure 2. These plots are very close to each other, and they indeed are asymptotically equivalent (i.e., ratio goes to unity) uniformly over $0 \le \delta \le \Delta$ as $\Delta \to +\infty$. The plots are minimal at "half-period" $\delta = \Delta/2$ ($\delta = (\Delta \pm 1)/2$ for odd Δ) and maximal at full period $\delta \in \{0, \Delta\}$. The minimal values have asymptotics $\mathsf{cor}_{\min} \underset{\Delta\to+\infty}{\sim} C \cdot 2^{-\Delta/2}$, where $C = 2$ for even Δ and $C = \frac{3}{2}\sqrt{2}$ for odd Δ. In the neighborhood of half-period the plots tend to the scaled hyperbolic cosine. In the neighborhood of full period adjacent values differ approximately twofold for large Δ.

2 Theory and technics of proofs

Presented results are obtained by methods of information geometry [7] applied to manifolds of Markov probability distributions (Markov manifolds for brevity) on the set $\mathbf{V}^{\mathbb{Z}} = \{\mathbf{x} = (\mathbf{x}_i)_{i\in\mathbb{Z}} : \mathbf{x}_i \in \mathbf{V}\}$ of two-sided infinite binary sequences.

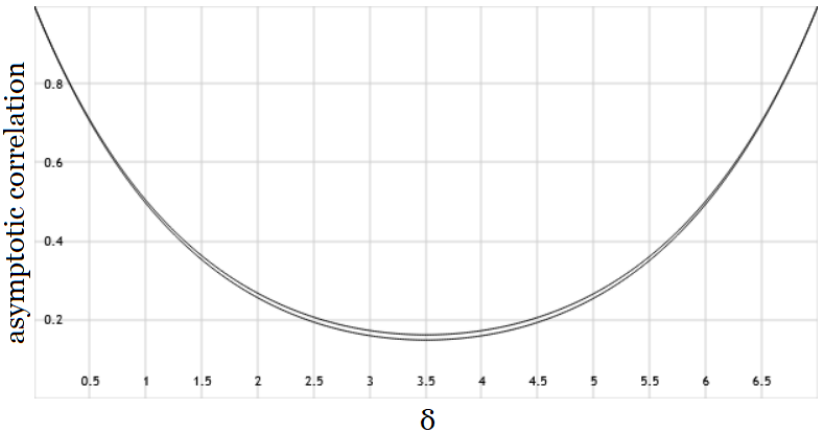


Figure 2: Asymptotic correlation coefficients (11), (12) plotted against δ ($\Delta = 7$)

2.1 Markov manifolds and tangent spaces

We consider two nested Markov manifolds.

The first Markov manifold $\mathcal{MC}$ consists of distributions of stationary Markov chains of all orders $s \in \mathbb{N}_0$ determined by stationarity condition and Markov equation:

$$\mathsf{P}\left\{\mathbf{x}_i = 0 | \mathbf{x}_{i-j} = q_j, j \in \mathbb{N}\right\} = \frac{1 + \boldsymbol{\omega}(q)}{2}, \quad q = (q_j)_{j\in\mathbb{N}} \in \mathbf{V}^{\mathbb{N}}, \quad i \in \mathbb{Z}, \tag{13}$$

where the function $\boldsymbol{\omega} : \mathbf{V}^{\mathbb{N}} \to \mathbb{R}$ depends only on finite number of components $\{q_j\}$ and $\|\boldsymbol{\omega}\|_\infty = \max_{q\in\mathbf{V}^{\mathbb{N}}} |\boldsymbol{\omega}(q)| < 1$. We call order of $\boldsymbol{\omega}$ and denote by $\mathtt{ord}(\boldsymbol{\omega})$ the minimal value $s \in \mathbb{N}_0$ such that $\boldsymbol{\omega}(q)$ does not depend on $\{q_j : j > s\}$. Manifold $\mathcal{MC}$ is in a bijection with the set of functions $\boldsymbol{\omega}$, so we identify them:

$$\mathcal{MC} = \{\boldsymbol{\omega} : \mathtt{ord}(\boldsymbol{\omega}) < +\infty, \|\boldsymbol{\omega}\|_\infty < 1\}.$$

The uniform distribution $\mathfrak{u} \in \mathcal{MC}$ corresponds to zero function $\boldsymbol{\omega}(q) \equiv 0$, $q \in \mathbf{V}^{\mathbb{N}}$. If we take off the norm limitation $\|\boldsymbol{\omega}\|_\infty < 1$, we get the model of tangent space [7] (linearized neighborhood) of uniform distribution $\mathfrak{u}$ within the Markov manifold $\mathcal{MC}$:

$$\mathbf{T} = \{\boldsymbol{\omega} : \mathtt{ord}(\boldsymbol{\omega}) < +\infty\}.$$

The fundamental object of information geometry defined on this tangent space is the Fisher-Riemann dot product [7]. It has the following form in our notation:

$$\langle \boldsymbol{\omega}, \boldsymbol{\omega}' \rangle = 2^{-s} \sum_{q\in\mathbf{V}^s} \boldsymbol{\omega}(q\|0^\infty)\boldsymbol{\omega}'(q\|0^\infty), \quad \boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathbf{T}, \quad s = \max\{\mathtt{ord}(\boldsymbol{\omega}), \mathtt{ord}(\boldsymbol{\omega}')\},$$

where "$\|$" is concatenation, $0^\infty \in \mathbf{V}^{\mathbb{N}}$ is one-sided zero binary sequence. We call elements $\boldsymbol{\omega} \in \mathbf{T}$ tangent vectors, that is the standard name for elements of tangent spaces.

Under "local information geometry" ("locality" refers to neighborhood of $\mathfrak{u}$) we understand here the structure of tangent space $\mathbf{T}$ with Fisher-Riemann dot product

on it, the objects defined on this structure and the meaningful quantities that can be computed based on it.

An orthonormal base of characters [8] $\boldsymbol{\chi} = \{\boldsymbol{\chi}^{(Q)}\}_{Q\in\mathbf{V}^+}$ is defined on $\mathbf{T}$:

$$\boldsymbol{\chi}^{(Q)} \in \mathbf{T}, \quad \boldsymbol{\chi}^{(Q)}(q) ::= \mathbf{e}\left(\bigoplus_{j\in\mathbb{N}} Q_j q_j\right), \quad Q \in \mathbf{V}^+, \quad q \in \mathbf{V}^{\mathbb{N}}, \tag{14}$$

where $\mathbf{V}^+ \subset \mathbf{V}^{\mathbb{N}}$ is the subset of one-sided binary sequences of the finite Hamming weight $\|Q\| ::= \sum_{j\in\mathbb{N}} Q_j < +\infty$ (with finite number of ones), $\mathbf{e}(z) ::= (-1)^z$ is a "discrete exponent". Infinite XOR in (14) is correct, because it contains only a finite number of ones die to $Q \in \mathbf{V}^+$. Note that $\mathbf{V}^+$ is a countable subset of continuous set $\mathbf{V}^{\mathbb{N}}$. Many important tangent subspaces $\mathcal{T} \subset \mathbf{T}$ are linear spans of subsets of the characters base (14): we call $\boldsymbol{\chi}$-canonic such subspaces $\mathcal{T}$. The notion of order is naturally transferred from $\mathbf{T}$ to $\mathbf{V}^+$ by (14):

$$\mathtt{ord}(Q) ::= \mathtt{ord}(\boldsymbol{\chi}^{(Q)}), \quad Q \in \mathbf{V}^+, \quad \mathbf{V}^+_s ::= \{Q \in \mathbf{V}^+ : \mathtt{ord}(Q) = s\}, \quad s \in \mathbb{N}_0.$$

Zero sequence has zero order: $\mathtt{ord}(0^\infty) = 0$. For nonzero sequences $Q \in \mathbf{V}^+$, $\|Q\| > 0$, the order is the position of the rightmost unit: $Q_{\mathtt{ord}(Q)} = 1$, $Q_i \equiv 0$, $i > \mathtt{ord}(Q)$.

Submanifolds $\mathcal{MC}(s) \subset \mathcal{MC}$ and subspaces $\mathbf{T}(s) \subset \mathbf{T}$ are defined as follows:

$$\mathcal{MC}(s) = \{\boldsymbol{\omega} \in \mathcal{MC} : \mathtt{ord}(\boldsymbol{\omega}) \le s\}, \quad \mathbf{T}(s) = \{\boldsymbol{\omega} \in \mathbf{T} : \mathtt{ord}(\boldsymbol{\omega}) \le s\}, \quad s \in \mathbb{N}_0.$$

Space $\mathbf{T}(s)$ is the tangent space of $\mathfrak{u}$ within the manifold $\mathcal{MC}(s)$. For any $0 \le s < s'$: $\mathfrak{u} \in \mathcal{MC}(s) \subset \mathcal{MC}(s')$, $\mathbf{T}(s) \subset \mathbf{T}(s')$. Denote by $\partial\mathbf{T}(s)$, $s > 0$, an orthogonal complement of $\mathbf{T}(s-1)$ within $\mathbf{T}(s)$ w.r.t. the Fisher-Riemann dot product. For $s = 0$ set $\partial\mathbf{T}(0) = \mathbf{T}(0)$. Spaces $\mathbf{T}(s)$, $\partial\mathbf{T}(s)$ are $\boldsymbol{\chi}$-canonic and have the following dimensions: $\dim(\mathbf{T}(s)) = 2^s$, $\dim(\partial\mathbf{T}(s)) = 2^{\max\{0,s-1\}}$, $s \in \mathbb{N}_0$. Base of characters for $\partial\mathbf{T}(s)$ is $\{\boldsymbol{\chi}^{(Q)} : Q \in \mathbf{V}^+_s\}$.

Let us call random sequence $\mathbf{x} \in \mathbf{V}^{\mathbb{Z}}$: Δ-stationary, if its probability distribution is invariant under shift by $\Delta \in \mathbb{N}$, i.e., $\mathcal{L}\{(\mathbf{x}_i)\} = \mathcal{L}\{(\mathbf{x}_{i+\Delta})\}$; semi-stationary, if it is Δ-stationary for some $\Delta \in \mathbb{N}$. Usual stationarity is 1-stationarity in these terms.

The second Markov manifold that we consider is $\mathcal{MC}^\pi = \cup_{\Delta\in\mathbb{N}}\mathcal{MC}^\pi\{\Delta\}$, where the submanifold $\mathcal{MC}^\pi\{\Delta\} \subset \mathcal{MC}^\pi$ consists of distributions of Δ-periodic Δ-stationary Markov chains determined by Δ-stationarity condition and Δ-periodic generalization of Markov equation (13):

$$\mathbf{P}\{\mathbf{x}_i = 0 | \mathbf{x}_{i-j} = q_j, j \in \mathbb{N}\} = \frac{1 + \boldsymbol{\omega}_i(q)}{2}, \quad q \in \mathbf{V}^{\mathbb{N}}, \quad i \in \mathbb{Z}, \tag{15}$$

$\boldsymbol{\omega}_i : \mathbb{Z} \to \mathcal{MC}$ is Δ-periodic w.r.t. $i \in \mathbb{Z}$ $\mathcal{MC}$-valued sequence: $\boldsymbol{\omega}_i \equiv \boldsymbol{\omega}_{i+\Delta}$, $i \in \mathbb{Z}$. We identify the set of such sequences with $\mathcal{MC}^\pi\{\Delta\}$, and with $\mathcal{MC}^\pi$ for all $\Delta \in \mathbb{N}$. Denote by $\mathbf{T}^\pi$ the tangent space of $\mathfrak{u}$ within $\mathcal{MC}^\pi$. Obviously $\mathcal{MC} \subset \mathcal{MC}^\pi$ and $\mathbf{T} \subset \mathbf{T}^\pi$. The model of $\mathbf{T}^\pi$ is the set of periodic $\mathbf{T}$-valued sequences $\boldsymbol{\omega}_i : \mathbb{Z} \to \mathbf{T}$, where the subspace $\mathbf{T}^\pi\{\Delta\} \subset \mathbf{T}^\pi$ of Δ-periodic ones forms the tangent space of $\mathfrak{u}$ within $\mathcal{MC}^\pi\{\Delta\}$ (we use braces $\mathbf{T}^\pi\{\cdot\}$, $\mathcal{MC}^\pi\{\cdot\}$ to avoid confusion with the notations $\mathbf{T}(\cdot)$, $\mathcal{MC}(\cdot)$). Spaces $\{\mathbf{T}^\pi\{\Delta\}\}$ are nested by periods divisibility: $\mathbf{T}^\pi\{\Delta\} \subset \mathbf{T}^\pi\{\Delta'\}$ for

any $\Delta|\Delta'$ (similarly for manifolds $\{\mathcal{MC}^\pi\{\Delta\}\}$). In particular, $\mathbf{T}^\pi\{1\} \subset \mathbf{T}^\pi\{\Delta\}$ for any $\Delta \in \mathbb{N}$, where $\mathbf{T}^\pi\{1\}$ is the subspace of constant sequences $\boldsymbol{\omega} = (\boldsymbol{\omega}_i \equiv \mathrm{const}_i)$. This subspace implements the embedding $\mathbf{T} \hookrightarrow \mathbf{T}^\pi$, i.e., $\mathbf{T}^\pi\{1\}$ is $\mathbf{T}$ embedded into $\mathbf{T}^\pi$. Denote by $\mathtt{per}(\boldsymbol{\omega}) ::= \min\{\Delta \in \mathbb{N} : \boldsymbol{\omega} \in \mathbf{T}^\pi\{\Delta\}\}$ the minimal period of $\boldsymbol{\omega} \in \mathbf{T}^\pi$. Fisher-Riemann dot product on $\mathbf{T}^\pi$ has the form:

$$\langle \boldsymbol{\omega}, \boldsymbol{\omega}' \rangle = \Delta^{-1} \sum_{i=1}^{\Delta} \langle \boldsymbol{\omega}_i, \boldsymbol{\omega}'_i \rangle, \quad \boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathbf{T}^\pi, \quad \Delta = \mathtt{lcd}(\mathtt{per}(\boldsymbol{\omega}), \mathtt{per}(\boldsymbol{\omega}')), \tag{16}$$

where $\mathtt{lcd}(\cdot,\cdot)$ is a lowest common dominator, $\langle \boldsymbol{\omega}_i, \boldsymbol{\omega}'_i \rangle$ is the Fisher-Riemann dot product on $\mathbf{T}$.

The space $\mathbf{T}^\pi$ is representable as a tensor product

$$\mathbf{T}^\pi = \mathbf{T} \otimes \mathbf{\Pi}, \tag{17}$$

where $\mathbf{\Pi} = \cup_{\Delta\in\mathbb{N}} \mathbf{\Pi}\{\Delta\}$ is a space of two-sided periodic real-valued sequences $a_i : \mathbb{Z} \to \mathbb{R}$, $\mathbf{\Pi}\{\Delta\}$ is its subspace of Δ-periodic sequences: $a_i \equiv a_{i+\Delta}$, $i \in \mathbb{Z}$. By tensor product $U \otimes V$ of two spaces we mean the linear span of tensor product $\{u_i\} \otimes \{v_j\} = \{u_i \otimes v_j\}$ (pairwise) of their bases $\{u_i\} \subset U$, $\{v_j\} \subset V$. Tensor product $\tilde{\boldsymbol{\omega}} = \boldsymbol{\omega} \otimes a \in \mathbf{T}^\pi$ of two single elements $\boldsymbol{\omega} \in \mathbf{T}$ and $a \in \mathbf{\Pi}$ is

$$\tilde{\boldsymbol{\omega}} = (\tilde{\boldsymbol{\omega}}_i)_{i\in\mathbb{Z}}, \quad \tilde{\boldsymbol{\omega}}_i = \boldsymbol{\omega} \cdot a_i. \tag{18}$$

In these terms $\mathbf{T}^\pi\{\Delta\} = \mathbf{T} \otimes \mathbf{\Pi}\{\Delta\}$, $\Delta \in \mathbb{N}$. On the space $\mathbf{\Pi}$ a dot product similar to (16) is defined:

$$\langle a, a' \rangle = \Delta^{-1} \sum_{i=1}^{\Delta} a_i a'_i, \quad a, a' \in \mathbf{\Pi}, \quad \Delta = \mathtt{lcd}(\mathtt{per}(a), \mathtt{per}(a')), \tag{19}$$

and under tensor product (18) dot products on $\mathbf{T}$ and on $\mathbf{\Pi}$ are multiplied:

$$\langle \tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}' \rangle = \langle \boldsymbol{\omega}, \boldsymbol{\omega}' \rangle \cdot \langle a, a' \rangle, \quad \boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathbf{T}, \quad a, a' \in \mathbf{\Pi}, \quad \tilde{\boldsymbol{\omega}} = \boldsymbol{\omega} \otimes a, \quad \tilde{\boldsymbol{\omega}}' = \boldsymbol{\omega}' \otimes a'. \tag{20}$$

Denote canonic orthonormal base $\mathfrak{d}^{(\Delta)} = \{\mathfrak{d}^{(t,\Delta)}\}_{t\in\mathbb{Z}_\Delta}$ of $\mathbf{\Pi}\{\Delta\}$:

$$\mathfrak{d}^{(t,\Delta)} \in \mathbf{\Pi}\{\Delta\}, \quad \mathfrak{d}^{(t,\Delta)}_i = \Delta^{1/2} \cdot \mathbb{1}\left\{i \overset{\Delta}{\equiv} t\right\}, \quad \Delta \in \mathbb{N}, \quad t \in \mathbb{Z}_\Delta. \tag{21}$$

Note that the combined set $\mathfrak{d} = \cup_{\Delta\in\mathbb{N}} \mathfrak{d}^{(\Delta)}$ is not orthonormal, and moreover, it is linearly dependent. Tensor product of the bases $\mathfrak{d}^{(\Delta)} \otimes \boldsymbol{\chi}$ forms orthonormal base of $\mathbf{T}^\pi\{\Delta\}$, $\Delta \in \mathbb{N}$. In these tensor terms the embedding $\mathbf{T} \hookrightarrow \mathbf{T}^\pi$ is made by tensor product with the constant sequence $\mathfrak{d}^{(0,1)} \in \mathbf{\Pi}$, $\mathfrak{d}^{(0,1)}_i \equiv 1$, $i \in \mathbb{Z}$: $\mathbf{T} \mapsto \mathbf{T} \otimes \mathfrak{d}^{(0,1)} = \mathbf{T}^\pi\{1\} \subset \mathbf{T}^\pi$. Further for brevity we omit this "$\otimes\mathfrak{d}^{(0,1)}$" for subspaces $\mathcal{T} \subset \mathbf{T}$ and write $\mathcal{T} \subset \mathbf{T}^\pi$ instead of $\mathcal{T} \otimes \mathfrak{d}^{(0,1)} \subset \mathbf{T}^\pi$.

Similarly to the characters base (14) on $\mathbf{T}$, many important tangent subspaces $\mathcal{T} \subset \mathbf{T}^\pi$ are linear spans of subsets of $\mathfrak{d}^{(\Delta)} \otimes \boldsymbol{\chi}$ for some $\Delta \in \mathbb{N}$. Therefore, combination of the characters base (14) and the tensor product representation (17) provides powerful tool for working in the tangent space $\mathbf{T}^\pi$. Below we illustrate it briefly.

2.2 Family of chi-square statistics "coded" by tangent subspaces

There is a family of chi-square statistics $S_{\mathcal{T}} = S_{\mathcal{T}}[\mathbf{x}_1^n]$ bijectively "coded" by finite-dimensional tangent subspaces $\mathcal{T} \subset \mathbf{T}^\pi$. More precisely, for each $\mathcal{T}$ there is a class of asymptotically equivalent statistics as $n \to +\infty$. These equivalent statistics behave asymptotically identically for uniformly distributed binary sequence $\{\mathbf{x}_i\}$ and in some "neighborhood" of this assumption called contigual asymptotics [1, 2], when $\mathcal{MC}^\pi \ni \mathcal{L}\{\mathbf{x}\} \to \mathfrak{u}$ along the Markov manifold $\mathcal{MC}^\pi$ with a "speed" $\mathcal{O}\left(n^{-1/2}\right)$ as $n \to +\infty$. Up to this equivalence, we can talk about statistics $S_{\mathcal{T}}$ without specifying which one is it from the equivalence class.

Let us call $\mathcal{T}$-test a chi-square statistical test of null-hypothesis $H_0 : \mathcal{L}\{\mathbf{x}\} = \mathfrak{u}$ associated with statistics $S_{\mathcal{T}}$. Under null-hypothesis $S_{\mathcal{T}}$ is asymptotically chi-square distributed with $d = \dim(\mathcal{T})$ degrees of freedom: $\mathcal{L}\{S_{\mathcal{T}}|H_0\} \to \chi^2_d$. Under "contigual" alternative H_1 (i.e., satisfying contigual asymptotics) $\mathcal{L}\{S_{\mathcal{T}}|H_1\} \to \chi^2_{d,\lambda}$, where $\chi^2_{d,\lambda}$ is noncentral chi-square distribution with noncentrality parameter $\lambda \geq 0$ depending on space $\mathcal{T}$ and on deviation of alternative from null-hypothesis (distance from $\mathcal{L}\{\mathbf{x}\}$ to $\mathfrak{u}$ in information metric). Asymptotic power of $\mathcal{T}$-test at the significance level $\alpha \in (0, 1)$ equals $1 - F_{d,\lambda}(F^{-1}_{d,0}(1-\alpha))$, where $F_{d,\lambda}(\cdot)$ and $F^{-1}_{d,\lambda}(\cdot)$ are respectively CDF and QF for $\chi^2_{d,\lambda}$. Table 1 presents some tangent subspaces $\mathcal{T} \subset \mathbf{T}^\pi$ and their $\mathcal{T}$-tests with equivalent analogues.

Table 1: Tangent subspaces $\mathcal{T} \subset \mathbf{T}^\pi$ and their $\mathcal{T}$-tests

Tangent subspace $\mathcal{T} \subset \mathbf{T}^\pi$	Degrees of freedom $\dim(\mathcal{T})$	Equivalent $\mathcal{T}$-tests
$\mathbf{T}(0)$	1	Monobit NIST Serial (type I, $m = 1$)
$\partial\mathbf{T}(1)$	1	Runs NIST Serial (type II, $m = 2$)
$\mathbf{T}(s)$, $s \geq 0$	2^s	NIST Serial (type I, $m = s + 1$) NIST Approx. Entropy ($m = s + 1$)
$\partial\mathbf{T}(s)$, $s \geq 1$	2^{s-1}	NIST Serial (type II, $m = s + 1$)
$\sum_{t\in\mathbb{Z}_4} \mathbf{T}(t) \otimes \mathfrak{d}^{(t,4)}$	15	FIPS Poker

Remarkable feature of the family of chi-square statistics $\{S_{\mathcal{T}}\}_{\mathcal{T}\subset\mathbf{T}^\pi}$ is that its asymptotic probabilistic properties are "consistent" with geometric properties of corresponding tangent subspaces $\mathcal{T} \subset \mathbf{T}^\pi$. Due to this feature, obtaining asymptotic probabilistic properties of $\{S_{\mathcal{T}}\}$ comes down to calculation of geometric quantities based on Fisher-Riemann dot product or combinatorial quantities, e.g., for large orthonormal bases considered as combinatorial objects.

Introduce two operations on the subspaces $\mathcal{T}, \mathcal{T}' \subset \mathbf{T}^\pi$: $\mathcal{T}' \boxplus \mathcal{T} ::= \mathcal{T}' + \mathcal{T}$ for direct sum of orthogonal subspaces $\mathcal{T} \perp \mathcal{T}'$; $\mathcal{T}' \boxminus \mathcal{T} ::= \mathcal{T}' \cap \mathcal{T}^\perp$ for orthogonal complement of nested subspace $\mathcal{T} \subset \mathcal{T}'$ within a larger one $\mathcal{T}'$. The use of these operations implies

that the spaces are obviously satisfy the properties of orthogonality (for "$\boxplus$") or nesting (for "$\boxminus$"). For instance, $\partial\mathbf{T}(s) = \mathbf{T}(s)\boxminus\mathbf{T}(s-1)$, $s \geq 1$, and $\mathbf{T} = \lim_{s\to+\infty}\mathbf{T}(s)$ breaks into a sum:

$$\mathbf{T} = \boxplus\limits_{s\in\mathbb{N}_0}\partial\mathbf{T}(s). \tag{22}$$

Theorem 4. *The following properties hold for subspaces $\mathcal{T},\mathcal{T}' \subset \mathbf{T}^\pi$:*

- *$S_\mathcal{T}$ and $S_{\mathcal{T}'}$ are asymptotically independent iff $\mathcal{T} \perp \mathcal{T}'$;*
- *$S_{\mathcal{T}'\boxplus\mathcal{T}} = S_{\mathcal{T}'} + S_\mathcal{T}$, $S_{\mathcal{T}'\boxminus\mathcal{T}} = S_{\mathcal{T}'} - S_\mathcal{T}$ (additivity).*

Let us call subspaces $\mathcal{T},\mathcal{T}' \subset \mathbf{T}^\pi$ semi-orthogonal, if $\mathcal{T}\boxminus\mathcal{T}'' \perp \mathcal{T}'\boxminus\mathcal{T}''$, $\mathcal{T}'' = \mathcal{T}\cap\mathcal{T}'$. Nested ($\mathcal{T} \subset \mathcal{T}'$) and orthogonal ($\mathcal{T} \perp \mathcal{T}'$) spaces are the special cases.

Corollary 1. *Let $\mathcal{T},\mathcal{T}' \subset \mathbf{T}^\pi$ be subspaces of equal dimensions $\dim(\mathcal{T}) = \dim(\mathcal{T}')$. Statistics $S_\mathcal{T}$, $S_{\mathcal{T}'}$ are asymptotically semi-independent with the asymptotic correlation coefficient*

$$\lim \mathbf{cor}\{S_\mathcal{T}, S_{\mathcal{T}'}\} = \frac{\dim(\mathcal{T}\cap\mathcal{T}')}{\dim(\mathcal{T})},$$

iff their spaces $\mathcal{T}$, $\mathcal{T}'$ are semi-orthogonal.

Let $\xi^{(i)} = (\xi_j^{(i)})_{j=1}^{d_i} \sim \mathcal{N}_{d_i}(0_{d_i}, \mathrm{id}_{d_i})$, $i \in \{1,2\}$, be two standard Gaussian vectors with cross-covariations $\mathbf{cov}\left\{\xi_{j_1}^{(1)}, \xi_{j_2}^{(2)}\right\} = \mathbb{1}\{j_1 = j_2 \leq d_3\}\cdot\theta_{j_1}$, $j_i \in \{1,\ldots,d_i\}$, $\theta_1,\ldots,\theta_{d_3} \neq 0$, $d_3 \leq \min\{d_1,d_2\}$. Squared Euclidean norms of these Gaussian vectors are chi-square distributed: $\sigma_i = \|\xi^{(i)}\|^2 \sim \chi^2_{d_i}$. Joint probability distribution of $\sigma = (\sigma_1,\sigma_2) \in \mathbb{R}^2_+$ depends on d_1, d_2 and on the multiset $\Lambda = \{\lambda_j\}_{j=1}^{d_3}$, $\lambda_j = \theta_j^2$. By multiset we mean possible multiplicity of its elements, e.g., 3-multiset $\{1,1,2\}$ consists of 1 (twice) and 2 (once). Denote this bivariate chi-square distribution $\chi^{2\times2}_{d_1,d_2}(\Lambda) ::= \mathcal{L}\{\sigma\}$. The case of independence of σ_1 and σ_2 is $\Lambda = \varnothing$. The case of semi-independence with correlation coefficient $\mathbf{cor}\{\sigma_1,\sigma_2\} = d_3/d_1$ is $d_1 = d_2$ and $\Lambda = \{1,\ldots,1\}$ consisting of d_3 ones. In general case $\mathbf{cov}\{\sigma_1,\sigma_2\} = 2\sum_{\lambda\in\Lambda}\lambda$.

Let: $\mathtt{proj}(\boldsymbol{\omega}|\mathcal{T})$ be an orthogonal projection of $\boldsymbol{\omega} \in \mathbf{T}^\pi$ onto finite-dimensional tangent subspace $\mathcal{T} \subset \mathbf{T}^\pi$ w.r.t. the Fisher-Riemann dot product; $\mathtt{proj}(\mathcal{T}|\mathcal{T}')$, $\mathcal{T},\mathcal{T}' \subset \mathbf{T}^\pi$, be linear operator $\mathcal{T} \to \mathcal{T}'$ of orthogonal projection of space $\mathcal{T}$ onto $\mathcal{T}'$; $\Lambda(\mathcal{T},\mathcal{T}')$ be a multiset of nonzero squared singular numbers of projection operator $\mathtt{proj}(\mathcal{T}|\mathcal{T}')$ (it is commutative $\Lambda(\mathcal{T},\mathcal{T}') = \Lambda(\mathcal{T}',\mathcal{T})$ as $\mathtt{proj}(\mathcal{T}|\mathcal{T}')$ and $\mathtt{proj}(\mathcal{T}'|\mathcal{T})$ are adjoint operators and have the same singular numbers). Let us further call $\Lambda(\mathcal{T},\mathcal{T}')$ Jordan multiset of a pair of tangent subspaces, because Jordan principal angles [9] are closely related to it.

Corollary 2. *Let $\mathcal{T}^{(1)},\mathcal{T}^{(2)} \subset \mathbf{T}^\pi$ be arbitrary tangent subspaces of finite dimensions $d_i = \dim(\mathcal{T}^{(i)})$. Pair $(S_{\mathcal{T}^{(1)}}, S_{\mathcal{T}^{(2)}})$ is asymptotically $\chi^{2\times2}_{d_1,d_2}(\Lambda(\mathcal{T}^{(1)},\mathcal{T}^{(2)}))$-distributed, and its asymptotic covariance admits the following equivalent expressions:*

$$\lim \mathbf{cov}\{S_{\mathcal{T}^{(1)}}, S_{\mathcal{T}^{(2)}}\}$$

$$= 2 \sum_{\lambda \in \Lambda(\mathcal{T}^{(1)}, \mathcal{T}^{(2)})} \lambda = 2 \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} \left\langle \boldsymbol{\omega}^{(j_1)}, \tilde{\boldsymbol{\omega}}^{(j_2)} \right\rangle^2 = 2 \left\| \mathit{proj}\left(\mathcal{T}^{(1)} | \mathcal{T}^{(2)}\right) \right\|_{\mathrm{Fr}}^2,$$

where $\{\boldsymbol{\omega}^{(j_1)}\}_{j_1=1}^{d_1} \subset \mathcal{T}^{(1)}$, $\{\tilde{\boldsymbol{\omega}}^{(j_2)}\}_{j_2=1}^{d_2} \subset \mathcal{T}^{(2)}$ *are arbitrary orthonormal bases,* $\|\cdot\|_{\mathrm{Fr}}$ *is a Frobenius norm of linear operator.*

Consider a class of tangent subspaces $\mathcal{T} \subset \mathbf{T}^\pi$ to illustrate the above theory. Let us fix some period $\Delta \in \mathbb{N}$. Due to (22), the space $\mathbf{T}^\pi\{\Delta\} = \mathbf{T} \otimes \mathbf{\Pi}\{\Delta\}$ breaks into a sum:

$$\mathbf{T}^\pi\{\Delta\} = \boxplus_{(t,h) \in \mathbf{C}_\Delta} \mathfrak{d}^{(t,\Delta)} \otimes \partial \mathbf{T}(h), \quad \mathbf{C}_\Delta ::= \mathbb{Z}_\Delta \times \mathbb{N}_0. \tag{23}$$

We call Δ-periodic cell of spatio-temporal index (t, h) (index for brevity, with temporal part t and spatial part h) the summand subspace under "$\boxplus$" in (23), and we call Δ-periodic cellular space of profile $C \subset \mathbf{C}_\Delta$, $|C| < +\infty$ a partial sum of Δ-periodic cells with indices from the profile C:

$$\mathbf{T}^\pi\{C, \Delta\} ::= \boxplus_{(t,h) \in C} \mathfrak{d}^{(t,\Delta)} \otimes \partial \mathbf{T}(h), \quad \dim(\mathbf{T}^\pi\{C, \Delta\}) = \mathtt{dim}(C), \tag{24}$$

$$\mathtt{dim}(C) ::= \sum_{(t,h) \in C} 2^{\max\{0, h-1\}},$$

where $\mathtt{dim}(\cdot)$ is an unbounded integer-valued "dimensionality measure" on $\mathbf{C}_\Delta$.

Example 1. In particular, chi-square statistics (7), (8) correspond to cellular spaces. Let us show how to get their profiles $C \subset \mathbf{C}_\Delta$ by example of FIPS Poker test [5] that corresponds to statistics (7) with $\Delta = 4$ and $s = 0$. The expression in Table 1 for tangent subspace $\mathcal{T}_{\mathrm{Poker}} \subset \mathbf{T}^\pi$ of Poker test is reduced to the form (24) by breaking $\mathbf{T}(t)$ into sum $\boxplus_{t'=0}^{t} \partial \mathbf{T}(t')$. The expression from Table 1, in its turn, is obtained as follows. Space $\mathcal{T}_{\mathrm{Poker}}$ is a tangent space of $\mathfrak{u}$ within 15-dimensional Markov model of i.i.d. binary 4-blocks. Without loss of generality, 4-block starts at zero time: $(\mathbf{x}_0, \ldots, \mathbf{x}_3) \in \mathbf{V}^4$. Let us represent this Markov model as a 4-periodic one of the form (15). Symbol $\mathbf{x}_0$ does not depend on prehistory $\{\mathbf{x}_i : i < 0\}$, whence $\boldsymbol{\omega}_0 \in \mathbf{T}(0)$ (zero Markov order). Symbol $\mathbf{x}_1$ under its fixed prehistory $\{\mathbf{x}_i : i < 1\}$ conditionally depends only on $\mathbf{x}_0$, whence $\boldsymbol{\omega}_0 \in \mathbf{T}(1)$ (first Markov order). Similarly $\boldsymbol{\omega}_i \in \mathbf{T}(i)$, $i \in \mathbb{Z}_4$. These conditions are equivalent to representation in Table 1. So the profile of 4-periodic cellular space $\mathcal{T}_{\mathrm{Poker}}$ is $C = \{(t, h) : t \in \mathbb{Z}_4, h \leq \delta\}$, and similarly for statistics (7) with any arbitrary $\Delta \in \mathbb{N}$ and $s = 0$ ("stair" shape on Figure 3). Profiles corresponding to statistics (8) have "double-stair" shapes (see Figure 4) obtained by set substraction of two stair shapes of different "heights".

Due to Corollary 1, asymptotic semi-independence of chi-square statistics is equivalent to semi-orthogonality of their tangent subspaces. Let us fix some $\Delta \in \mathbb{N}$ and two profiles $C, C' \subset \mathbf{C}_\Delta$ of equal dimensionality measure $d = \mathtt{dim}(C) = \mathtt{dim}(C')$. Corresponding Δ-periodic cellular spaces of equal dimensions d are semi-orthogonal

by definition, whence chi-square statistics $S_{\mathbf{T}^\pi\{C,\Delta\}}$, $S_{\mathbf{T}^\pi\{C',\Delta\}}$ are asymptotically semi-independent, and finding their asymptotic correlation coefficient $\tilde{d}/d$ comes down to calculation of dimensionality measure of intersection profile:

$$\tilde{d} = \dim(\mathbf{T}^\pi\{C,\Delta\} \cap \mathbf{T}^\pi\{C',\Delta\}) = \dim(\mathbf{T}^\pi\{C \cap C',\Delta\}) = \mathtt{dim}(C \cap C').$$

Let us fix now some profile $C \subset \mathbf{C}_\Delta$ and consider corresponding chi-square statistics in two versions: undistorted $S_{\mathbf{T}^\pi\{C,\Delta\}}[\mathbf{x}_1^n]$ and δ-distorted $S_{\mathbf{T}^\pi\{C,\Delta\}}[\mathbf{x}_{1+\delta}^n]$ (see Subsection 1.3). Reasoning similarly to Example 1, it is easy to show that δ-distortion leads to cyclic δ-shifting of profile along the temporal axis:

$$S_{\mathbf{T}^\pi\{C,\Delta\}}[\mathbf{x}_{1+\delta}^n] = S_{\mathbf{T}^\pi\{C',\Delta\}}[\mathbf{x}_1^n], \quad C' = \mathtt{R}^\delta C,$$

where $\mathtt{R}^\delta : \mathbf{C}_\Delta \to \mathbf{C}_\Delta$, $\mathtt{R}^\delta(t,h) ::= (t \overset{\Delta}{+} \delta, h)$, $(t,h) \in \mathbf{C}_\Delta$. Obviously dimensionality measure is invariant under this cyclic shift $\mathtt{dim}(C) \equiv \mathtt{dim}(\mathtt{R}^\delta C)$ as its weight $2^{\max\{0,h-1\}}$ depends only on the spatial part h of index (t,h). So finally, for chi-square statistics corresponding to Δ-periodic cellular spaces the asymptotic correlation coefficient between the undistorted and δ-distorted versions admits the following expression:

$$\lim \mathbf{cor}\left\{S_{\mathcal{T}}[\mathbf{x}_1^n], S_{\mathcal{T}}[\mathbf{x}_{1+\delta}^n]\right\} = \frac{\mathtt{dim}(C \cap \mathtt{R}^\delta C)}{\mathtt{dim}(C)}, \quad \mathcal{T} = \mathbf{T}^\pi\{C,\Delta\}. \tag{25}$$

The values (11), (12) in Theorem 3 are obtained from (25) for stair and double-stair profiles $C \subset \mathbf{C}_\Delta$ (see Example 1). For these two types of profiles Figures 3–4 illustrate components of the right-hand side fraction (25): original profile C, cyclically shifted one $\mathtt{R}^\delta C$, and their intersection $C \cap \mathtt{R}^\delta C$. Weights of dimensionality measure are plotted on the vertical axis: their sum over profile (black cells) is equal to dimensionality of corresponding cellular space.

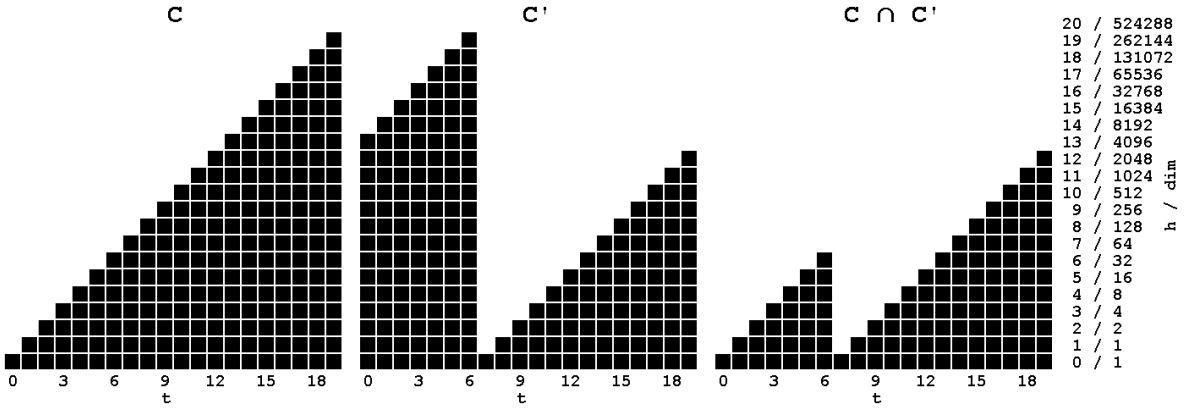


Figure 3: Illustration to (25) for "stair" profile C and its cyclic shift C': $\Delta = 20$, $\delta = 7$

The result of Theorem 2 concerns joint asymptotic distribution of a pair of chi-square statistics $(S_{\mathbf{T}^\pi\{C,\Delta\}}, S_{\mathbf{T}^\pi\{C',\Delta'\}})$ corresponding to cellular spaces of two arbitrary periods $\Delta, \Delta' \in \mathbb{N}$. In the case $\Delta \neq \Delta'$ there is no more guaranteed semi-orthogonality of cellular spaces, and as a consequence, asymptotic semi-independence of chi-square

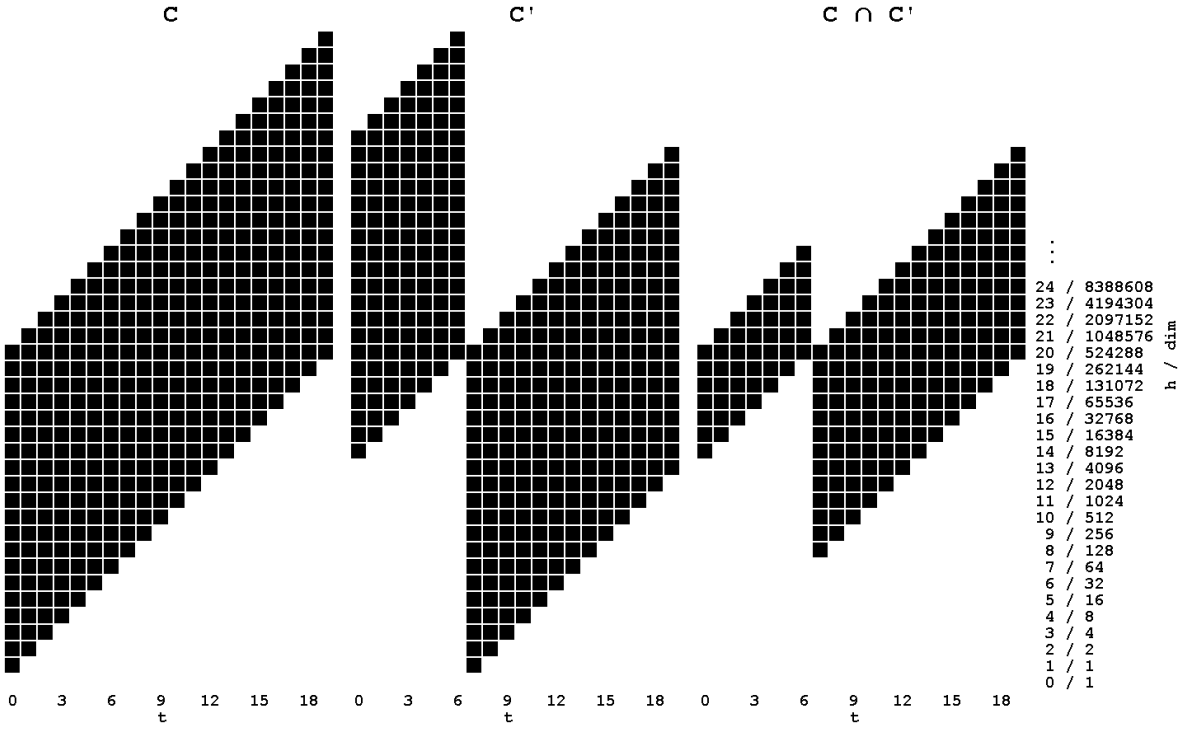


Figure 4: Illustration to (25) for "double-stair" profile C and its cyclic shift C': $\Delta = 20$, $\delta = 7$

statistics. Therefore, we need more general Corollary 2 to describe joint asymptotic distribution of chi-square statistics. The following trivial factorization Lemma is also used.

Lemma 1. *Let two sums $\mathcal{T}^{(i)} = \boxplus_{j \in J^{(i)}} \mathcal{T}^{(i,j)}$, $i \in \{1,2\}$, of tangent subspaces satisfy "diagonality" property: $\mathcal{T}^{(1,j)} \perp \mathcal{T}^{(2,j')}$ for any $j \neq j'$. Then their Jordan multiset breaks into union:*

$$\Lambda(\mathcal{T}^{(1)}, \mathcal{T}^{(2)}) = \bigcup_{j \in J^{(1)} \cap J^{(2)}} \Lambda(\mathcal{T}^{(1,j)}, \mathcal{T}^{(2,j)}).$$

Let us break $2^{\max\{0,h-1\}}$-dimensional Δ-periodic cell of index $(t,h) \in \mathbf{C}_\Delta$ into a sum of one-dimensional orthogonal subspaces:

$$\mathfrak{d}^{(t,\Delta)} \otimes \partial\mathbf{T}(h) = \boxplus_{Q \in \mathbf{V}_h^+} \mathtt{span}\left\{\mathfrak{d}^{(t,\Delta)} \otimes \boldsymbol{\chi}^{(Q)}\right\}. \tag{26}$$

By analogy with Δ-periodic cells, we call Δ-periodic atom of spatio-temporal index (index for brevity) $(t,Q) \in \mathbf{A}_\Delta ::= \mathbb{Z}_\Delta \times \mathbf{V}^+$ the summand subspace under "$\boxplus$" in (26): Q is a spatial part of atom's index. Rewrite cellular space (24) as a sum of atoms, grouping summands by spatial parts Q:

$$\mathbf{T}^\pi\{C,\Delta\} = \boxplus_{(t,h) \in C} \boxplus_{Q \in \mathbf{V}_h^+} \mathtt{span}\left\{\mathfrak{d}^{(t,\Delta)} \otimes \boldsymbol{\chi}^{(Q)}\right\} = \boxplus_{h \in \mathbf{h}^{(C)}} \boxplus_{Q \in \mathbf{V}_h^+} \boldsymbol{\chi}^{(Q)} \otimes \boldsymbol{\Pi}\{\mathbf{t}_h^{(C)}, \Delta\}, \tag{27}$$

$$\boldsymbol{\Pi}\{\mathbf{t},\Delta\} ::= \mathtt{span}\left\{\mathfrak{d}^{(t,\Delta)} : t \in \mathbf{t}\right\} \subset \boldsymbol{\Pi}\{\Delta\},$$

$$\mathbf{t}_h^{(C)} ::= \{t : (t,h) \in C\} \subset \mathbb{Z}_\Delta, \quad \mathbf{h}^{(C)} ::= \{h : \exists (t,h) \in C\} \subset \mathbb{N}_0.$$

Here $\mathbf{t}_h^{(C)}$ is a "temporal slice" of profile C at the spacial level h, $\mathbf{h}^{(C)}$ is a "spatial projection" of profile C, $\mathbf{\Pi}\{\mathbf{t},\Delta\}$ is a canonic subspace of $\mathbf{\Pi}\{\Delta\}$ corresponding to "temporal profile" $\mathbf{t}$. Consider now a similar sum (27) for another one cellular space $\mathbf{T}^\pi\{C',\Delta'\}$. Diagonality property of Lemma 1 holds for these sums over Q: it follows from (20) and characters orthogonality. Hence:

$$\begin{aligned}
&\Lambda(\mathbf{T}^\pi\{C,\Delta\},\mathbf{T}^\pi\{C',\Delta'\}) \\
&\quad = \bigcup_{h\in\mathbf{h}^{(C)}\cap\mathbf{h}^{(C')},Q\in\mathbf{V}_h^+} \Lambda\left(\boldsymbol{\chi}^{(Q)}\otimes\mathbf{\Pi}\{\mathbf{t}_h^{(C)},\Delta\},\boldsymbol{\chi}^{(Q)}\otimes\mathbf{\Pi}\{\mathbf{t}_h^{(C')},\Delta'\}\right) \\
&\quad = \bigcup_{h\in\mathbf{h}^{(C)}\cap\mathbf{h}^{(C')}} \Lambda^{\times 2^{\max\{0,h-1\}}}\left(\mathbf{\Pi}\{\mathbf{t}_h^{(C)},\Delta\},\mathbf{\Pi}\{\mathbf{t}_h^{(C')},\Delta'\}\right), \quad (28)
\end{aligned}$$

where "$\Lambda^{\times k}$" means multiset Λ taken with multiplicity k. Summation of elements (squared singular numbers) over Jordan multiset (28) for stair and double-stair profiles C, C' (see Example 1) leads to expression (10).

For Jordan multiset of the form $\Lambda\left(\mathbf{\Pi}\{\mathbf{t},\Delta\},\mathbf{\Pi}\{\mathbf{t}',\Delta'\}\right)$ in the final part of (28) the sum of its elements, and moreover, these elements themselves can be found explicitly. From (19), (21) in notation of Theorem 2:

$$\left\langle \mathfrak{d}^{(t,\Delta)},\mathfrak{d}^{(t',\Delta')}\right\rangle = \mathbb{1}\left\{t \overset{\kappa}{\equiv} t'\right\}\cdot(aa')^{-1/2}, \quad t\in\mathbb{Z}_\Delta, \quad t'\in\mathbb{Z}_{\Delta'}, \qquad (29)$$

i.e., $\left\langle \mathfrak{d}^{(t,\Delta)},\mathfrak{d}^{(t',\Delta')}\right\rangle \equiv 0$ for $t \overset{\kappa}{\not\equiv} t'$. Let us break $\mathbf{\Pi}\{\mathbf{t},\Delta\}$ into a sum over integers modulo κ: $\mathbf{\Pi}\{\mathbf{t},\Delta\} = \boxplus_{\tau\in\mathbf{t}/\kappa\mathbb{Z}}\mathbf{\Pi}\{\mathbf{t}_\tau,\Delta\}$, $\mathbf{t}_\tau = \{t\in\mathbf{t} : t\overset{\kappa}{\equiv}\tau\}$, and similarly for $\mathbf{\Pi}\{\mathbf{t}',\Delta'\}$. These sums satisfy diagonality property of Lemma 1, whence:

$$\Lambda\left(\mathbf{\Pi}\{\mathbf{t},\Delta\},\mathbf{\Pi}\{\mathbf{t}',\Delta'\}\right) = \bigcup_{\tau\in(\mathbf{t}/\kappa\mathbb{Z})\cap(\mathbf{t}'/\kappa\mathbb{Z})} \Lambda\left(\mathbf{\Pi}\{\mathbf{t}_\tau,\Delta\},\mathbf{\Pi}\{\mathbf{t}'_\tau,\Delta'\}\right). \qquad (30)$$

Due to (29), projection operator $\mathtt{proj}\left(\mathbf{\Pi}\{\mathbf{t}_\tau,\Delta\}|\mathbf{\Pi}\{\mathbf{t}'_\tau,\Delta'\}\right)$ is represented by a matrix $|\mathbf{t}_\tau|\times|\mathbf{t}'_\tau|$ of all entries equal to $(aa')^{-1/2}$ having a single nonzero singular number $\sqrt{\frac{|\mathbf{t}_\tau|}{a}\cdot\frac{|\mathbf{t}'_\tau|}{a'}}$, whence (30) takes the form:

$$\Lambda\left(\mathbf{\Pi}\{\mathbf{t},\Delta\},\mathbf{\Pi}\{\mathbf{t}',\Delta'\}\right) = \bigcup_{\tau\in(\mathbf{t}/\kappa\mathbb{Z})\cap(\mathbf{t}'/\kappa\mathbb{Z})} \Lambda\left(\mathbf{\Pi}\{\mathbf{t}_\tau,\Delta\},\mathbf{\Pi}\{\mathbf{t}'_\tau,\Delta'\}\right). \qquad (31)$$

$$\Lambda\left(\mathbf{\Pi}\{\mathbf{t},\Delta\},\mathbf{\Pi}\{\mathbf{t}',\Delta'\}\right) = \left\{\frac{|\mathbf{t}_\tau|}{a}\cdot\frac{|\mathbf{t}'_\tau|}{a'} : \tau\in(\mathbf{t}/\kappa\mathbb{Z})\cap(\mathbf{t}'/\kappa\mathbb{Z})\subset\mathbb{Z}_\kappa\right\}.$$

This means, in particular, that Jordan multiset of any two cellular spaces of periods $\Delta,\Delta'\in\mathbb{N}$ contains only rational numbers of the form $\frac{b}{a}\cdot\frac{b'}{a'}$, $1\le b\le a$, $1\le b'\le a'$, where $\frac{a}{a'}$ is an irreducible fraction equals to $\frac{\Delta}{\Delta'}$. From (28), (31) we also obtain the following semi-orthogonality criterion for arbitrary cellular spaces (i.e., criterion for Jordan multiset of all units).

Theorem 5. *In the above and Theorem 2 notations cellular spaces* $\mathbf{T}^{\pi}\{C, \Delta\}$ *and* $\mathbf{T}^{\pi}\{C', \Delta'\}$ *are semi-orthogonal iff for each simultaneously nonempty equivalence classes* $\mathbf{t}_{h,\tau}^{(C)}, \mathbf{t}_{h,\tau}^{(C')} \neq \varnothing$, $h \in \mathbf{h}^{(C)} \cap \mathbf{h}^{(C')}$, $\tau \in (\mathbf{t}_h^{(C)}/\kappa\mathbb{Z}) \cap (\mathbf{t}_h^{(C')}/\kappa\mathbb{Z})$, *these classes are both maximal:* $\left|\mathbf{t}_{h,\tau}^{(C)}\right| = a$, $\left|\mathbf{t}_{h,\tau}^{(C')}\right| = a'$.

This criterion obviously holds for $\Delta = \Delta'$ as $a = a' = 1$ and nonemptiness is equivalent to maximality.

References

1. Voloshko, V.A., Kharin, Yu.S., Trubey, A.I. (2022). On power comparison for some tests on pure randomness under Markov high-order dependencies. *Proc. XIII Int. Conf. "Computer Data Analysis and Modeling"*. P. 211–217.

2. Voloshko, V.A. (2023). On asymptotic properties for a family of χ^2-tests of pure randomness of binary sequence. *Proc. II Int. Conf. "Theoretical and Applied Cryptography"*. P. 15–43. (In Russian)

3. Kharin, Yu.S., Petlitskii, A.I. (2007). A Markov chain of order s with r partial connections and statistical inference on its parameters. *Discrete Mathematics and Applications*. Vol. **17**, No. **3**. P. 295–317.

4. Maltsew, M.V., Kharin, Yu.S. (2023). On statistical estimation of multivariate entropy for quality testing of cryptographic generators. *Proc. II Int. Conf. "Theoretical and Applied Cryptography"*. P. 140–147. (In Russian)

5. NIST (2001). *Security requirements for cryptographic modules*. FIPS PUB 140-2.

6. Rukhin, A. [et. al.] (2010). *A statistical test suite for random and pseudorandom number generators for cryptographic applications*. NIST SP 800-22 Rev. 1a.

7. Amari, S., Nagaoka, H. (2000). *Methods of information geometry*. Oxford University Press.

8. Luong, B. (2009). *Fourier analysis on finite Abelian groups*. Birkhauser: Boston.

9. Jordan, C. (1875). Essai sur la geometrie a n dimensions. *Buletin de la Societe Mathematique de France*. Vol. **3**. P. 103–174.

PARAMETER ESTIMATION FOR MMPP WITH TWO STATES OF THE CONTROLLING MARKOV CHAIN

S.E. Vorobejchikov[1], Yu.B. Burkatovskaya[2]

[1,2] *Tomsk state University*
[2] *Tomsk Polytechnic University*
Tomsk, RUSSIA
e-mail: [1]sev@mail.tsu.ru, [2]tracey@tpu.ru

The paper considers the problem of estimating parameters of MMPP with two regimes. A combined estimator based on the method of moments and the maximum lekelihood method is proposed. A cumulative sum algorithm is used to detect changes in the control Markov chain parameters. The proposed method has a low computational complexity and sufficient accuracy of the results.

Keywords: Markov modulated Poisson process, moment estimation, maximum likelihood estimation

1 Introduction

Stochastic processes with changing parameters, in particular, processes with change points, are widely used to approximate real nonlinear processes in different applications. From a statistical point of view, a change point is a place or time point such that the observations follow one distribution up to that point and another distribution after that point. Multiple change points problems can also be defined similarly. Processes with a single change point are usually used to simulate a withdrawal from stable conditions or an equipment error; whereas processes with multiple change points describe real processes with several states, where switches between states occur in unknown instants.

In queue theory, processes with multiple change points are applied to describe queue systems with different regimes of customer arrivals; as the simplest model, a process with two states corresponding to "usual" and "peak" time can be considered. These two states are characterized of different rates of arrivals; in the first case, events occur in general much less than in the second. When the switching times are controlled by a Markov chain and the flow of events have Poisson distribution, the model is named Markov-modulated Poisson process (MMPP).

2 Problem Statement

The Markov-Modulated Poisson Process (MMPP) is a doubly stochastic Poisson process whose rate varies according to a Markov chain. We consider a process with two rates determined by the states of a non-observable controlling Markov chain. The sojourn time in state $\{j\}$, $j \in 1, 2$, is determined by exponentially distributed random variables x_i, $i = 1, 2, ...$ with parameters μ_j, where x_i for odd i and for even i are distributed with different parameters.

Denote $x_1 + ... + x_m$ as T_m and $T_0 = 0$. At the interval $[T_{i-1}, T_i)$ one can observe a Poisson process with rate λ_j, where j is the state of the controlling Markov chain at the interval. The instants T_i can be considered as the change points of the observed Poisson process, where the rate of the process changes. We denote the arrival times of the observed process as t_k, $k = 1, 2, ...$, $t_0 = 0$.

The problem is to estimate the parameters $\{\lambda_1, \lambda_2, \mu_1, \mu_2\}$ by the observations t_k.

3 Parameter Estimation

Consider the process $\{\tau_i\}_{i\geq 1}$, where $\tau_i = t_i - t_{i-1}$ is the length of the i-th interval between arriving events in the observed flow. The values τ_i for MMPP are independent and exponentially distributed with one of two possible parameters. It allows us to use the hyperexponential distribution as a model for the observed data. The hyperexponential distribution has the following density function

$$f(\tau) = p\lambda_1 e^{-\lambda_1 \tau} + (1-p)\lambda_2 e^{-\lambda_2 \tau}, \quad p \in (0,1), \tag{1}$$

which is a mixture of two exponential distributions.

To estimate the parameters λ_1, λ_2 and p of the hyperexponential distribution, we propose a combined estimation on the basis of the method of moments and the maximum likelihood method. Let us have a sample of the $\{\tau_1, ..., \tau_N\}$ obeyed distribution (1). First we calculate two first moments

$$m_1 = \frac{1}{N}\sum_{k=1}^{N} \tau_k, \quad m_2 = \frac{1}{N}\sum_{k=1}^{N} \tau_k^2, \tag{2}$$

and equating them to the corresponding theoretical moments, one obtains the equations

$$m_1 = \frac{p}{\lambda_1} + \frac{1-p}{\lambda_2}, \quad m_2 = \frac{2p}{\lambda_1^2} + \frac{2(1-p)}{\lambda_2^2} \tag{3}$$

Introducing the following notations $x = \frac{p}{\lambda_1}$, $y = \frac{1-p}{\lambda_2}$ one obtains the quadratic equation

$$x^2 - 2m_1 p x + p\left(m_1^2 - \frac{1-p}{2} m_2\right) = 0.$$

If the inequality $m_2 - 2m_1^2 \geq 0$ holds true then the equation has two roots. Taking the largest one as x one can obtain the values λ_1, λ_2 as functions of the parameter p

$$\hat{\lambda}_1 = \frac{1}{m_1 + \sqrt{\frac{1-p}{p}}\sqrt{\frac{m_2}{2} - m_1^2}}, \quad \hat{\lambda}_2 = \frac{1}{m_1 - \sqrt{\frac{p}{1-p}}\sqrt{\frac{m_2}{2} - m_1^2}}. \tag{4}$$

It should be noted that $\hat{\lambda}_1 < \hat{\lambda}_2$ because we choose the largest root of the quadratic equation for x. If one takes the smallest root for x, he obtains the same estimators for λ_1, λ_2, except of $\hat{\lambda}_2 < \hat{\lambda}_1$.

As both roots should be positive, the parameter p should obey the inequality:

$$m_1 - \sqrt{\frac{p}{1-p}}\sqrt{\frac{m_2}{2} - m_1^2} > 0.$$

From the equation, one obtains

$$p < \frac{2m_1^2}{m_2}. \tag{5}$$

Next we define the log-likelihood function

$$L(p) = \sum_{n=1}^{N} \log\left(p\lambda_1 e^{-\lambda_1\tau_n} + (1-p)\lambda_2 e^{-\lambda_2\tau_n}\right), \tag{6}$$

where the parameters λ_k are calculated according (4). Maximizing the function $L(p)$ on p, where $0 < p < \frac{2m_1^2}{m_2}$ and using relations (4) one obtains the estimators of the parameters $\lambda_k,\ k = 1, 2$.

Value $m_2 - 2m_1^2$ where m_1, m_2 are defined equation (3) satisfy the following properties when $N \to \infty$

$$E\left(m_2 - 2m_1^2\right) \to 2p(1-p)\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)^2. \tag{7}$$

The variance of $m_2 - 2m_1^2$ has the following form, where C is a constant depending on the process parameters:

$$Var\left(m_2 - 2m_1^2\right) = \frac{C}{N} + O\left(\frac{1}{N^2}\right). \tag{8}$$

4 Change-point Detection Algorithm

Previously, we developed a cumulative sum algorithm for the detection of changes in the Poisson process when the parameters λ_j are known. Define statistics z_i as the logarithm of the likelihood ratio for exponential distribution

$$z_i = \ln\left(\frac{\lambda_2 \exp(-\lambda_2\tau_i)}{\lambda_1 \exp(-\lambda_1\tau_i)}\right) = \ln\frac{\lambda_2}{\lambda_1} - (\lambda_2 - \lambda_1)\tau_i. \tag{9}$$

Let $\lambda(t)$ be the rate of the arrival process at the moment t. Introduce two hypothesis:

$$H_j = \{\lambda(t) = \lambda_j,\ j = 1, 2\}. \tag{10}$$

Suppose $\lambda_1 < \lambda_2$ and denote λ_2/λ_1 by δ, where $\delta > 1$. The statistics z_i have the following form:

$$z_i = \ln\delta - \lambda_1(\delta - 1)\tau_i. \tag{11}$$

Under the hypothesis H_j, the statistics have the following properties:

$$\begin{aligned} E\left[z_i \middle| H_1\right] &= \ln\delta - (\delta - 1) < 0; \\ E\left[z_i \middle| H_2\right] &= \ln\delta - \left(\frac{\delta - 1}{\delta}\right) > 0, \end{aligned} \tag{12}$$

hence, the mean of statistics z_i changes the sign with the change the rate of the process.

In our algorithm, we use estimators (4) obtained above instead of the parameter values. We introduce positive values $h^{(1)}$, $h^{(2)}$ as the algorithms thresholds and construct the cumulative sum S_i which is recalculated at the instants of arrivals:

$$\begin{gathered} S_0 = 0; \\ S_i = \begin{cases} \max\{0, S_{i-1} + c_{i-1}z_i\}, & \text{if} \quad S_{i-1} + c_{i-1}z_i < h_i, \quad i \geq 1; \\ 0, & \text{if} \quad S_{i-1} + c_{i-1}z_i \geq h_i, \quad i \geq 1; \end{cases} \\ z_i = \ln \frac{\hat{\lambda}_2}{\hat{\lambda}_1} - (\hat{\lambda}_2 - \hat{\lambda}_1)\tau_i; \\ c_0 = 1; \\ c_i = \begin{cases} c_{i-1}, & \text{if} \quad S_{i-1} + c_{i-1}z_i < h_i, \quad i \geq 1; \\ -c_{i-1}, & \text{if} \quad S_{i-1} + c_{i-1}z_i \geq h_i, \quad i \geq 1; \end{cases} \\ h_i = \begin{cases} h^{(1)}, & \text{if} \quad c_{i-1} = 1; \\ h^{(2)}, & \text{if} \quad c_{i-1} = -1. \end{cases} \end{gathered} \tag{13}$$

If $c_i = 1$, the sum detects the increase in the rate, i.e. the change in hypothesis from H_1 to H_2 when it reaches the threshold h_1; if $c_i = -1$, the sum detects the decrease in the rate when it reaches the threshold h_2.

Let the sequence $\{\sigma_m\}_{m\geq 0}$ be the sequence of the instants when the cumulative sum reaches the threshold h, i.e.

$$\begin{gathered} \sigma_0 = 0; \\ \sigma_m = \min\left\{t_i > \sigma_{m-1} : S_i \geq h\right\}, \quad m \geq 1. \end{gathered} \tag{14}$$

Consider a sequence $\{n_m\}_{m\geq 0}$ associated with the sequence $\{\sigma_m\}_{m\geq 0}$ as follows

$$\begin{gathered} n_0 = 0; \\ n_m = \max\left\{t_i \leq \sigma_m : S_i > 0, S_{i-1} = 0\right\}, \quad m \geq 1. \end{gathered} \tag{15}$$

Thus, the instant n_m is the first instant when the cumulative sum becomes positive to reach the threshold.

The algorithm for change-point detection is described as follows. Calculate the cumulative sum given by equation (13). Then construct the sequences $\{\sigma_m\}$, $\{n_m\}$ defined by equations (14), (15). If for the last instant n_m one has $m = 2l + 1$, one sets $n_{2l+2} = t_N$; if $m = 2l$, one sets $n_{2l+1} = t_N$. Here, the odd instants n_{2l+1} are the estimators of the instants when the rate changes from λ_1 to λ_2, and the even instants n_{2l+2} are the estimators of the instants when the rate changes from λ_2 to λ_1. So, at intervals $[n_{2l}, n_{2l+1}]$ we consider the Markov chain to be in the first state; when at intervals $[n_{2l+1}, n_{2l+2}]$ we consider it to be in the second state.

STATISTICAL PROBLEMS AT COMPUTING ONLINE CONTROLLED EXPERIMENTS AT SCALE

G.V. Zasko[1], V.V. Kharlamov[2], R.K. Filev[3]

[1,2,3] *T-Bank, Applied Statistics Laboratory*
Moscow, RUSSIA
e-mail: [1]zasko.gr@bk.ru, [2]vi.v.kharlamov@gmail.com, [3]romanfilev@gmail.com

Online controlled experiments, or A/B tests, are a reliable tool for data-driven decision making in industrial applications. We discuss several applied statistical problems related to the A/B testing at scale. The list of problems includes detecting inherent biases in samples, reducing metric variance to accelerate A/B tests, and improving statistical power by multivariate hypothesis testing. For each problem, we propose a practical solution and show the results in numerical experiments.

Keywords: A/B tests, data-driven decision making, variance reduction, multivariate hypothesis testing

1 Introduction

Online controlled experiments, or A/B tests, are the most reliable way to assess the impact of product changes and make data-driven business decisions [5]. In an A/B test, customers serve as a randomization unit, and traffic is randomly assigned with a given allocation ratio between a control variation and one or more treatment variations. For each randomization unit, the values for a list of business metrics are computed; and these values are treated as random variables. Statistical hypotheses, typically about the equality of means, are then tested for the collected samples.

Industrial A/B tests have unique features. The first one is scale: tens of thousands of tests per year need to be analyzed, with the sample size ranging from thousands to millions of units per test, and hundreds of metrics computed in each test. Second, the samples accumulate incrementally as customers interact with a given splitting scheme. As a result, the sample size is not known in advance, and the experimental design must include both a power analysis to determine the required sample size and an estimate of how many days the experiment should last.

In the present report, we discuss statistical tools, developed at T-Bank, for A/B testing at scale. In particular, we focus on several statistical problems closely related to this issue.

2 Considered problems

Assume that customers are randomly assigned to one of two variations and a product change with potential business impact is introduced. Suppose that this change occurs at the specific time t and affects only one of the variations, hereafter referred to as

the test variation; while the other variation remains unaffected and is referred to as the control variation. Let T be a binary indicator variable such that $T = 0$ denotes assignment to the control variation and $T = 1$ denotes assignment to the test variation.

For each $k = 1, \ldots, M$, denote by Y_k^+ the k-th business metric computed over a time interval following the time t, and by Y_k^- the same metric computed over a time interval preceding the time t. For each k, we test the hypothesis

$$H_{0,k} : \mathsf{E}[Y_k^+ \mid T = 1] = \mathsf{E}[Y_k^+ \mid T = 0]$$

against the two-sided alternative

$$H_{1,k} : \mathsf{E}[Y_k^+ \mid T = 1] \neq \mathsf{E}[Y_k^+ \mid T = 0].$$

Rejection of $H_{0,k}$ provides evidence that the product change has a measurable impact on the k-th business metric.

2.1 Testing whether the sample is not inherently biased

However, mainly due to issues with the randomization process, the business metrics could differ significantly between the variations even before the experiment begins. Such experiments are called inherently biased, and valid statistical inferences cannot be made from their results [2]. More precisely, if these biased experiments remain undetected, they can lead to wrong business decisions.

We are able to formalize this applied business problem as a hypothesis testing problem as follows. To detect biased experiments, we test the following set of hypothesis

$$\hat{H}_{0,k} : \mathcal{L}[Y_k^- \mid T = 1] = \mathcal{L}[Y_k^- \mid T = 0]$$

against the corresponding two-sided alternatives, where $\mathcal{L}[Y_k^- \mid T = i]$ is the distribution law of Y_k^- conditional on $T = i$.

To test $\hat{H}_{0,k}$ at given k, we develop an effective combined criterion that merges the advantages of the Anderson–Darling test and the chi-squared test. The proposed criterion controls the Type I error rate and exhibits high power for both discrete and continuous metrics.

2.2 Variance reduction techniques

By accelerating experiments, one can assess more product changes in the same amount of time. Note that the main bottleneck is the time required to collect samples, not the time required to compute business metrics. Therefore, the most effective way to accelerate A/B testing is to reduce the required sample size [5]. This, in turn, can be achieved by reducing the variance of the target business metrics.

Among variance reduction techniques, we focus on a method that shows good performance according to our numerical experiments. Specifically, let Y^+ be the target

metric, and let X_l^- for $l = 1, \ldots, n$ be a set of other metrics, which are called covariates. The superscripts $+$ and $-$ keep the notation introduced above, indicating the post- and pre-experimental periods, respectively. Consider the linear transformation

$$\hat{Y}^+ = Y^+ - \sum_{l=1}^{n} \theta_l X_l^-,$$

where $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$. It is easy to show that

$$\mathsf{D}[\hat{Y}^+] \geq (1 - \rho^2)\mathsf{D}[Y^+],$$

where $\rho^2 = \left(\mathrm{y}^T \mathrm{S}^{-1} \mathrm{y}\right) / \mathsf{D}[Y^+]$, and $\mathrm{S} = [s_{ij}]$ is the symmetric positive definite matrix with the entries $s_{ij} = cov(X_i^-, X_j^-)$, and y is the n-component column with the entries $y_j = cov(Y^+, X_j^-)$. Moreover, the equality holds if and only if $\mathrm{S}\theta = \mathrm{y}$. Thus, by choosing $\theta = \mathrm{S}^{-1}\mathrm{y}$, one achieves a variance reduction factor of $(1 - \rho^2)$ by leveraging the pre-experimental data. Note that we are able to choose the set of covariates so that the matrix S is positive definite and hence non-singular.

This variance reduction technique is justified in the context of A/B experiments if the hypothesis testing for the transformed metric $\hat{Y}^+$ is equivalent to that for the original metric Y^+. This equivalence holds under the condition

$$\mathsf{E}\left[\sum_{l=1}^{n} \theta_l X_l^- \,|\, T = 1\right] = \mathsf{E}\left[\sum_{l=1}^{n} \theta_l X_l^- \,|\, T = 0\right]$$

that is satisfied by unbiased experiments (see Section 2.1).

Note that the proposed method generalizes the well-known CUPED approach [1, 6] *Controlled Experiments Using Pre-Experiment Data* which employs only a single covariate. Likewise, we refer to the proposed method as Multi-CUPED.

2.3 Testing the multivariate hypothesis

In industrial A/B tests, hundreds of business metrics are computed for each test. Consequently, one tests the family of null hypothesis $H_{0,k}$ for $k = 1, \ldots, M$, where M can be very large. Testing multiple hypotheses reduces statistical power due to the need for Bonferroni-type corrections or other adjustments to control the family-wise error rate.

However, in certain business scenarios one can test the hypothesis

$$H_0 : \mathsf{E}[Y_k^+ \,|\, T = 1] = \mathsf{E}[Y_k^+ \,|\, T = 0] \,\text{for all}\, k = 1, \ldots, M$$

against the alternative

$$H_1 : \exists k_* : \mathsf{E}[Y_{k_*}^+ \,|\, T = 1] \neq \mathsf{E}[Y_{k_*}^+ \,|\, T = 0].$$

This problem can be equivalently formulated as testing equality of mean vectors for the multivariate random variable $\mathrm{Y} = (Y_1^+, \ldots Y_M^+)$. This framework is known as multivariate hypothesis testing [3]. The statistical tests for the multivariate hypothesis

deal with multivariate random variables and account for pairwise correlations through the covariance matrices. These tests, for example, the Hotelling T-squared test, often assume that the covariance matrices are equal among the variations, while this assumption does not hold in applications.

To test H_0, we implement a criterion based on the work [4]. We show that the proposed criterion controls the Type I error rate in all scenarios, including those with unequal sample sizes, and those with unequal covariance matrices between variations. Using numerical experiments with both real and synthetic data, we demonstrate when the proposed criterion outperforms metric-wise T-tests with corrections for the multiple testing.

3 Conclusion

In this report, we demonstrate how business-motivated problems arising in industrial A/B testing can be effectively formalized as precise mathematical problems and subsequently solved with statistical methods. Several statistical problems are discussed, including detecting inherent biases in samples, reducing metric variance to accelerate A/B tests, and improving statistical power by multivariate hypothesis testing. For each problem, we propose a practical solution and show the results in numerical experiments.

References

1. Deng A., Xu Y., Kohavi R., and Walker T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *The 6-th ACM International Conference on Web Search and Data Mining*. pp. 123-132.

2. Fabijan A., Dmitriev P., Holmstrm Olsson H., Bosch J., Vermeer L., Lewis D. (2019) Three Key Checklists and Remedies for Trustworthy Analysis of Online Controlled Experiments at Scale. *IEEE/ACM 41-st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. pp. 1-10.

3. Johnson R.A., and Wickern D.W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.

4. Kim S.J. (1992) A practical solution to the multivariate Behrens-Fisher problem. *Biometrika*. Vol. **79**, Num. **1**, pp. 171-176.

5. Kohavi R., Tang D., Xu Y. (2020). *Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press.

6. Xie H., and Aurisset J. (2016). Improving the sensitivity of online controlled experiments: Case studies at NETFLIX. *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 645-654.

THE LIGHTWEIGHT PARTS DESIGN PROCESSES IMPROVEMENT BASED ON MODELING STRUCTURES WITH CYLINDRICAL CELLS[1]

B.A. Zhalezka[1], K.S. Korolenok[2], I.A. Shamardzina[3]

[1,2,3]*Belarusian National Technical University*
Minsk, BELARUS
e-mail: [1]boriszh@yandex.by, [2]kskorolenok@mail.ru, [3]shamardina.i@bntu.by

The work is focused on studying methods and models that describe the patterns of design and production of lightweight parts. Additionally, it involves constructing a functional and structural model for the hardware implementation of graphical construction of parts and cells within them. One of the key challenges in manufacturing lightweight parts with a cellular structure is ensuring the required technical and operational properties. To address this issue, it is proposed to utilize the capabilities of engineering stress-strain state analysis of the resulting parts using SolidWorks API methods.

Keywords: additive manufacturing, material consumption, cellular structure, lightweight parts, CAD

1 Introduction

In the context of the digital economy, integrating information technologies into all aspects of life is becoming increasingly important, particularly in developing software packages for digitalizing project tasks [1]. Additive manufacturing and IT-technologies for automating and controlling this process are key technologies of the fourth industrial revolution.

Additive manufacturing is a promising industrial development, fundamentally changing part design. Additive technologies (AT) enable the creation of complex and functionally optimized products that traditional methods cannot produce, and they virtually eliminate post-processing.

Replacing monolithic parts with "lightweight" ones featuring hollow cell regions is a crucial approach to reducing material consumption [2]. In a number of high-tech industries, such as aircraft and astronautics, biomedicine and implantology, robotics, and innovative mechanical engineering, reducing product weight has become critically important. This reduction significantly impacts product efficiency, functionality, and cost-effectiveness.

The increasing complexity of cellular structures and their design in CAD (Computer-Aided Design) systems highlights the need for developing new methods and tools to automate the design of cellular structures. This is particularly relevant for structures based on cylindrical cells [3]. Additionally, sections of cylindrical cells

[1]The work was carried out as part of a research project on GB 21-266 "Marketing support for industrial enterprises of the Republic of Belarus in the context of regional integration and digitalization of the global economy" for 2020-2025.

and straight polyhedral cells are topologically equivalent, allowing cylindrical cells to serve as a reference when compared with polyhedral cells.

The primary objective was to develop a methodology for computer-aided design of parts using cylindrical cells. This methodology includes analyzing the technical and operational properties of the parts, as well as programming models and algorithms for their design, and performing engineering analysis of the designed parts.

2 General characteristics of the used approach

To automate multiple changes to the original geometric model of a solid part and integrate cellular structures into it based on engineering analysis, it is proposed to use SolidWorks CAD in combination with C # libraries for dynamic management of system components via the API.

As a result, a Windows Forms-based software tool was developed, which, using SolidWorks API methods, automates the construction of a test part sample in the form of a parallelepiped and creates an internal cellular structure (e.g., cylindrical) without altering the external geometric characteristics. This reduces the part's total mass while maintaining its technical and operational properties.

The tool's functionality allows dynamic control of the cylindrical cellular structure's geometric configuration by setting parameters, automatically performs engineering analysis of the part in a CAD environment, and exports resulting output data values. Ultimately, it automates multiple studies on the effect of embedded cellular structures' geometric configuration on the part's stress-strain state.

Additionally, to ensure the computer-aided design process, the system includes a number of special features: control over cell visualization, access to rebuilding and deleting the body and cells, the ability to create an analysis study for stress-strain state parameters of the part, including the choice of fixation side and force application, as well as the force itself, access to stress and deformation data, and other technical and operational characteristics.

Constructing a test part in the form of a parallelepiped is a versatile option because any part can be identified as having a certain area of this shape, for which a cellular structure can be calculated. On the other hand, a parallelepiped-shaped sample is a standard part for conducting physical experimental compression work.

The study of cell design methods allowed us to choose two construction directions with different initial data: the construction of cylindrical cells according to a known cell radius, or according to a known constant volume occupied by the cellular structure, i.e. by the percentage of the volume of material to be removed from the part. To fine-tune the creation of cylindrical cells, the designer is given the opportunity to change the construction parameters: select the faces for fixing the part and the face of the applied force, as well as the magnitude of the applied force itself, set any values of radius and volume, change the values of height and radius separately for each cell or not to build it at all.

For the correct construction of a structure with cylindrical cells, two problems were solved for calculating the parameters necessary for constructing a cellular structure:

the first one is based on a given radius of one cell, and the second one is based on a given volume of the cellular structure and the number of cells.

To construct a cellular structure, it is necessary to calculate the wall thicknesses between the cells and between the cell faces and the original part. For the first task (building cells by radius), it will be sufficient to consider building the bases of 3D cells, i.e. circles, within a rectangle defined by the length and width of the body, since in this case the height of the cells is a constant value and does not affect the calculation.

Thus, the values of wall thickness in length and width are equal:

$$l_1 = \frac{b - 2rn_1}{n_1 + 1} \tag{1}$$

$$l_2 = \frac{a - 2rn_2}{n_2 + 1} \tag{2}$$

where a and b are the length and width of the body, n1 and n2 are the number of cells in the length and width directions, respectively, and r is the radius of a single cell.

In the second task, we need to keep the total volume and height of the cells constant, as well as keep the number of cells constant. However, the dimensions of the cylinders will change. To automate the construction of this cellular structure, we need to calculate the radii of each cylinder as well as the thickness of the walls between cells and between cell faces and the original body. The formulas for calculating the thickness of walls l1 and l2 will be:

$$l_1 = \frac{b - 2n_1}{n_1 + 1} \cdot \sqrt{\frac{V^0_{\text{hole}}}{\pi h n}} \tag{3}$$

$$l_2 = \frac{a - 2n_2}{n_2 + 1} \cdot \sqrt{\frac{V^0_{\text{hole}}}{\pi h n}} \tag{4}$$

where n is the number of holes and h is the height of the hole.

The algorithm for constructing a cellular structure is implemented in the CellsDrawer.cs class, the drawCells() method. Drawing a digital model of a cellular structure consists of performing the following basic procedures: selecting a face to create a sketch; creating a sketch; building circles in the sketch that define the bases of 3D cells; exiting the sketch; cutting out the material using the "Elongated Cut" operation. One of the main methods for implementing these procedures using the Solid Works API is described as follows (mas1, mas2, and mas3 are arrays of cell existence, radii, and height, respectively):

3 Main results and analysis

An example of designing a part with a cellular structure, evenly filled with cylindrical cells, using the developed software is shown in Figure 1.

After creating a 3D model of a part with cellular structures, an engineering analysis is performed. The process of engineering analysis consists of performing the following

```
public void drawCells(double[,] mas1, decimal[,] mas2, double[,] mas3)
        {
            faces = bodyDrawer.GetFacesArray();
            var ent = faces.GetValue(1) as Entity;
            ent.Select(true);
            swSketchManager.InsertSketch(false);
            activeSketch = application.swModel.GetActiveSketch2
           double x_current = (-body.GetLenght()) / 2.0 +
           FreeClass.cellCalc.GetRadius() + FreeClass.cellCalc.GetDelta1();
           double y_current = (-body.GetWidth()) / 2.0 +
           FreeClass.cellCalc.GetRadius() + FreeClass.cellCalc.GetDelta2();
           double leftHoleCenterX = x_current, leftHoleCenterY = y_current;
           double delta1 = FreeClass.cellCalc.GetDelta1();
           double delta2 = FreeClass.cellCalc.GetDelta2();
           double rad = FreeClass.cellCalc.GetRadius();
           int row = (int)FreeClass.cellCalc.GetK1(), collumn =
           (int)FreeClass.cellCalc.GetK2();
           for (int i = 0; i < row; i++)
           {
               for (int j = 0; j < collumn; j++)
               {
                   if (mas1[i, j] == 1)
                   {
                       double rrad = mas3[i, j];
                        application.swModel.SketchManager.CreateCircleByRadi
                        us(x_current, y_current, 0, rrad);
                       y_current = y_current + 2 * rad + delta2;
                       double hh = Convert.ToDouble(mas2[i, j]);
                       cut = featureCut(hh);
                       ent.Select(true);
                   }
                   else { y_current = y_current + 2 * rad + delta2; }
               }
              y_current = leftHoleCenterY;
               x_current = leftHoleCenterX + 2 * rad + delta1;
               leftHoleCenterX = x_current; leftHoleCenterY = y_current;
           }
           application.swModel.ClearSelection();
              }.
```

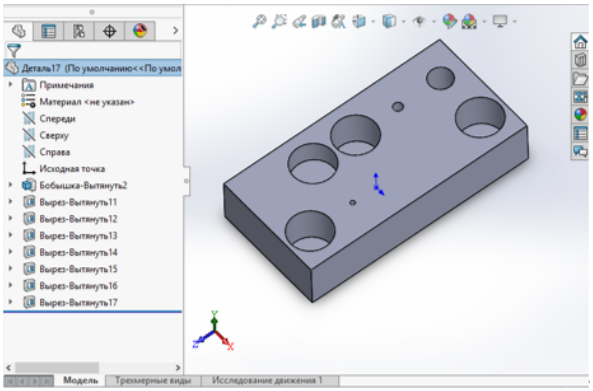


Figure 1: Construction of a part with a cellular structure

main stages: creating an instance of the study; specifying the part material; fixing the faces that must remain stationary during operation; applying loads; building a finite element grid for a given 3D model; launching and performing engineering calculations; recording the output parameters of calculations. The results of the study obtained in SolidWorks are a study of the stress-strain state of a monolithic body and a body with a constructed cellular structure (Figure 2). The stress and strain indicators are entered in the table on the form (Figure 3):

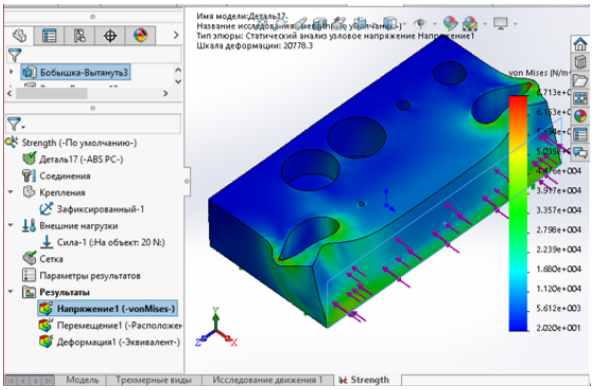


Figure 2: Research of a part with a cellular structure

In the course of experimental studies, a periodic pattern of stress changes was established with a decrease in cell size and a simultaneous increase in their number. Presumably, this pattern is a consequence of the significant influence of the wall thickness between the cells relative to the size of the cell itself (the higher the ratio of these values, the closer the stress and strain values tend to the corresponding values for a monolithic part), as well as the consequence of the different distribution of a given load on the cells, which depends on the size of the cells. To study this theory, graphs of stress and strain dependencies on the percentage of wall thickness to cell radius were constructed (Figure 4), and an experiment was conducted to increase the wall thickness to which force is applied by reducing the radii of the outermost row of cells (the experimental cellular structure contains 98 cells) and the results of the study of this

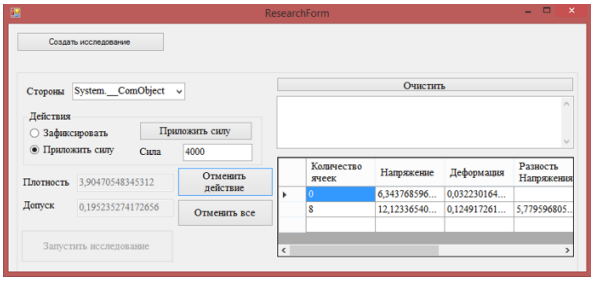
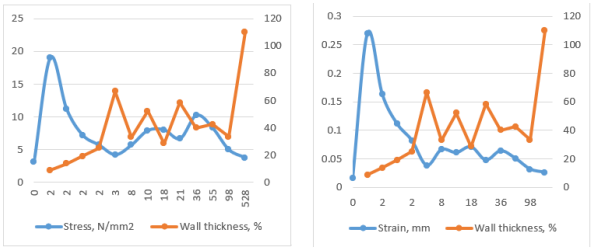


Figure 3: Research results

cellular structure were obtained.

The data obtained suggest that, indeed, with an increase in wall thickness from 1mm to 3mm, the stress and strain values decreased, namely by 30.2% and 32.7% of the initial values, respectively. The updated data is shown in the graphs in Figure 4.



a. Graph of stress dependence

b. Graph of strain dependence

Figure 4: Graphs of stress and strain dependence on the percentage of wall thickness to cell radius on a different number of cells (lower axis)

4 Conclusion

The developed software makes it possible to automatically carry out multiple studies of test parts with cylindrical cells, the configuration of which can vary depending on the size, number of cells, as well as on the specified volume per cellular structure, vary the parameters of the loading schemes of the obtained models of parts, and obtain the values of engineering analysis of the stress-strain state of the parts. Preliminary experimental studies have been conducted on the effect of various configurations of cylindrical cell structures and wall thickness on the stresses and deformations of a particular test part during loading.

The data obtained make it possible to set the task of finding the optimal configuration of a cellular structure when designing it in practical conditions, as well as the possibility of determining the boundaries in which the area of solving this optimization problem is located [4]. The optimization criteria, as well as the limits of the parameters of the stress-strain state of the part, must be set by the customer [5].

Thus, the implementation of additive technologies in industrial production, including the sequential design of parts with cellular structures, allows for a significant material savings while controlling the flexibility and strength characteristics of the part

and reducing the weight of both intermediate and final products. This is particularly important during the transition to Industry 4.0 technologies. In the process of digital transformation of the economy, we should take into account the experiences of advanced industrial countries [6] and implement these technologies to achieve sustainable economic growth.

References

1. Zhalezka B. A., Shamardzina I. A. (2024). *Interaction of the higher professional education system, R&D and the labor market in the context of digitalization of the economy of the Russian Federation and the Republic of Belarus: a collective scientific monograph.* Azhur Publishing House: Yekaterinburg, 274 p. (in Russian).

2. Kochyn V. P. (2022). Design of complex integrated systems. *Computer data analysis and modeling: stochastics and data science.* Belarusian State University: Minsk, pp. 69-72.

3. Korolenok K. S. (2019). Prospects of using additive technologies in economics. *In: World economy and business administration of small and medium-sized enterprises: the 15th International Scientific Seminar held within the framework of the 17th International Scientific and Technical Conf. Science for Education, Production, Economics.* Law and Economics: Minsk, pp. 122-123 (in Russian).

4. Petrov M.A. (2021). Obtaining, characteristics and application of cellular structures in mechanical engineering. *RHYTHM of mechanical engineering.* Vol. **1**, Num. **6**, pp. 20-27 (in Russian).

5. 3D printers in mechanical engineering [Electronic resource] Mode of access: `https://globatek.ru/3d-wiki/ /otrasli-primeneniya-3d-printerov/machinery` Date of access: 05.05.2025 (in Russian).

6. Oh D.H., Danilchanka A., Zhalezka B., Siniauskaya V. (2021). The Transition of Economy from Analogue to Digital in the XXI Century by the case of the Republic of Korea. *Eastern European Journal of Regional Studies.* Vol. **7**, Num. **1**, pp. 109-134.

APPROXIMATE FORMULA FOR THE MATHEMATICAL EXPECTATION OF THE SOLUTION OF A SPECIAL TYPE OF SDE WITH JUMPS

A.V. Zherelo[1]

[1]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]zherelo@bsu.by

The paper presents a formula for the approximate calculation of the mathematical expectation from the solution of a stochastic differential equation containing jumps. The presented formula belongs to the so-called weak approximations and is based on an approximate calculation of the moments of the solution of the equation.

Keywords: stochastic processes with discontinuities, stochastic differential equation, approximate formula, Itô integral

1 Introduction

There are many processes in environment, that can be described only by stochastic processes with jumps. The Poisson process is often used to model processes with jumps. But it is often necessary to construct a more complex model than the model directly containing the Poisson process. In such cases, stochastic differential equations are often used (see, e.g. [1, 2]).

The object of study of this work is an equation of the form:

$$X_t = X_0 + \int\limits_0^t \alpha(X_{s-}, s)ds + \int\limits_0^t \beta(X_{s-}, s)d\tilde{P}_s, \tag{1}$$

where $t \in [0, 1]$, $X_0 \in \mathbb{R}$, $\tilde{P}_t = P_t - \lambda t$ is the compensated Poisson process, P_t is the Poisson process with a parameter $\lambda \in \mathbb{R}$. The stochastic integral at the right side is the Itô integral

This report propose the approximate formula, which can be used for the approximate calculation of the mathematical expectations of functional of the form $\mathbb{E}[G(X_{(\cdot)})]$, where $X \equiv X_t$ is a solution of the equation (1). Here and below the symbol $(\cdot)$ is used to indicate, that G may depends on a trajectory of the solution of the equation.

2 Approximate formula

In this work we assume, that the following conditions are met:

$$|\alpha(y_1, t) - \alpha(y_2, t)|^2 + |\beta(y_1, t) - \beta(y_2, t)|^2 \le K_1|y_1 - y_2|^2,$$
$$|\alpha(y_1, t)|^2 + |\beta(y_1, t)|^2 \le K_2(1 + |y_1|^2),$$

where $K_1, K_2 \in \mathbb{R}$ are constants, $y_1, y_2 \in \mathbb{R}$, and that G has a Fréchet derivative, which we denote by G', and $|G'(x)| \leq C$, where $C \in \mathbb{R}$, for any $x \in \mathbb{R}$.

The proposed approximate formula has the form (see [3]):

$$\mathbb{E}G \approx J(G) = \sum_{j_1,j_2=1}^{2} A_{j_1} B_{j_2} \int_0^1 \int_0^1 \int_0^1 G[Y_{j_1,j_2}(\cdot, u_1, u_2, u_3)] du_1 du_2 du_3, \tag{2}$$

where

$$A_1 + A_2 = 1,$$

$$a_{1,1} = \frac{1}{2}\left(1 - \sqrt{-\frac{A_2}{A_1}}\right),\ a_{1,2} = \frac{1}{2}\left(1 + \sqrt{-\frac{A_2}{A_1}}\right),$$

$$a_{2,1} = \frac{1}{2}\left(1 - \sqrt{-\frac{A_1}{A_2}}\right),\ a_{2,2} = \frac{1}{2}\left(1 + \sqrt{-\frac{A_1}{A_2}}\right),$$

$$B_1 = \frac{1}{2\pi(\mathbb{R})}\left(1 + \frac{1}{\sqrt{1 + 4\pi(\mathbb{R})}}\right),\ B_2 = \frac{1}{2\pi(\mathbb{R})}\left(1 - \frac{1}{\sqrt{1 + 4\pi(\mathbb{R})}}\right),$$

$$b_1 = \frac{1}{2}\left(1 - \sqrt{1 + 4\pi(\mathbb{R})}\right),\ b_2 = \frac{1}{2}\left(1 + \sqrt{1 + 4\pi(\mathbb{R})}\right),$$

$$\begin{aligned}
Y_{j_1,j_2}(t) \equiv\ & Y_{j_1,j_2}(t, u_1, u_2, u_3) = X_0 \\
& \alpha\Bigg(X_0 + \alpha\left(X_0 + \beta(X_0, u_3)\rho^{(2)}_{j_2}(u_2-, u_3), u_2\right)\rho^{(1)}_{j_1,2}(u_1-, u_2) + \\
& \qquad \beta\left(X_0 + \alpha(X_0, u_2)\rho^{(1)}_{j_1,2}(u_3-, u_2), u_3\right)\rho^{(2)}_{j_2}(u_1-, u_3), u_1\Bigg)\rho^{(1)}_{j_1,1}(t, u_1) + \\
& \alpha\Bigg(X_0 + \alpha\left(X_0 + \beta(X_0, u_3)\rho^{(2)}_{j_2}(u_1-, u_3), u_1\right)\rho^{(1)}_{j_1,1}(u_2-, u_1) + \\
& \qquad \beta\left(X_0 + \alpha(X_0, u_1)\rho^{(1)}_{j_1,1}(u_3-, u_1), u_3\right)\rho^{(2)}_{j_2}(u_2-, u_3), u_2\Bigg)\rho^{(1)}_{j_1,2}(t, u_2) + \\
& \beta\Bigg(X_0 + \alpha\left(X_0 + \alpha(X_0, u_2)\rho^{(1)}_{j_1,2}(u_1-, u_2), u_1\right)\rho^{(1)}_{j_1,1}(u_3-, u_1) + \\
& \qquad \alpha\left(X_0 + \alpha(X_0, u_1)\rho^{(1)}_{j_1,1}(u_2-, u_1), u_2\right)\rho^{(1)}_{j_1,2}(u_3-, u_2), u_3\Bigg)\rho^{(2)}_{j_2}(t, u_3),
\end{aligned}$$

and

$$\rho^{(1)}_{j_1,k}(s, u_k) = a_{j_1,k} 1_{[u_k,1]}(s), \quad k = 1, 2, \quad \rho^{(2)}_{j_2}(s, u_3) = b_{j_2} 1_{[u_3,1]}(s),$$

$$1_{[u_k,1]}(s) = \begin{cases} 1, & s \in [u_k, 1], \\ 0, & \text{overwise.} \end{cases}$$

The accuracy of the proposed formula was assessed.

Theorem. The following estimate of the error of the formula (2) is valid

$$\mathbb{E}[G] - J(G) \leq \frac{4}{3} C \sum_{j_1, j_2 = 1}^{2} |A_{j_1}||B_{j_2}| \sqrt{K_2(1 + X_0^2)} t^{3/2} + o(t^{3/2}).$$

The example of application of the formula (2) is proposed.

References

1. Bruti-Liberati N., Platen E. (2006). *Approximations of Jump Diffusions in Finance and Economics*. Quantitative Finance Research Center, University of Technology, Sydney.

2. Yongfeng Wu, Xue Liang (2018). Vasicek model with mixed-exponential jumps and its applications in finance and insurance. *Advances in Difference Equations*. Vol. **138**, pp (2018). https://doi.org/10.1186/s13662-018-1593-z.

3. Zherelo A. (2025). On an Approximate Formula for Functionals with Respect to a Solution of Stochastic Differential Equation with a Drift and Random Process with Jumps. *Nonlinear Phenomena in Complex Systems*. Vol. **28(2)** , pp. 137 – 143.

SOME APPROXIMATIONS OF REPLICATING PORTFOLIO

N.M. Zuev[1], P.M. Lappo[2]

[1,2]*Belarusian State University*
Minsk, BELARUS
e-mail: [1]zuevnm@bsu.by, [2]lappopm@bsu.by

We consider (B, S)-market [1] with N risky and one riskless assets. To estimate financial derivatives we construct an approximation of replicating portfolio that minimizes an expectation of penalty function.

Keywords: replicating portfolio, (B, S)-market, derivatives pricing

1 Introduction

In the first part we consider one-period market model with discrete time and finite number states of the world. The approximation of replicating portfolio is built as the portfolio that minimizes the expectation of penalty function. In the second part we analyze the multi-period model with discrete time and use dynamic programming method to built the approximation portfolio.

2 One-period Model

We begin with a finite number N of risky securities or assets $S_1, \ldots, S_N$, and $B_0 = S_0$ is riskless security (bank account) [2]. In this part we consider only their values at times 0 and 1. At time 0 the investors know the time-0 values, but the time-1 values are random variables on probability space $(\Omega, \mathcal{F}, P)$. The time-0 prices of the securiries are assumed to be strictly positive. Since $S_j(0, \omega)$ is the same for all $\omega \in \Omega$ we simply denote this common value as $S_j(0)$ and consider the row vector

$$S(0) = [S_1(0), \ldots, S_N(0)]^T.$$

The time-1 prices are random variables defined on probability space $(\Omega, \mathcal{F}, P)$. Let us denote the vector

$$S = S(1) = (S_1, \ldots, S_N).$$

Investors select a portfolio of the assets at time 0. The number of the units of the asset j held from the time 0 to time 1 is denoted by the numbers $\theta_j, j = 0, \ldots, N$. If θ_j is positive, θ_j units of security j are purchased. If θ_j is negative, $|\theta_j|$ units of security j are sold short. We denote the portfolio as the column (trading strategy)

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{bmatrix}.$$

Then the value of corresponding portfolio at time 0 is

$$S(0)\theta = \theta_0 S_0(0) + \theta_1 S_1(0) + \cdots + \theta_N S_N(0).$$

The time-1 value of this portfolio will depend on the state of nature. If the state ω occurs, then the time-1 value is

$$\theta_0 S_0(1,\omega) + \theta_1 S_1(1,\omega) + \cdots + \theta_N S_N(1,\omega).$$

Let we have the derivative security $f(S_j(1,\omega))$, where f is a measurable function. The replicating portfolio θ for it is determined as the portfolio for which for all ω

$$f(S_j(1,\omega)) = \theta_0 S_0(1,\omega) + \theta_1 S_1(1,\omega) + \cdots + \theta_N S_N(1,\omega).$$

If our market is arbitrage-free, then the value of this portfolio at time 0 is the derivative price at time 0. If our market is complete, then this price is unique.

Let us denote the penalty function as $U(x,y)$. We propose to evaluate derivative price as the time-0 value of the portfolio that minimizes expected value

$$E(U(f(S_j(\omega), \theta_0 S_0(\omega)) + \theta_1 S_1(\omega) + \cdots + \theta_N S_N(\omega))) \to \min_{\theta}. \qquad (1)$$

As the functions $U(x,y)$ we can take $U(x,y) = (x-y)^2$, $U(x,y) = |x-y|$ or others. In the case of the two first functions, if the minimum expectation is 0, then θ that minimizes the expectation is the replicating portfolio. Otherwise we have some approximation. Let us denote the covariance matrix S by Σ.

Theorem 1. *If the function $U(x,y) = (x-y)^2$, then the approximating portfolio, minimizing the left side of (1) is given by the equality*

$$\overline{\theta} = \Sigma^{-1} E(f(S_j, S)),$$

and the derivative price approximation at time 0 is

$$\overline{f}(S_j) = \Sigma^{-1} E(f(S_j, S)) \cdot S(0).$$

The proof is based on differentiation of the left hand side of (1) w.r.t. θ and equating the derivatives to 0. As the measure P we can use empirical distribution.

3 The Multi-period Model

A construction of an approximation of the replication portfolio in multi-period model with two assets was considered by authors in [3]. In the situation with $N+1$ assets the approximating portfolio could be constructed using the dynamic programming method and conditional mathematical expectations.

References

1. Shiryaev A.N. *Osnovy stokhasticheskoi finansovoi matematiki*. M.: Fazis. T.2: Teoriya. 1998. S. 482-1106.

2. Financial Economics with Applications to Investments, Insurance and Pensions / P.P. Boyle [et al] // *The Actuarial Foundation, Schaumburg, Illinois*. 1998. 670 P.

3. Zuev N.M., Lappo P.M. *O postroyenii portfelya investitsiy minimiziruyushchego srednekvadraticheskoye otkloneniye ot funktsii vyplat*. XIV Belorusskaya matematicheskaya konferentsiya. Minsk, 2024. S. 136-137.

Index

Scientific edition

COMPUTER DATA ANALYSIS AND MODELING: STOCHASTICS AND DATA SCIENCE

Proceedings of the Fourteenth International Conference

September 24–27, 2025, Minsk

In the author's edition

Responsible for Issue V.A. Voloshko

Signed in print August , 2025. Format 60x84 1/8. Offset paper.
Digital printing. Conventional printed sheets 41.85. Publisher's signatures 40.6.
Circulation of copies. Order

Belarusian State University.
Certificate of state registration of the publisher, manufacturer,
distributor of printed publications No. 1/270 of 03.04.2014.
4 Independence ave., 220030, Minsk.