APPLIED METHODS OF STATISTICAL ANALYSIS. NONPARAMETRIC APPROACH

PROCEEDINGS
OF THE INTERNATIONAL WORKSHOP

14-19 September 2015

Novosibirsk & Belokurikha
2015

UDC 519.22(063)
A 67

Editors:

Prof. Boris Lemeshko, Prof. Mikhail Nikulin,
Prof. Narayanaswamy Balakrishnan

A 67 **Applied Methods of Statistical Analysis. Nonparametric Approach** - AMSA'2015, Novosibirsk, Russia, 14-19 September, 2015: Proceedings of the International Workshop. - Novosibirsk: NSTU publisher, 2015. - 462 pp.
ISSN 2313-870X

ISSN 2313-870X **UDC 519.22(063)**

Applied Methods of Statistical Analysis. Nonparametric Approach

Chairmen:

Narayanaswamy Balakrishnan,	McMaster University, Canada
Mikhail Nikulin,	University of Bordeaux, France
Aleksey Vostretsov,	Novosibirsk State Technical University
Boris Lemeshko,	Novosibirsk State Technical University
Evgeny Tsoy,	Novosibirsk State Technical University
Felix Tarasenko,	Tomsk State University
Aleksandr Medvedev,	Siberian State Aerospace University

Scientific Committee:

A. Antonov,	Institute of Nuclear Power Engineering, Russia
N. Balakrishnan,	McMaster University, Canada
G. Koshkin,	Tomsk State University, Russia
M. Krnjajić,	National University of Ireland, Ireland
B. Lemeshko,	Novosibirsk State Technical University, Russia
N. Limnios,	Universite de Technologie de Compiegne, France
A. Medvedev,	Siberian State Aerospace University, Russia
V. Melas,	St. Petersburg State University, Russia
G. Mikhailov,	Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
M. Nikulin,	University of Bordeaux, France
V. Ogorodnikov,	Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
B. Ryabko,	Siberian State University of Telecommunications and Information Sciences, Russia
V. Solev,	St.Petersburg Department of Steklov Mathematical Institute RAS, Russia
F. Tarasenko,	Tomsk State University, Russia
V. Timofeev,	Novosibirsk State Technical University, Russia
A. Vostretsov,	Novosibirsk State Technical University, Russia

Organizing Committee:

Ekaterina Chimitova, Mariia Semenova, Victoria Volkova,
Evgenia Chetvertakova, Stanislav Vozhov, Victor Demin

Preface

The Third International Workshop “Applied Methods of Statistical Analysis. Nonparametric Approach” – AMSA’2015 is organized by Novosibirsk State Technical University. It took place in the resort Belokurikha located at the foothills of Altai, Russia. The purpose of our Workshop is to organize interesting meeting on different statistical problems of interest. This seminar aims to provide an overview of recent results in applied mathematical statistics and primarily on testing statistical hypotheses, statistical methods in reliability and survival analysis, nonparametric methods, robust methods of statistical analysis, statistical simulation of natural processes, econometric methods and modeling, information and statistical analysis of complex systems.

Within the framework of AMSA’2015, the XV International Symposium on Nonparametric Methods in Cybernetics and System Analysis was organized by Siberian State Aerospace University called after academician M.F. Reshetnev and Tomsk State University. The first such Symposium was held in 1976 in Tomsk, and since then, after each two or three years, it was taken at various places in Siberia, collecting participants from all the Soviet Union, and later – from other countries. The Symposium is devoted to the development of modern mathematical methods for building intellectual computer systems for various purposes operating under incomplete knowledge of the studied process and problems of system analysis.

The First International Workshop “Applied Methods of Statistical Analysis. Simulations and Statistical Inference” – AMSA’2011 and the Second International Workshop “Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control” – AMSA’2013 took place in Novosibirsk, Russia. This city is very well known for its fundamental contributions to the development of theory of the probability, mathematical statistics, stochastic processes and statistical simulation. These meetings had been focused on recent research in the areas of survival analysis, reliability, quality of life, and related topics, from both statistical and probabilistic points of view. The great attention is paid to applications of statistical methods in survival analysis, reliability and quality control.

The Workshop proceedings would certainly be interesting and useful for specialists, who use statistical methods for data analysis in various applied problems arising from engineering, biology, medicine, quality control, social sciences, economics and business.

Boris Lemeshko

Contents

Rank as Proxy for the Observation in Statistical Procedures

F.P.Tarasenko and V.P. Shulenin
National Research Tomsk State University, Russia

Abstract

The properties of rank tests are discussed and it is shown that besides computational convenience, in many cases they have advantages over their counterparts on observations.

Keywords: Statistical procedures, effectiveness and efficiency of procedures, rank tests, statistical properties of ranks.

Introduction

Ranks often are preferred to actual observation values in processing experimental data. There are a few good reasons for that:

- Ranks are pure whole numbers and, hence, are very convenient to calculate. In contrast to this, observations often are continuous values that need rounding (with unpredictable consequences), and registered in various measuring scales (with each scale having different set of allowed operations over its values).
- Ranks are related to observations and, hence, contain some of the same (sought by observer) information as well as observations themselves.
- Relation between the sample value and its rank becomes even stronger with growth of a sample size; this promises the good asymptotic properties to procedures based on ranks.
- Last but not least: some distribution-free properties of ranks insure robustness to the rank procedures, – much appreciated property in statistical practice.

Here follows a brief survey of old and a few new results on these issues.

1 Basic Distributions

Let $\vec{X} = (X_1, ..., X_n)$ be a sample from p.d.f. $F_X(x)$ with a density $f_X(x)$, $x \in R^1$. Let, then, $\vec{X}_{(\cdot)} = (X_{(1)}, ..., X_{(n)})$ be the ordered statistics, and $\vec{R} = (R_1, ..., R_n)$ be a vector of ranks for the sample$\vec{X} = (X_1, ..., X_n)$. Between the sample $\vec{X}$ and the pair $\left\{\vec{X}_{(\cdot)}, \vec{R}\right\}$ there exists mutual one-to-one correspondence, which means that the information contained in observations $\vec{X} = (X_1, ..., X_n)$ maybe split into two parts. One part belongs to order statistics $\vec{X}_{(\cdot)} = (X_{(1)}, ..., X_{(n)})$, the other – to ranks $\vec{R} = (R_1, ..., R_n)$. Therefore, a seeking the same aim statistical procedures may be built either on raw observations $\vec{X} = (X_1, ..., X_n)$, or on order statistics $\vec{X}_{(\cdot)} = (X_{(1)}, ..., X_{(n)})$, or on ranks $\vec{R} = (R_1, ..., R_n)$.

The vector random variable of a "mixed" type (i.e. consisting of discrete and continuous components [1]), which our pair $\left\{\vec{X}_{(\cdot)}, \vec{R}\right\}$ belongs to, is characterized by corresponding probability distributions:

C.d.f. for i.i.d.r.v. $\vec{X} = (X_1, ..., X_n)$ is equal to

$$F_{X_1,...,X_n}(x_1, ..., x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) = \prod_{i=1}^{n} F_X(x_i). \tag{1}$$

C.d.f. for r-th order statistics ($1 \le r \le n$, $x \in R^1$) is

$$F_{X_{(r)}}(x) = P\left\{X_{(r)} \le x\right\} = \sum_{i=r}^{n} C_n^i F_X^i(x)(1 - F_X(x))^{n-i} = I_{F(x)}(r, n - r + 1), \tag{2}$$

where $I_p(n\ ,\ m)$ is the incomplete beta-function tabulated in [2]. Corresponding density is

$$f_{X_{(r)}}(x) = nC_{n-1}^{r-1} F_X^{r-1}(x)(1 - F_X(x))^{n-r} f_X(x). \tag{3}$$

The joined p.d.f. of random vector $\vec{X}_{(.)} = (X_{(1)}, ..., X_{(n)})$ is

$$f_{\vec{X}_{(\cdot)}}(x_{(1)}, ..., x_{(n)}) = \begin{cases} n! f_{\vec{X}}(x_{(1)}, ..., x_{(n)}) = n! \prod\limits_{i=1}^{n} f_X(x_{(i)})\ , x_{(1)} < ... < x_{(n)} \\ 0\ ,\ \ otherwise \end{cases} \tag{4}$$

In case of symmetrical (invariant to permutations of arguments) distribution of X, order statistics and rank vector are independent:

$$f_{\vec{X}_{(.)}\vec{R}}(x_{(1)}, ..., x_{(n)}; r_1, ..., r_n) = f_{X_{(1)},...,X_{(n)}}(x_{(1)}, ..., x_{(n)}) \cdot P\left\{\vec{R} = \vec{r}\right\} \tag{5}$$

and their conditional and unconditional distributions coincide [3].

If the distribution $g_{X_1,...,X_n}(x_1, ..., x_n)$ is non-invariant to permutations, the famous Hoeffding's Theorem [4] holds:

$$P_g\{\vec{R} = \vec{r}\} = P_g\{R_1 = r_1, ..., R_n = r_n\} = \frac{1}{n!} M \left\{ \frac{g_{\vec{X}}(X_{(r_1)}, ..., X_{(r_n)})}{f_{\vec{X}}(X_{(r_1)}, ..., X_{(r_n)})} \right\} \tag{6}$$

which in case of independence of variables takes the form of

$$P_g\{\vec{R} = \vec{r}\} = \frac{1}{n!} M \left\{ \prod_{i=1}^{n} \frac{g_{X_i}(X_{(r_i)})}{f_X(X_{(r_i)})} \right\} \tag{7}$$

The joint d.f. $F_{R_i X_i}(x, y)$ of random variables R_i and X_i is [5]

$$F_{R_i X_i}(x, y) = n^{-1} \sum_{j=1}^{n} C(x - j) \cdot \int_{-\infty}^{y} f_{X_i|R_i=j}(x\,|j\,) dx =$$

$$n^{-1} \sum_{j=1}^{n} C(x - j) \cdot F_{X_{(j)}}(y), 1 \le i, j \le n \tag{8}$$

and a formal expression for joint density of the mixed type random variable (X, Y) is

$$f_{XY}(x, y) = \sum_{i-1}^{n} p_X(\tilde{x}_i) f_{Y|X}(y|\tilde{x}_i)\delta(x - \tilde{x}_i). \qquad (9)$$

2 Some Characteristics of Independence between Observations and Their Ranks

Suitableness of ranks for coming out as a proxy of the sample measurements in statistical processing of experimental data, clearly depends on how tight is the connection between them. Most general presentation of interdependence between random variables is given by their joint distribution function (8). Its one-sided presentations are made by conditional distributions of each of them conditioned by value of the other one. Such conditional distributions may be obtained by corresponding integration of d.f. (8).

But there are several particular indicators characterizing different aspects of the statistical connectedness. Let us describe some of these quantitative indices for observations and their ranks.

2.1 Regression

The regression function determines the relationship between a random variable and corresponding values of dependent value. If both regression lines coincide, it means that the relationship between the two variables is strictly functional. The more they differ, the weaker is the relationship. In case of independency the lines are orthogonal to each other.

Let us denote a regression of the observation X_i of its rank R_i as $M(X_i\,|R_i = j\,)$, $1 \le i, j \le n$, and regression of the rank R_i of X_i as $M(R_i = j\,|X_i = x\,)$, $1 \le i, j \le n$, $x \in R^1$. It can be shown [5] that

$$M(R_i = j\,|X_i = x\,) = 1 + (n-1)F_X(x), x \in R^1, \qquad (10)$$

$$M(X_i\,|R_i = j\,) = M(X_{(j)}), 1 \le i, j \le n. \qquad (11)$$

Quantitative and qualitative analyses of these lines behavior for different distributions show [5] that the lines are crossing under a certain angle which is monotonously decreases with sample size increasing. It means that interdependence between rank and observation becomes only stronger under enlarging n.

2.2 Correlation

The correlation coefficient is a measure of connexion, which is very popular among data analysis practitioners. Its calculation for observations and ranks gives a result [5]:

$$\rho_{X_iR_i}(F) = \frac{\sqrt{3}}{2}\left(\frac{n-1}{n+1}\right)^{1/2}\frac{\Delta(F)}{S(F)}, \forall i \in (1, ..., n), \tag{12}$$

where $\Delta(F)$ is the Geeny's average difference defined as

$$\Delta(F) = \int_{-\infty}^{+\infty}\int |x-y|dF(x)dF(y) \tag{13}$$

and $S(F)$ is standard deviation defined as

$$S(F) = \left(\frac{1}{2}\int_{-\infty}^{+\infty}\int (x-y)^2 dF(x)dF(y)\right)^{1/2}. \tag{14}$$

It turns out that correlation between observation and its rank is always positive, equal for any observation in a sample, fast approaches, with growing n, to a value typical for the length of tails of the distribution. Here are values of $\rho_{XR}(F)$ for some distributions:

$F(x)$	Uniform	Gaussian	Logistic	Laplasian
$\rho_{XR}(F)$	1,00	0,98	0,95	0,92

The longer tails of a distribution are, the less correlated are ranks and observations. This explains, in a way, difference between effectiveness of the same rank procedure being applied to data from different distributions.

2.3 Information

Various “quantities of information” are used for estimating degree of connexion tightness. In our case of considering ties inside a pair (X_i, R_i), the Shannon's quantity of information

$$I(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y)\ln\left[\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right]dxdy \tag{15}$$

after cumbersome calculations, appeared to be

$$I(X,Y) = \ln n - \left(\sum_{k=1}^{n-1}\ln k + \frac{n-1}{2} - \frac{2}{n}\sum_{k=1}^{n-1}k\ln k\right), \tag{16}$$

or, asymptotically, with accuracy of $ARE_F(U:t) = 3$, is

$$I(X,R) = \ln\sqrt{ne/2\pi}. \tag{17}$$

So, quantity of information in ranks about observations does not depend neither on index i of the observation, nor on its d.f. $W_1, ..., W_k$, and increases, together with n, with velocity $\ln\sqrt{n}$. This ensures, that qualities of the rank statistical procedures will asymptotically approximate merits of procedures based on observations themselves.

3 On Some Advantages of Ranks over Observations

It was already mentioned that ranks have attracted interest from statisticians and data analysts due to their content (they share the information with observations) and to their form (they are integers, which are very convenient to work with). But it does not mean that the straightforward replacement of observations by their ranks in a statistical procedure will bring a desired effect. First, observations and their ranks usually belong to different measuring scales, with different permissible operations for their processing. This restricts usage of direct similarity of procedures to the case of their containing equivalent permissible operations only. Second, ranks of sample values contain the same kind of information as the values themselves if only this information is connected with own size of each value (when large-sized value receives higher rank). But if the information of interest is about other relations between observations, then another, the appropriate way of ordering values is required to map the information onto ranks. And the third, last but not least: the algorithms (sequences of operations) of statistical processing of data depend on a'priori knowledge of stochastic nature of the data. This is why the same sample must be treated much differently under conditions of parametric, non-parametric, and robust statistics. And here again an important role belongs to proper way of put observations in order to preserve useful information on ranks. But the most surprising and admiring feature of ranks manifests itself in complicated circumstances of robust statistics: rank test could be more effective than its counterpart based on observations.

Let us discuss briefly the abovementioned peculiarities of ranks and give some illustrative examples.

3.1 Ordering that transfers target information from observations onto ranks

Usefulness of ranks as substitutes to observations is primarily based on their attachment to the values of observations. But sometimes a statistical procedure is designed to extract from the sample such information that is indirectly defined by the values of observations but directly by their relevancy to other random events. In such a case, neither the sample alone, nor its rank vector are valid for achieving the purpose of data processing.

Typical example is homogeneity tests. The purpose is to reveal the identity or distinction between two distributions, judging by a comparison of the samples taken from them. The test is made by combining the two samples into one, and detecting a degree of their overlapping. If distributions are different then observations from one sample will dominate in number over another one in those regions where their probability is higher. For instance, if distributions are shifted (differ in location parameter) then observations from one of them will overwhelm the other in number at one side of the whole range of values; if distributions differ in scale parameter, then the observations from the wider one will outnumber those from narrower at both far ends of the range. The same will happen to the ranks of observations, if ordering was made on the whole joined sample but with retained information of belonging observations to their distributions.

3.2 Comparison of rank tests with their counterparts based on observations

The general theory of rank tests is presented in books by Lehman [6], Hayek and Shidak [3], Pury and Sen [4], Hettsmanspreger [7]. Here we give only a few examples revealing merits of rank tests in comparison with analogous tests based on observations.

The notion of the Pitman asymptotic relative efficiency (ARE) is widely used for comparison of two tests, T_n and S_n. $ARE_F(T_n : S_n)$ characterizes the ratio of sample sizes n_1 and n_2 under which T_n and S_n with equal levels of significance ensure equal ARE against the same sequence of contigual alternatives converging to zero hypothesis.

For the Wilcoxon sign rank test S^+ and Student's t-test

$$ARE_F\,(S^+ : T(\vec{X})) = 12\sigma^2 \left[\int\limits_{-\infty}^{\infty} f^2(x)dx \right]^2 . \tag{18}$$

In Table 1 the numerical values of $ARE_F\,(S^+ : T(\vec{X}))$ are presented for some symmetric distributions.

Table 1

Distribution $F(x)$	$ARE_F(S^+ : T(\vec{X}))$
Uniform	1
Gaussian	$3/\pi = 0,955$
Logistic	$\pi^2/9 = 1,097$
Double exponential	1,5

For the Wilcoxon sign rank test S^+ and the sign test S

$$ARE_F\,(S : S^+) =^2_F (S)/^2_F(S^+)) = 4\sigma^2 f^2(0)/3 \cdot \left[\int f^2(x)dx \right]^2 . \tag{19}$$

Its numerical values are given in Table 2.

Calculations of ARE for many other pairs of tests were made (e.g. in [7 – 12]). Some general conclusions follow from their consideration:

– In most cases ARE does not depend on scale parameter and is connected to the distributions' family type only.

Table 2

Distribution $F(x)$	$ARE_F(S:S^+)$
Uniform	$1/3$
Gaussian	$2/\pi = 0,637$
Logistic	$\pi^2/12 = 0,822$
Double exponential	2

– AREs may take various values not limited from above, but have non-zero lower limits. For instance, $ARE_F(U:t) \geq 0,864$, which means that in two-sampled problem of shift we may loose in efficiency not more than $13,6\%$ using Wilcoxon's test instead of Student's one. Under Gaussian distribution the loss is 5% only. The most favorable distribution ($ARE_F(U:t) = 3$) is gamma-distribution with $p = 1$. So, under these circumstances the Wilcoxon test is always preferable among other tests.

Robust statistics is an approach to designing statistical procedures at an intermediate (between parametric and non-parametric) level of a priori knowledge about stochastic nature of observations. Underlining them distribution is considered as known approximately: it belongs to a "supermodel", a certain vicinity of some parametric function. The procedures are designed that remain effective ("robust") until actual distribution lies inside the vicinity; there are among those the rank procedures, too. And they demonstrate certain advantages.

For example, efficiency of H-test of Kruskall-Wallis against its Gaussian competitor, Fisher's F-test is [7]

$$ARE_F(H:F) = 12\sigma_f^2 \left[\int_0^1 f(F^{-1}(u)du) \right]^2, \qquad (20)$$

and this formula is valid for several other counterparts of tests [8]. Numerical values of it for Gaussian model with a scale obstruction

$$F \in \Im_{\varepsilon,\tau}(\Phi) = \{F : F_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau)\}, 0 \leq \varepsilon < 1/2, \tau \geq 1$$

are given in Table 3.

Table 3

	ε	0.00	0.01	0.03	0.05	0.08	0.10	0.15	0.20
$ARE_{F_{\varepsilon,\tau}}(H:F)$	$\tau = 3$	0.955	1.009	1.108	1.196	1.309	1.373	1.497	1.575
	$\tau = 5$	0.955	1.150	1.505	1.814	2.201	2.412	2.795	3.006
	$\tau = 7$	0.955	1.369	2.115	2.759	3.553	3.977	4.724	5.099

It is seen that H-test looses in efficiency only 5% to the optimal F-test of Fisher in Gaussian case, but much overwhelms it under deviations from normality.

References

[1] Wilks S. Mathematical statistics. – M.: Nauka, 1967. – 632 p.

[2] K. Pearson. Tables of the Incomplete Beta-Function. Cambridge, England: The University Press, 1934.

[3] Gaek Ya., Shidak Z. Theory of rank tests (in Russian). – M.: Nauka, 1971. – 376 p.

[4] Puri M.L., Sen P.K. Nonparametric methods in Multivariate Analysis. John Wiley, N.Y., 1970, 440 p.

[5] Shulenin V.P., Tarasenko F.P. Regression functions for observations and their ranks (In Russian) // Tomsk State University Journal of Control and Computer Science 2003, N. 280, pp. 213 – 216.

[6] Lehmann E. L. Nonparametric: Statistical Methods Based on Ranks. Holden – Day. San Francisco, 1975. - 326 p.

[7] Hettmansperger T.P. Statistical inference based on ranks. John Wiley and Sons, New York. 1984. 323 p.

[8] Kendall M, Stuart A. Statistical inference and communication (in Russian). – M.: Nauka, 1973. – 899 p.

[9] Gibbons J.D. Nonparametric Statistical Inference. New York, McGraw-Hill, 1971.

[10] Randles R.H. and Wolfe D.A. Introduction to the theory of nonparametric statistics. New York: Wiley. 1979.

[11] Tarasenko F.P. Nonparametric statistics (in Russian). Tomsk. TSU Publisher, 1976. – 292p.

[12] Tarasenko F.P., Shulenin V.P. On statistical relation between observations and its rank (in Russian) // Trudy SFTI pri Tomskom universitete. 1971, Vol. 60 , p. 220 – 228.

[13] Hollander M., Wolfe D. A. Nonparametric Statistical Methods: John Wiley and Sons, New York. 1973. 503 p.

Smooth Estimation of Multivariate Reliability Function

IRINA L. FUKS[1] AND GENNADY M. KOSHKIN[2]
[1] *Department of Computer Science,*
National Research Tomsk State University, Tomsk, Russia
[2] *Department of Applied Mathematics and Cybernetics,*
Laboratory of Geological Informatics of Computer Science Department,
National Research Tomsk State University, Tomsk, Russia
e-mail: fooxil@sibmail.com, kgm@mail.tsu.ru

Abstract

Empirical distribution and reliability functions are discrete that often does not correspond to real random variables in technical and physical applications. Smooth nonparametric estimators of the multivariate reliability function based on the product of finite and Laplace kernel functions are suggested. The asymptotic mean square error (MSE) of the estimator and its limiting distribution are presented that allow a new interval estimator of the reliability function to be constructed. Advantages of the suggested estimators over the well-known parametric algorithms are discussed.

Keywords: Multivariate reliability function, smooth kernel estimation, mean square error, asymptotic normality, interval estimation.

Introduction

Design, construction, and operation of complex instrumental and software systems and complexes require insurance of their reliability [1, 2]. Researchers who are engaged in the prediction of reliability of objects of research on experimental stands during field experiments [3] and estimation of the reliability of semiconductor optoelectronic devices [4], lasers [5], and their components [6] also face these problems.

To calculate the reliability and to predict failures, the simple characteristic of efficiency of non-restorable elements $S(t) = 1 - F(t)$, $t \in R^1$, is often used, where $F(t)$ is the distribution function of the failure time T for the examined element. The function $S(t)$ describes the probability of failure-free operation of the non-restorable element up to the moment t and is called the reliability function. To calculate the reliability, it is convenient to use values of the failure characteristics of individual elements, because the formulas so derived are simple and convenient for engineering practice [7]. A more complicated problem is estimation of the strength reliability of the element consisting in determining the probability of failure-free operation (see [1], p. 14) which is expressed through the reliability function:

$$P\{S(s-z)\} > 0 = S(0),$$

where the random variable specifies the ultimate stress and the random variable determines the tension in the element under the action of an external load.

In parametric statistics, a function depending on a finite number of the unknown parameters that are to be estimated is chosen as the distribution function of the failure time $F(t)$. Some distributions describe sufficiently accurately the occurrence of failures of these or other elements. For example, the exponential distribution $F(t, \lambda) = 1 - e^{-\lambda t}$ for $t \geq 0$ describes moments of failure of elements whose residual lifetime is independent of the period of preceding operation. According to [8], the Weibull distribution $F(t, \lambda, \alpha) = 1 - e^{-(\lambda t)^{\alpha}}$, where $t \geq 0$, and $\lambda, \alpha \geq 0$, is used to describe the fatigue phenomena [6] and failures of electronic devices [9]. In [10], the behavior of the reliability functions was investigated when the occurrence of failures in the sequence of tests was described by a Markovian process. The methods of maximum likelihood, moments, and least squares allow one to estimate sufficiently efficiently the unknown parameters from observations of random variables [11].

In problems of estimation of the reliability of complex systems, the moments of failure of the examined elements are statistics; as a rule, they are determined as a result of expensive experiments. In this case, researchers often do not have sufficient information on the elements themselves and on the nature of occurrence of their failures that complicates and sometimes makes even impossible the construction of an adequate parametric model of actual object. In some cases, it is required to improve significantly the reliability of the evaluation, for example, for potentially dangerous equipment. In this case, for small volume of statistical data and unknown distribution law, parametric models can inadequately describe actual failures, which can cause catastrophic consequences. Therefore, the problem of the development and investigation of nonparametric methods of analysis of system reliability from the data on failures of products and devices becomes urgent.

The main advantage of the nonparametric procedures compared with the parametric ones consists in the fact that they remain efficient when prior information on the distributions does not allow one to take advantage of any parametric family of distributions to construct a mathematical model of the object. Thus, actual probability densities of random variables and can have several extreme values (see [1], p. 16), thereby complicating the strength reliability estimation of the elements. At the same time, the application of discrete empirical distribution and reliability functions which are nonparametric estimators leads to deterioration of the accuracy of the algorithms so obtained when solving many reliability problems. Additional problems can arise at the boundaries of the definition regions, at which the estimates can take zero or unity values, though their true values differ from zero or unity.

In the present work extending the results of [12] to the multivariate case, a class of smooth estimators of the multivariate reliability functions is considered that in addition, have no disadvantages of the empirical reliability function at the boundaries of the definition region. It should be mentioned that the smooth estimator was first suggested for the one-dimensional distribution function in 1964 [13]. Since then, the properties of such estimators have been studied by various authors (for example, see [14]–[22]). To solve the important practical problem of calculating the bandwidth for a smooth estimator of the distribution function and hence for a smooth estimator of the reliability function from empirical data (see formula (3) below), some approaches have been employed, including methods of leave-one-out cross-validation [23], plug-in [24], and cross-validation [25]. It should be noted that analogous problem was solved in [26]–[28] by other methods.

Let R^l be the l-dimensional Euclidean space and $T = (T_1, T_2, \ldots, T_l)$ be the failure-free operating vector-period of an system of l elements. One of the reliability indicators of the non-restorable system of l elements is time of its failure-free operation. The probability that such system will operate till vector-moment $t = (t_1, \ldots, t_l)$ is expressed through the reliability function

$$S(t) = P(T > t) = 1 - F(t). \tag{1}$$

Function (1) allows other probabilities to be calculated. Thus, the probability of failure of l elements in l-dimensional parallelepiped $(t, t+x)$ is expressed through the difference $S(t) - S(t, t+x) = P(t < T \leq (t+x))$.

The uncertainty in the failure moment for a separate prototype is the main source of randomness in the evaluation of its reliability. Dealing with a homogeneous group of sufficiently large number of l-dimensional systems of elements, we are within the framework of probability theory — science of mass random phenomena. Observing such group of n systems and fixing the moments of their failures $T_1, \ldots, T_n$, $T_i \in R^{l+} = [0, \infty) \times \cdots \times [0, \infty)$, $T_i = (T_{1i}, \ldots, T_{li})$, we obtain a sample of independent identically distributed random vectors.

The present work is aimed at construction of the smooth kernel estimators of the multivariate reliability function from the sample $T_1, \ldots, T_n$, investigation of their asymptotic properties for kernels of various classes, finding limiting distributions, and interval estimators of the reliability function.

1 Multivariate empirical reliability function

Let us designate by the symbol $\Rightarrow$ the convergence in distribution and by $N_1\{\mu, \sigma^2\}$ the one-dimensional random variable distributed normally with mean μ and variance σ^2, where $0 \geq \sigma < \infty$ and symbols $\mathbf{E}$ and $\mathbf{D}$ denote the mathematical expectation and variance.

Since $S(t) = P\left(T > t\right) = P\left(T_1 > t_1, \ldots, T_l > t_l\right)$, it is natural for the sample of n independent and identically distributed random vectors $\{T_i \geq 0, i = 1, \ldots, n\}$ representing periods of failure-free operation of n l-dimensional systems of elements to take as the simplest estimators

$$S_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(T_i > t\right) = \frac{1}{n}\sum_{i=1}^{n}\prod_{k=1}^{l} I\left(T_{ki} > t\right), \tag{2}$$

where $I(A)$ is an indicator of an event A. The estimator $S_n(t)$ is called the empirical reliability function.

Let us present the properties of estimator $S_n(t)$ [29]:

1. Unbiasedness: $\mathbf{E}S_n(t) = S(t)$.

2. Variance: $\mathbf{D}S_n(t) = \dfrac{1}{n}S(t)\left(1 - S(t)\right).$

3. According to the Central Limit Theorem $(S_n(t) - S(t)) \Rightarrow N_1\left\{0, S(t)(1 - S(t))\right\}.$

The estimator $S_n(t)$ has two disadvantages:

1) $S_n(t)$ is discontinuous at points $T_1, \ldots, T_n$,

2) $S_n(t) = 0$ in a region $\Omega_\infty = (T_1 > t) \bigcap \cdots \bigcap (T_n > t)$.

Let $f_n(t)$ be an estimator of l-dimensional distribution density. Then an estimator of a hazard rate function $f_n(t)/S_n(t)$, with allowance for disadvantage 2 in the region Ω_∞, is unusable [19-21].

2 Smooth kernel estimator of the reliability function and its asymptotic unbiasedness

Let us introduce the class of functions-kernels $\mathbb{S}$.

Definition 1. *The Borel l-dimensional function $V(u) = V(u_1, \ldots, u_l)$ belongs to the class $\mathbb{S}$ if $V(u)$ is a continuous strictly monotonically decreasing function by each component, such that $V(\cdot): \mathbf{R}^l \to \mathbf{R}^1$, $V(-\infty, -\infty, \ldots, -\infty) = 1$, and $V(\infty, t_2, \ldots, t_l) = V(t_1, \infty, \ldots, t_l) = \ldots = V(t_1, t_2, \ldots, \infty) = 0$.*

Define the smooth empirical reliability function as

$$\tilde{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} V\left(\frac{t - T_i}{a_n}\right) = \frac{1}{n}\sum_{i=1}^{n} V\left(\frac{t_1 - T_{1i}}{a_{1n}}, ..., \frac{t_l - T_{li}}{a_{ln}}\right), \tag{3}$$

where $V(u) \in \mathbb{S}$ and the sequences of positive numbers $a_{kn} \downarrow 0$, $k = 1, ..., l$. The function $V(u)$ is called the kernel of estimate (3). Note that $\tilde{S}_n(t)$ has no disadvantages of the estimate $S_n(t)$ (2) and $\tilde{S}_n(t)|_{a_n=0} = S_n(t)$.

As a function $V(u) = \prod_{k=1}^{l} V(u_k) \in \mathbb{S}$, we can take the product of the Laplace kernels

$$V_L(u_k) = \begin{cases} 1 - 0.5e^u_k, & -\infty < u_k < 0, \\ 0.5e^{-u_k}, & 0 \le u_k < \infty, \end{cases} \tag{4}$$

or, taking as a basis the standard normal distribution, the kernels of the form

$$V_G(u_k) = \frac{1}{2}\left[1 - \operatorname{erf}\left(\frac{u_k}{\sqrt{2}}\right)\right]$$

where $\operatorname{erf}(x) = \dfrac{2}{\sqrt{\pi}}\displaystyle\int_0^x e^{-u_k^2} du_k$ is the error function.

Let us elucidate when estimate (3) is asymptotically unbiased for $S(t)$. Further, to simplify the proofs of Lemmas and Theorems, put in (3) $a_{kn} = a_n$, $k = 1, ..., l$.

Lemma 1. *If the reliability function $S(z)$ is continuous at a point t, $V(u) \in \mathbb{S}$, and the sequence of real numbers $a_n \downarrow 0$, then*

$$\lim_{n\to\infty} \mathbf{E}\tilde{S}_n(t) = S(t). \tag{5}$$

Proof. By the definition of mathematical expectation, considering that the function $V(\cdot)$ is continuous at a point t, we have

$$\mathbf{E}\tilde{S}_n(t) = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} V\left(\frac{t-T_i}{a_n}\right)\right] = \int\limits_{\mathbf{R}^{l+}} V\left(\frac{t-y}{a_n}\right) dF(y) =$$

$$= \int\limits_{0}^{t_1} \dots \int\limits_{0}^{t_l} V\left(\frac{t-y}{a_n}\right) dF(y) + \int\limits_{t_1}^{\infty} \dots \int\limits_{t_l}^{\infty} V\left(\frac{t-y}{a_n}\right) dF(y).$$

Since

$$\lim_{n\to\infty} V\left(\frac{t-y}{a_n}\right) = \begin{cases} 0, & \text{if} \quad y < t, \\ 1, & \text{if} \quad y > t, \end{cases}$$

according to the Lebesgue dominated convergence theorem (see [30], p. 284),

$$\lim_{n\to\infty} \int\limits_{\mathbf{R}^{l+}} V\left(\frac{t-y}{a_n}\right) dF(y) = \int\limits_{t_1}^{\infty} \dots \int\limits_{t_l}^{\infty} dF(y) = 1 - F(t) = S(t). \tag{6}$$

Thus we have proved the validity of statement (5).

3 Convergence order of estimator bias with the Laplace kernel

Let us demonstrate that the estimator $\tilde{S}_{nL}(t) = \dfrac{1}{n}\sum\limits_{i=1}^{n} V_L\left(\dfrac{t-T_i}{a_n}\right)$ with the Laplace kernel $V_L(u) = \prod\limits_{k=1}^{l} V_L(u_k)$ (see (4)) has the convergence order $O\left(a_n^l\right)$ of the bias $\mathbf{b}\left(\tilde{S}_{nL}(t)\right) = \mathbf{E}\tilde{S}_{nL}(t) - S(t)$. Let $f(t) = \dfrac{\partial^l S(t)}{\partial t_1 \cdots \partial t_l}$ be the distribution density of the random vector T.

Lemma 2. *If $S(z)$ is continuous at a point t, $\sup\limits_{t\in R^{l+}} f(t) \le C < \infty$, and $a_n \downarrow 0$, then for $n \to \infty$*

$$\left|\mathbf{b}\left(\tilde{S}_{nL}(t)\right)\right| = O\left(a_n^l\right). \tag{7}$$

Proof. To prove formula (7), we take advantage of the following representation:

$$\mathbf{E}\tilde{S}_{nL}(t) = \int\limits_{\mathbf{R}^{l+}} V_L\left(\frac{t-y}{a_n}\right) dF(y) = S(t) + \int\limits_{0}^{t} V_L\left(\frac{t-y}{a_n}\right) dF(y) + \int\limits_{t}^{\infty} \left[V_L\left(\frac{t-y}{a_n}\right) - 1\right] dF(y). \tag{8}$$

Having substituted (4) into (8), we obtain

$$\mathbf{E}\tilde{S}_{nL}(t) = S(t) + \int\limits_{0}^{t} \prod_{k=1}^{l} \left[0.5e^{-\left(\frac{t_k-y_k}{a_n}\right)}\right] f(y)dy + \int\limits_{t}^{\infty} \prod_{k=1}^{l} \left[1 - 0.5e^{\left(\frac{t_k-y_k}{a_n}\right)} - 1\right] f(y)dy.$$

Changing the variables $u = \dfrac{t-y}{a_n}$ in the integrals and considering that $\sup\limits_{t\in R^{l+}} f(t) \le C$, we obtain

$$\left|\mathbf{b}\left(\tilde{S}_{nL}(t)\right)\right| \le C\frac{a_n^l}{2}\left(\prod_{k=1}^{l}\int\limits_{0}^{t_k/a_n} e^{-u_k}du_k + \prod_{k=1}^{l}\int\limits_{-\infty}^{0} e^{u_k}du_k\right) = C\frac{a_n^l}{2}\left(\prod_{k=1}^{l} e^{-t_k/a_n} + 2\right) = O\left(a_n^l\right).$$

Thus, the validity of formula (8) has been proved.

4 Asymptotic variance and MSE

Lemma 3. *If the reliability function $S(z)$ is continuous at a point t, $V(u) \in \mathbb{S}$, and the sequence of real numbers $a_n \downarrow 0$, then the variance of the smooth estimator $\tilde{S}_n(t)$ is*

$$\mathbf{D}\tilde{S}_n(t) = \frac{1}{n}S(t)\left(1 - S(t)\right) + o\left(\frac{1}{n}\right). \tag{9}$$

Proof. Indeed, by the definition of the variance, taking into account the independence of random vectors $T_1, \ldots, T_n$, arguing as in the derivation of (5), we have

$$\mathbf{D}\tilde{S}_n(t) = \frac{1}{n}\mathbf{D}V\left(\frac{t-T_1}{a_n}\right) = \frac{1}{n}\left\{\int\limits_{\mathbf{R}^{l+}} V^2\left(\frac{t-y}{a_n}\right)dF(y) - \left[\int\limits_{\mathbf{R}^{l+}} V\left(\frac{t-y}{a_n}\right)dF(y)\right]^2\right\}. \tag{10}$$

Furthermore, applying to the integrals procedures used to prove statement (6), we obtain

$$\mathbf{D}\tilde{S}_n(t) = \frac{1}{n}\left[S(t) - S^2(t) + o(1)\right] = \frac{1}{n}S(t)\left(1 - S(t)\right) + o\left(\frac{1}{n}\right).$$

It is obvious that statement (9) is also valid for the estimator $\tilde{S}_{nL}(t)$. Let us find the principal term of the asymptotic MSE for $\tilde{S}_{nL}(t)$.

Theorem 1. *If $S(z)$ is continuous at a point t, $\sup\limits_{t\in R^{l+}} f(t) \le C$, and $a_n = o\left(n^{-1/2l}\right)$, then for $n \to \infty$*

$$u^2\left(\tilde{S}_{nL}(t)\right) = \frac{1}{n}S(t)\left(1 - S(t)\right) + o\left(\frac{1}{n}\right). \tag{11}$$

Proof. Statement (11) immediately follows from the MSE representation $u^2\left(\tilde{S}_{nL}(t)\right) = \mathbf{D}\tilde{S}_{nL}(t) + \mathbf{b}^2\left(\tilde{S}_{nL}(t)\right)$ in the form of the sum of the variance and the squared bias including formulas (7) and (9).

Thus, according to (11), the principal term of the MSE for the smooth estimator $\tilde{S}_{nL}(t)$ coincides with the variance $n^{-1}S(t)\left(1 - S(t)\right)$ of the empirical reliability function $S_n(t)$ (2).

5 Asymptotic normality

Let us designate by $\{\xi_{j,n}\}_{j=1}^{n}$, $n = 1, 2, \ldots$, the sequence of independent and identically distributed random variables in the scheme of series (the distribution of the random variable $\xi_{j,n}$ depends on n).

Theorem 2. *If $S(z)$ is continuous at a point t, $\sup\limits_{t\in R^{l+}} f(t) \le C$, and $a_n = o\left(n^{-1/2l}\right)$, then for $n \to \infty$*

$$\sqrt{n}\left[\tilde{S}_{nL}(t) - S(t)\right] \Rightarrow N_1\left\{0, S(t)(1 - S(t))\right\}. \tag{12}$$

Proof. Let us represent

$$\sqrt{n}\left[\tilde{S}_{nL}(t) - S(t)\right] = \sqrt{n}\left[\tilde{S}_{nL}(t) - \mathbf{E}\tilde{S}_{nL}(t)\right] + \sqrt{n}\mathbf{b}\left(\tilde{S}_{nL}(t)\right). \tag{13}$$

It is clear that the second term in the right side of (13), according to (7), converges to zero when $n \to \infty$:

$$\sqrt{n}\mathbf{b}\left(\tilde{S}_{nL}(t)\right) = \sqrt{n}\left[o\left(n^{-1/2}\right)\right] \to 0. \tag{14}$$

Let us demonstrate that all conditions of the Central Limit Theorem in the scheme of series are satisfied for the first term in the right side of (13) (see [28], p. 435). Let we have

$$\xi_{j,n} = \frac{1}{\sqrt{n}}\left[V_L\left(\frac{t - T_j}{a_n}\right) - \mathbf{E}V_L\left(\frac{t - T_j}{a_n}\right)\right].$$

Then $\tilde{S}_{nL}(t) - \mathbf{E}\tilde{S}_{nL}(t) = \dfrac{1}{\sqrt{n}}\sum\limits_{j=1}^{n}\xi_{j,n}$. It is obvious that $\mathbf{E}\xi_{j,n} = 0$ and, considering formula (9),

$$\mathbf{E}\xi_{j,n}^2 = \frac{1}{n}\mathbf{D}V_L\left(\frac{t - T_j}{a_n}\right) < \infty.$$

Also, according to formula (9), $\lim\limits_{n\to\infty} n\mathbf{E}\xi_{1,n}^2 = S(t)\left(1 - S(t)\right)$.

Let us check the validity of the Lindeberg condition. Since $\sup\limits_{u\in\mathbf{R}^l} V_L(u) \le 1$, than for any $\tau > 0$,

$$\begin{aligned}\kappa_n = n\mathbf{E}\left(\left|\xi_{1,n}\right|^2, \left|\xi_{1,n}\right| > \tau\right) &< \frac{n}{\tau}\mathbf{E}\left|\xi_{1,n}\right|^3 \\ \le \frac{C}{\sqrt{n}}\left[\mathbf{E}\left|V_L\left(\frac{t - T_1}{a_n}\right)\right|^3 + \left|\mathbf{E}V_L\left(\frac{t - T_1}{a_n}\right)\right|^3\right] &< \frac{2C}{\sqrt{n}},\end{aligned}$$

where C is a positive constant. Hence, $\kappa_n = O\left(n^{-1/2}\right) \to 0$ for $n \to \infty$, i.e. the Lindeberg condition is satisfied. So, according to the Central Limit Theorem in the scheme of series,

$$\sum_{j=1}^{n}\xi_{j,n} \Rightarrow N_1\left\{0, S(t)(1 - S(t))\right\}.$$

Considering formula (14), we obtain statement (12).

6 Interval estimation of the reliability function

Formula (12) allows us to find transformation of smooth estimators of the reliability function that has limiting standard normal distribution. Thus, according to [31], for sufficiently large n the variance

$$\mathbf{D}\left(2\sqrt{n}\arcsin\sqrt{\tilde{S}_{nL}(t)}\right)\approx 1.$$

Note that the transformation

$$\arcsin\sqrt{\left(\sum_{i=1}^{n}V_L\left(\frac{t-T_i}{a_n}\right)+\frac{3}{8}\right)\Big/\left(n+\frac{3}{4}\right)}$$

for moderately large n provides a more stable variance [32]. Taking into account the foregoing and formula (12), if $S(z)$ is continuous at a point t, $\sup\limits_{t\in R^{l+}} f(t)\leq C$, and $a_n = o\left(n^{-1/2l}\right)$, then for $n\to\infty$

$$2\sqrt{n}\left[\arcsin\sqrt{\tilde{S}_{nL}(t)}-\arcsin\sqrt{S(t)}\right]\Rightarrow N_1\left\{0,1\right\}$$

from which the inequality follows

$$2\sqrt{n}\left|\arcsin\sqrt{\tilde{S}_{nL}(t)}-\arcsin\sqrt{S(t)}\right|<u_{1-\frac{\alpha}{2}},$$

where $u_{1-\frac{\alpha}{2}}$ is the quantile of the level $1-\frac{\alpha}{2}$ of the standard normal distribution. Thus, the interval estimator with the preset reliability $1-\alpha$ for $S(t)$ assumes the form

$$\left[\sin\left(\arcsin\sqrt{\tilde{S}_{nL}(t)}-\frac{u_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right)\right]^2<S(t)<\left[\sin\left(\arcsin\sqrt{\tilde{S}_{nL}(t)}+\frac{u_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right)\right]^2. \quad (15)$$

It is important to note that interval estimator (15) is not expressed through the unknown reliability function $S(t)$.

Conclusions

Let us list the main results of this work.

1. Such classes of kernel functions have been determined for which orders of convergence to zero can be found for the biases of smooth estimators.

2. Expressions for principal terms of the MSEs of smooth estimators with Laplace and finite kernels have been derived. It was established that the principal terms of the MSEs of such estimators coincide with the variance of the empirical reliability function that, as is well known, is the optimal nonparametric estimator of the reliability function $S(t)$.

3. The asymptotic normality of difference $\sqrt{n}\left[\tilde{S}_{nL}(t) - S(t)\right]$ and has been proved, which allows the researchers and experimenters to construct the interval estimators whose characteristics are independent of the unknown reliability function $S(t)$.

As a result of investigations, it was established that in comparison with the well-known parametric and discrete nonparametric algorithms, the advantage of the suggested estimators in calculations of the strength reliability is that they allow more reliable data to be obtained on the reliability of technical products and their residual lifetimes to be estimated. Thus, the use of smooth estimators of the reliability function allows, in particular, to obtain additional information based on smooth nonparametric estimators of the hazard rate function $\tilde{\lambda}_n(t) = f_n(t)/\tilde{S}_{nL}(t)$ when calculating the probability of failure-free operation of a pipeline from the data of strength tests of steels (see [1], p. 163). In addition, interval estimators with preset reliability can be constructed for the hazard rate function $\lambda(t)$ using the smooth estimator $\tilde{\lambda}_n(t)$ [33].

Also, the proposed algorithms and the results obtained can be used in solving the problem of increasing the reliability of various systems processing, transmitting and storing information. Here are some examples of such use:

— synthesis of better tests for new fault models;

— synthesis of logic circuits to mask faults of individual classes;

— study of temporal models of the components of information systems;

— analysis and synthesis of controllers used in modern transport systems;

— construction of mixed diagnostic tests for hybrid intelligent training and testing system [34].

Acknowledgements

This work was supported by the Russian Foundation for Basic Research, projects 13-08-00744, 13-07-98037 and was performed under the Program of Improving TSU International Competitiveness.

References

[1] Syzrantsev V. N., Nevelev Ya. P., Golofast S. L. (2008). *Calculation of the Strength Reliability of Products Based on the Methods of Nonparametric Statistics.* Nauka, Novosibirsk (in Russian).

[2] Chernyaev V. D. (1997). *System Reliability of HydrocarbonTransport Pipelines.* Nedra, Moscow (in Russian).

[3] Tikhomirov V. P. (2009). *Physical Experiment and Modeling in Mechanical Engineering.* Publishing House of Orel State Technical University, Orel (in Russian).

[4] Zhuravlev O. V., Ivanov A. V., Kurnosov V. D. et al. (2010). Estimator of Reliability of Semiconductor Radiator. *Fiz. Tekh. Poluprovodn.* Vol. **44**, no. 3, pp. 377–382 (in Russian).

[5] Herrick R. W. (2004). Failure Analysis and Reliability of Optoelectronic Devices. In: *Microelectronics Failure Analysis Desk Reference.* AMS International, Materials Park, Ohio, pp. 230–254.

[6] Soloviev A. D. (1967) Theory of Aging Elements. *Proc. 5th Berkeley Symp. on Math. Stat. Probability.* University of California Press, Berkeley and Los Angeles. Vol. **3**, pp. 313–324.

[7] Barzilovich E. Yu., Belyaev Yu. K., Kashtanov V. A. et al. (1983). *Problems of Mathematical Reliability Theory.* Radio Svyaz, Moscow (in Russian).

[8] Barlow R. E., Proschan F. (1975). *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart, and Winston, New York.

[9] Pollyak Yu. G. (1963). On Errors of Reliability Prediction Due to the Statistical Dependence Between Failures Elements. *Elektrosvyaz.* No. 4, pp. 3–9 (in Russian).

[10] Bogdanoff J. L., Kozin F. (1985). *Probabilistic Models of Cumulative Damage.* John Wiley & Sons, Inc., New York, NY.

[11] Hartman K., Letski E., Shaefer W. (1977) *Planning of Experiments in the Study of Technological Processes.* (Russ. transl). Mir, Moscow (1977) (in Russian).

[12] Koshkin G. M. (2014). Smooth Estimators of the Reliability Functions for Non-Restorable Elements. *Russian Physics Journal.* Vol. **57**, no. 5, pp. 672–681.

[13] Nadaraya E. A. (1964). Some New Estimates of Distribution Functions. *Theory Probab. Appl.* Vol. **9**, no. 3, pp. 497–500.

[14] Azzalini A. (1981). A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method. *Biometrika.* Vol. **68**, no. 1, pp. 326–328.

[15] Reiss R.-D. (1981). Nonparametric Estimation of Smooth Distribution Functions. *Scand. J. Statist.* Vol. **8**, pp. 116–119.

[16] Falk M. (1983). Relative Efficiency and Deficiency of Kernel Type Estimators of Smooth Distribution Functions. *Statist. Neerlandica.* Vol. **37**, pp. 73–83.

[17] Swanepoel J. W. H. (1988). Mean Integrated Squared Error Properties and Optimal Kernels When Estimating a Distribution Function. *Comm. Statist. Theory Methods.* Vol. **17**, no. 11, pp. 3785–3799.

[18] Jones M. C. (1990). The Performance of Kernel Density Functions in Kernel Distribution Function Estimation. *Statist. Probab. Lett.* Vol. **9**, pp. 129–132.

[19] Shirahata S., Chu I. S. (1992). Integrated Squared Error of Kernel-Type Estimator of Distribution Function. *Ann. Inst. Statist. Math.* Vol. **44**, no. 3, pp. 579–591.

[20] Kitayeva A.V., Koshkin G.M. (1997). Stable Multidimensional Intensity Function: Stable Nonparametric Estimation with Refined Convergence Rate. *Automat. and Remote Control.* Vol. **58**, no. 5, pp. 876–886.

[21] Vaal V. A., Koshkin G. M. (1999). Nonparametric Estimation of the Hazard Rate Function and its Derivatives. *Russian Physics Journal.* Vol. **42**, no. 3, pp. 362–366.

[22] Vaal V. A., Vexler A., Koshkin G. M. (2013). On Nonparametric Estimation of Hazard Function and its Derivatives. *Journal of Control and Computer Science.* Tomsk State University, Tomsk, no. 1(22), pp. 32–39 (in Russian).

[23] Sarda P. (1993). Smoothing Parameter Selection for Smooth Distribution Functions. *J. Statist. Plann. Inf.* Vol. **35**, no. 5, pp. 65–75.

[24] Altman N., Leger C. (1995). Bandwidth Selection for Kernel Distribution Function Estimation. *J. Statist. Plann. Inf.* Vol. **46**, pp. 195–214.

[25] Bowman A., Hall P., Prvan T. (1998). Trust Bandwidth Selection for the Smoothing of Distribution Functions. *Biometrika.* Vol. **85**, no. 4, pp. 799–808.

[26] Chu I. S. (1995). Bootstrap Smoothing Parameter Selection for Distribution Function Estimation. *Math. Japon.* Vol. **41**, no. 1, pp. 189–197.

[27] Shao Y., Xiang X. (1997). Some Extensions of the Asymptotics of a Kernel Estimator of a Distribution Function. *Statist. Probab. Lett.* Vol. **34**, pp. 301–308.

[28] Una-Alvarez J., Gonzalez-Manteiga W., Cadarso-Suarez C. (2000). Kernel Distribution Function Estimation under the Koziol-Green Model. *J. Statist. Plann. Inf.* Vol. **87**, pp. 199–219.

[29] Borovkov A. A. (1984). *Mathematical Statistics: Estimation of Parameters. Testing of Hypotheses.* Nauka, Moscow (in Russian).

[30] Kolmogorov A. N., Fomin S. V. (1972). *Elements of the Theory of Functions and Functional Analysis.* Nauka, Moscow (in Russian).

[31] Rao C. R. (1965). *Linear Statistical Inference and its Applications.* John Wiley & Sons, Inc., New York, NY.

[32] Anscomb F. J. (1948). The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika.* Vol. **35**, pp. 246–254.

[33] Vaal V. A., Koshkin G. M. (1998). Interval Nonparametric Estimates of Hazard Function. In: *Mathematical Modeling and Probability Theory: Collection of Scientific Works of Tomsk University*. Tomsk State University, Tomsk, pp. 147–149 (in Russian).

[34] Yankovskaya A. E., Fuks I. L., Dementyev Y. N. (2013). Mixed Diagnostic Tests in Construction Technology of the Training and Testing Systems. *International Journal of Engineering and Innovative Technology (IJEIT)*. Vol. **3**, no. 5, pp. 169–174.

About the Dual Non-parametric Control of Dynamic Systems

Bannikova A.[1], Korneeva A.[1], Kornet M.[2]
[1] Siberian Federal University, Krasnoyarsk, Russia
[2] Siberian State Aerospase University, Krasnoyarsk, Russia
e-mail: anna.korneeva.90@mail.ru

Abstract

The tasks of nonparametric identification and dual control of dynamic objects with discrete-continuous nature of the process is considered. The methods of dynamic processes modeling and control, based on the nonparametric algorithms are offered. The complexity of dynamic process modeling and control under condition of incomplete information is discussed. The purpose of the given work is to develop and investigate the algorithms of identification and control of dynamic processes by both case nonparametric and partially parametric classes of the model. The results of computing experiment are explicitly presented which show efficiency of this method for the case of solving the tasks. The scientific researches in this field will help improving the control and identification quality.

Keywords: dynamic processes, nonparametric identification, adaptive systems.

Introduction

At the present moment the parametric theory is widely spread. The problem of parametrical identification and control is investigated by different authors in particular Cypkin Ja in his theory of adaptive systems [1]. The parametrical theory based on the statistical solution is analyzed by Feldbaum A. in his publication [2]. In these works, the stage of posing the identification and control tasks of parametric structure of the dynamic process model, selected by means of different methods is defined model structure up to the parametric value.

However, the issue of identification and control should be analyzed from the point of non-parametric theory. The problems of identification and control under condition of incomplete information is very topical, because many of the dynamic processes are not deeply studied. The factor of unknown distribution random noises causes the complexity of solving the identification and control tasks. In the case of insufficiency a priori information for selecting the structure of a parametric model of the dynamic process the theory of nonparametric systems is applied [2, 3]. In comparison with the parametric theory, the nonparametric theory is applied for identification tasks if only the qualitative characteristics of the system are known.

The purpose of the given work consists in developing and researching the algorithms of identification and dual control of dynamic processes by both case: nonparametric and partially parametric classes of the model.

The tasks of the work are to develop the extended algorithms for modeling and control of dynamic objects and to carry out experimental research of the real objects and their comparison with the presented objects of the model. The main idea of this research is to reduce the problem of identification to a mathematical modeling by using nonparametric model of a regression function.

1 The level of priory information

Different levels of prior information are considered by A. Feldbaum [4]. In this paper the following levels of prior information is analyzed [2].

The level of parametric uncertainty is the first levels of prior information, which is conceded below. The parametric level of prior information means, that the parametric structure of the model and some characteristics of random noises with zero mathematical expectation and limited dispersion are known. The iterative probable procedures are used for estimating various parameters. Under these conditions the problem of identification is solved in “narrow sense”.

The following level of prior information is the level of nonparametric uncertainty. Nonparametric level of prior information doesn’t imply knowledge about this parametric model, but applies that some information of qualitative character of dynamic processes is known, for example the linearity for dynamic processes or the nature of its nonlinearity is required. The methods of nonparametric statistics are applied to the solution of the identification tasks (identification in “all-inclusive sense” [1]).

The level of parametric and nonparametric uncertainty is the level under conditions of the amount of information, which does not correspond to any of the types described above. In this respect, solving the task of identification is formulated in conditions of both case parametric, and nonparametric prior information. The models represent interdependent system, of parametric and nonparametric ratios. The solution of identification problems in this level is important from the point of practical problem solving.

2 Nonparametric identification

Let’s consider the dynamic object from the point of different levels of a priori information. The first level implies the determination of linearity of the dynamic object, but the structure of the parametric model is unknown. The order of the equation can’t be determined from a priori information.

In the second case, the dynamic process is described by the equation:

$$x_t = f(x_{(t-1)}, x_{(t-2)}, ..., x_{(t-k)}, u_t) \tag{1}$$

where $f(.)$ is unknown functional, x_t is the output variable of the process, u_t is control actions, k is the known “depth” of memory [4], which is found based on a priori information. The form of the function is not defined with the precision of parameters.

The block diagram of the simulation of the process is show in Figure 1.

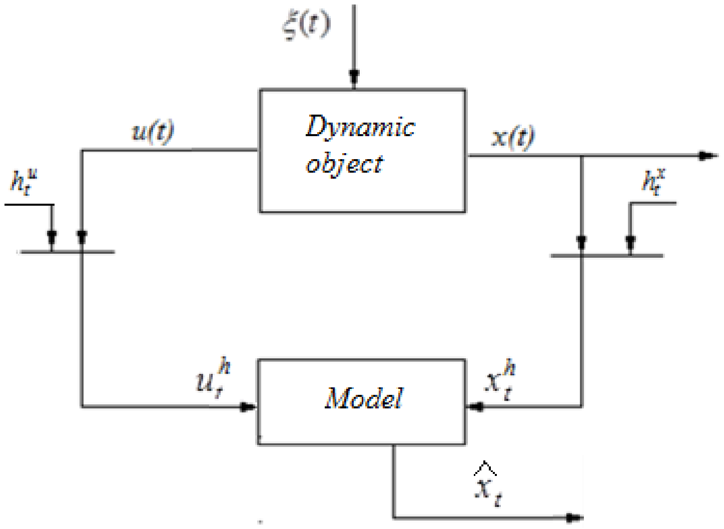


Figure 1: Block - scheme of modeling the dynamic object

The notation is accepted in Figure 1: (t) is continuous time; index t is discrete time, $\hat{x}_t$ is the output model of the object, random noise measurement h_t^x, h_t^u corresponding to the process variables, $\xi(t)$ is vector random interference.

Let the object be described by a linear differential equation of unknown order. In this case, $x(t)$ under zero initial conditions is:

$$x(t) = \int_0^t h(t-\tau)u(\tau)d\tau \tag{2}$$

where $h(t-\tau)$ is the weight function of the system, which is a derivative of the transfer function: $h(t) = k`(t)$. It is known that the inverse operator (2) is the operator [5]:

$$u(t) = \int_0^t v(t-\tau)x(\tau)d\tau \tag{3}$$

where $v(t-\tau)$ is the weight function of the object in the "output - input" direction and $v(t) = \omega`(t)$, where $\omega(t)$ is a transfer function of the system in the same direction. Therefore, the problem now is to find the weight functions $h(t)$, $v(t)$. One way of solving this problem is to measure the transient function and the evaluation of the weighting function using the results of the measurements: $\left\{x_i = k_i, t_i, i = \overline{1,s}\right\}$

The nonparametric model (2) has the form:

$$x_s(t) = \int_0^t (h_s(t-\tau), \overline{k_s}, \overline{t_s})u(\tau)d\tau \tag{4}$$

where $\overline{k_s}, \overline{t_s}$ is time vectors: $\overline{k_s} = (k_1, ..., k_s), \overline{t_s} = (t_1, ..., t_s)$, and $h_s(.)$ is:

$$h_s(t) = \frac{1}{sc_s}\int_0^t k_i H(\frac{t-t_i}{c_s})dt \tag{5}$$

where $H(.)$ is a bell-shaped (nuclear) function, c_s is a blur parameter satisfying, the certain conditions of convergence [4].

3 Nonparametric dual control

The deficiency a priory information results in the necessity to combine learning and controlling the object. This type of control is called the dual control. The problem of the dual control was investigated by A.Feldbaum. This parametrical theory was developed based on the theory of statistical solution in this publication [2]. The given paper presents the analysis of developing the dual control by using nonparametric dual control theory [3].

Let's consider the dynamic object by both case nonparametric and partially parametric classes of the model. As described bellow, the first case implies the determination of linearity of the dynamic object, where the structure of the parametric model is unknown and the order of the equation can't be determined from a priori information. In the second case, the dynamic process is described by the equation (1).

The block diagram of the control process is shown in Figure 2.

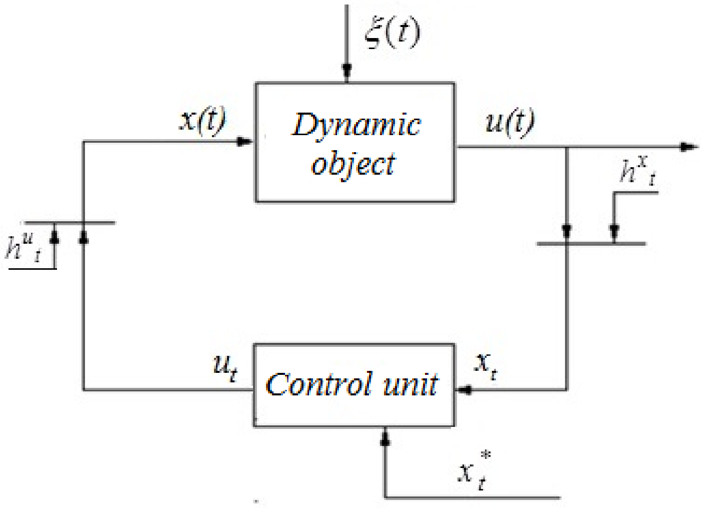


Figure 2: Control scheme of a dynamic object

The notation is accepted in Figure 2: x^*_t is task for the control unit. The unknown operation A of an object describes the processes, i.e. $x(t) = A < u(t) >$ where $x(t)$ is the output variable of the process, $u(t)$ is a set of control actions. If the operation A^{-1} is defined: $AA^{-1} = 1$, then:

$$A^{-1}x(t) = a^{-1}, A < u(t) >, u(t) = A^{-1}x(t) \tag{6}$$

Setting the trajectory $x(t) = x^{*}(t)$, the ideal value $u^{*}(t)$ is found from (12). In this case, the operation A was found from the nonparametric models (4), and the inverse operator A is the operator A^{-1} and can be found from the equation (3). Then, the inverse operator A is a set of control impact:

$$u^{*}(t) = \int_{0}^{t} \frac{1}{sc_s} \sum_{j=1}^{s} \omega_j H'(\frac{t-\tau-t_i}{c_s}) x^{*}(\tau) d\tau \tag{7}$$

where $x^{*}(\tau)$ is a task control. The integral of equation (13) is taken numerically. The unknown operations A and A^{-1} are calculated by using the weight function and the transfer function of the system in the class of nonparametric statistic [4], because the equation of the processes is unknown. Then, nonparametric dual control algorithms of linear dynamic system has the form:

$$u_{s+1} = u_s^{*} + \delta u_{s+1} \tag{8}$$

where u_s^{*} is (13), and $\delta u_{s+1} = \epsilon(x_{s+1}^{*} - x_s)$ is the “search step”. The duality of the algorithm is exhibited here. If the structure of the dynamic process may be described by the partially parametric classes of the model, i.e. the process is described by the equation $x(t) = f(x(t-1), x(t-2), ..., x(t-k), u(t))$, where k is known, then u_s^{*} is described by the nonparametric evaluation of the regression function using the results of the measurements $\{x_i, u_i, i = \overline{1,s}\}$

$$u_s^{*} = \frac{\sum_{i=1}^{s} u_i \Phi(\frac{x_{s-j}^{*} - x_{i-j}}{c_s}) \prod_{j=1}^{k} \Phi(\frac{x_{s-j} - x_{i-j}}{c_s})}{\sum_{i=1}^{s} \Phi(\frac{x_{s-j}^{*} - x_{i-j}}{c_s}) \prod_{j=1}^{k} \Phi(\frac{x_{s-j} - x_{i-j}}{c_s})} \tag{9}$$

where $\Phi(.)$ is a bell-shaped (nuclear) function, and c_s has the form:

$$c_s = \alpha \left| x_s - x_{s-1}^{0} \right| \tag{10}$$

where x_{s-1}^{0} - the closest element to x_s.

4 The computation experiment

The verification of the nonparametric identification and control algorithms is carried out by statistical modeling. For the purpose of computational experiment the object is described by equations of the form: $x_t = 0.4x_{t-1} - 0.3x_{t-3} + u_t$, where x_t is the output variable of the process, u_t is the input process variable.

Transient response of the object is shown in Figure 3. The input control variable is defined by the equation: $u(t) = sin(0.5t)$. The model of the object is constructed by using a non-parametric model (4). The simulation results are shown in Figure 4.

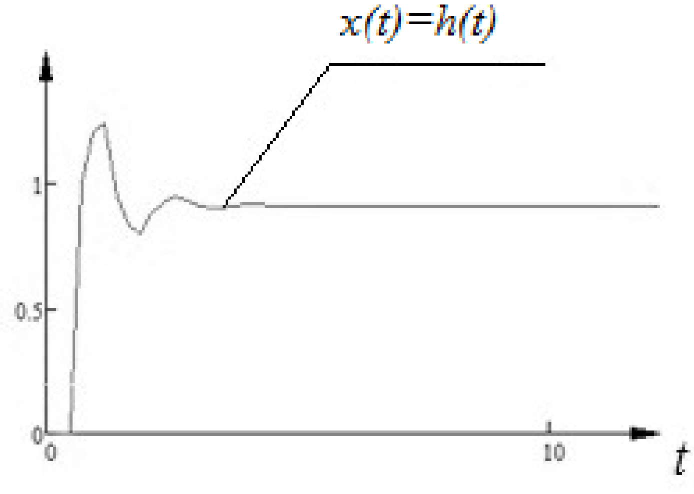


Figure 3: The transfer function of the dynamic process

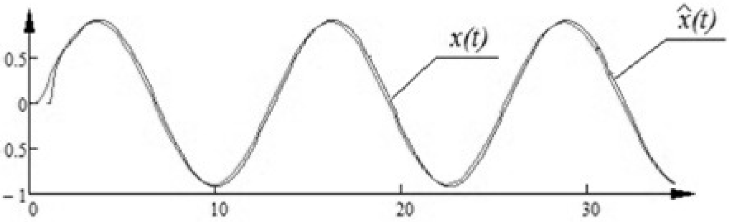


Figure 4: Results of the identification process using the model (4)

The notation is accepted in Figure 4, where $x(t)$ is output of the object, $\hat{x}(t)$ is output of the model. The square error of the simulation is 0.015. The model of the object is constructed by using a non-parametric model (10).

The use of this model is acceptable, when the parametric structure of the object is partially known. The simulation results are presented in Figure 5.

The square error of modeling is 0,023.

The comparison of nonparametric dual control algorithms with the typical control algorithms defined as PI-algorithms is represented in the computational experiment. The amount of sampling (u_i, x_i) is 100. The control results are shown in Figure 6, when the task control is stepwise impact:

In Figure 6 the notation is accepted: $x(t)$ is output of the object, when the control unit is a nonparametric dual control regulator, $\overline{x}(t)$ is output of the object, when the control unit is the PI regulator, $x^*(t)$ is a control task. The square error of the control for the nonparametric regulator it is 0.07, for the PI regulator is 0.34.

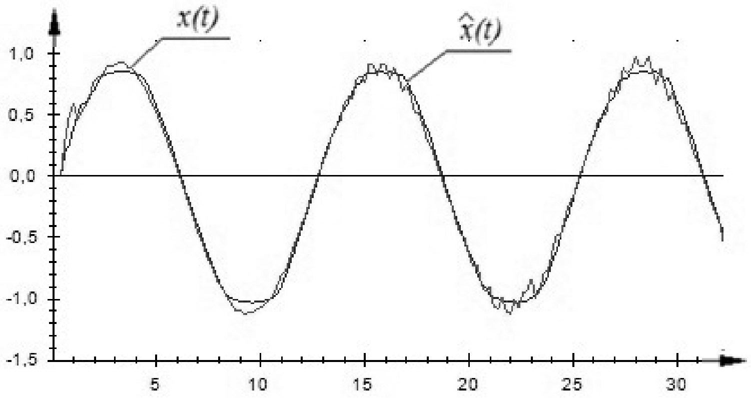


Figure 5: The results of the identification process using the model (10)

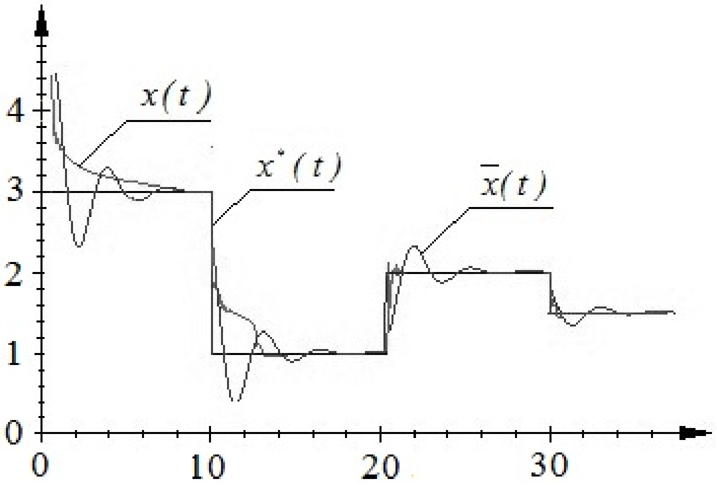


Figure 6: The control results, when the task control is a stepwise impact

5 Conclusion

In the article the analysis of algorithms for nonparametric identification and control under condition of non-parametric uncertainty is carried out, i.e. the case where a priori information about the object is small and do not allow choosing the parametric model of the object. In this case, the Duhamel integral is used for describing the process. The problem is reduced to the solution of nonparametric estimation of the weight function of the system because of the observations “input-output” of the object. The non-parametric algorithms under partial nonparametric uncertainty are shown in the computational experiment.

Acknowledgements

The authors wish to express their gratitude to the supervisor, Prof. Dr. Medvedev A.V. who was abundantly helpful and offered invaluable assistance, support and guidance.

References

[1] Cypkin Ja. (1968). *Adaptation and training in automated systems*. Nauka, Moscow.

[2] Medvedev A.V. (2010). The theory of nonparametric systems. Modeling.*Vestnik SibGAU*. Vol. **30**, pp. 4-9.

[3] Medvedev A.V. (2010). The theory of nonparametric systems. Processes.*Vestnik SibGAU*. Vol. **29**, pp. 4-9.

[4] Nadaraya E. (1983). *Non-parametric estimation of the probability density and the regression curve.* . izd. Tbil. un-ta, Tbilisi.

Adaptive Regression Estimates. Semi-nonparametric Models

V.A. Simakhin and O.S. Cherepanov
Kurgan State University, Kurgan, Russian Federation
e-mail: sva_full@mail.ru, ocherepanov@inbox.ru

Abstract

In the present work, semi-nonparametric estimates of regression by the weighted maximum likelihood method are considered. Their efficiency is investigated for a class of distributions of residues with different degrees of tail stretching in the presence of outliers described by the Tukey model. It is demonstrated that the given estimates are efficient.

Keywords: Robust, Adaptive Estimates, Weighted Maximum Likelihood Method, Nonparametric, Semi-Nonparametric, Regression.

Problem formulation. Introduction

Let us consider a problem of local regression. Let we have the vector $x = (x_1, \dots, x_p)^T \in X \in R_p$ of independent variables and the dependent variable $y = m(x) + \varepsilon$, where $m(x)$ is an unknown regression function and T denotes transposition. It is required to find the regression estimate $m(x_0)$ at any point $x_0 \in X$ from the available independent observations $(x_i, y_i), i = 1, \dots, N$.

Let us designate by $G_1(x) \subset I_1 \leftrightarrow g_1(x)$ the *a priori* distribution function and the density of random vector x; $G_2(\varepsilon) \subset I_2 \leftrightarrow g_2(\varepsilon)$ the *a priori* distribution function and the density of the random variable ε independent of x; $F(x, y) \subset I_3, \leftrightarrow f(x, y)$, $F(y/x) \subset I_4 \leftrightarrow f(y/x)$ actual experimental joint and conditional distribution functions and the density of random vector $z = (x^T, y)^T$; $m(x) \subset I_5$; $F_N = (x, y)$ and $F_N(y/x)$ the empirical distribution functions of the vector z and the nonparametric estimate $F(y/x)$.

Let us consider below the following classes $\langle I \rangle = (I_1, I_2, I_3, I_4, I_5)$: class I_1 of distributions with finite Fisher information; class I_2 of unimodal symmetric distributions with finite Fisher information; class I_3 of distributions of the form (the Tukey supermodel):

$$F(x, y) = [(1 - p)G_2(y - m(x_0)) + pH(y - m(x_0))]\, G_1(x), \tag{1}$$

where p is the fraction of outliers and H_1 is their distribution; without loss of generality, we consider $x = x_1 \in R_1$.

All methods and algorithms for solving the given problem depend on the *a priori* information on the classes $\langle I \rangle = (I_1, I_2, I_3, I_4, I_5)$ we have. The intersection of the given types of *a priori* information $\langle I \rangle$ determines also a concrete algorithm for finding estimates.

The problems with unknown $m(x)$ and $G_1(x)$ are conventionally referred to the class of nonparametric problems, but for supermodels Eq. (1), a number of robust nonparametric problems arise in which a part of supermodels can be determined parametrically, and another part nonparametrically - a class of semi-nonparametric models.

Classification by levels of *a priori* information allows $\langle I \rangle$ semi-parametrical estimates of the regression to be ordered as already considered in the literature, and new estimates to be synthesized for scanty *a priori* information $\langle I \rangle$.

Our analysis of the literature shows that by the present time, a large number of robust nonparametric regression algorithms have been synthesized or heuristically suggested (for example, see [1-12]; these references include only works supplied with extensive bibliography on the subject). The classical nonparametric Nadaraya-Watson regression estimates (see references in [1] and [2]) were nonrobust; therefore, by analogy with them, nonparametric estimates of the conditional median and conditional mode were suggested [1]. Later on, the synthesis of the robust nonparametric regression estimates was mostly based on local adaptation methods (LAM) [3],[4] and local maximum likelihood method (LMLM) [12] using conditional M-estimates. Conditional R- and L- estimates [17] are not widespread.

According to the ideas of robust statistics [3], [4], [7], [14], the estimators for robust estimates are determined based on minimax solutions. Ya. Z. Tsypkin [4] called **optimal on a class** such robust estimates synthesized on the basis of the entropy criterion. However, as demonstrated investigations, such **estimates optimal on a class** can have amazingly low efficiency for a concrete situation. Exactly this fact stimulates a search for adaptive estimates adjusted to search for an optimal solution in a given concrete situation [8], [9], [13] based on distributions, outliers, and bandwidth parameter [3], [11]. An analysis of *a priori* information $\langle I \rangle = (I_1, I_2, I_3, I_4, I_5)$ [8], [8], [13]shows that the adaptive algorithm must be adjusted based on the form of *a priori* information on the distribution $G_2(\varepsilon) \subset I_2$ (global adaptation) and on the shape of outliers in supermodel Eq. (1) which, as a rule, takes into account unknown information on the shape and fraction of the outliers (local adaptation). It is important to note that adaptation algorithms must be nonparametric inherently [8].

In the present work, adaptive semi-nonparametric estimates of local regression are synthesized based on the weighted maximum likelihood method [8], [9] for different levels of *a priori* information. This work further develops our work [13] in which semi-parametrical global regression estimates were synthesized.

1 Weighted maximum likelihood method. Adaptation

1.1 Weighted maximum likelihood method

Let us consider the problem of a search for a local estimate $m_N(x_0)$ for the Tukey supermodel given by Eq. (1). For this purpose, we take advantage of the LMLM [12].

Equations for LMLM estimates in our case are written in the form:

$$\sum_{i=1}^{N} \Psi(z_i, m_N(x_0)) K_1 \left[\frac{x_0 - x_i}{h_{1N}}\right] = 0, \tag{2}$$

where $z = (x, y)^T$, $\Psi(z, m(x_0))$ is the estimator, $K_1(u)$ is the kernel function [2], [3], [8] $K_1(-u) = K_1(u)$, $\int K_1^2(u)du < \infty$, $\int uK_1(u)du = 0$, h_{1N} is the bandwidth parameter.

The estimator $\Psi(z, m(x_0))$ from Eq. (2) can be represented in the following a form:

$$\Psi(z, x_0) = \left[\frac{\frac{\partial}{\partial m} g_2(y - m(x_0))}{f(z)}\right] = \left[\frac{\frac{\partial}{\partial m} g_2(y - m(x_0))}{g_2(y - m(x_0))}\right] \cdot \left[\frac{g_2(y - m(x_0))}{f(z)}\right] =$$

$$= \frac{\partial}{\partial m} \ln g_2(y - m(x_0)) \cdot \left[g_2(y - m(x_0))\right]^l, \quad l = \frac{\ln f(\cdot) + \ln g_2(\cdot)}{\ln g_2(\cdot)},$$

$$\Psi(z, m(x_0)) = U(g_2) \cdot W(F), \tag{3}$$

where $U(g_2) = \frac{\partial}{\partial m} \ln g_2(y - m(x_0))$ is the function of contribution of the conditional *a priori* distribution $g_2(y - m(x_0))$, $W = [g_2(y - m(x_0))]^l$ is the weight function, l is the radical parameter responsible for information on deviation of the actual distribution F from the *a priori* (ideal) distribution G_2, concentrating all information on outliers in **one**, in principle unknown, radical parameter.

Hence, LMLM estimator Eq. (3) can be written in terms of the WMLM estimator in the form

$$\sum_{i=1}^{N} \frac{\partial}{\partial m} \ln g_2(y_i - m_N(x_0)) \cdot \left[g_2(y - m_N(x_0))\right]^l K_1 \left[\frac{x_0 - x_i}{h_{1N}}\right] = 0. \tag{4}$$

1.2 Properties of estimates by the weighted maximum likelihood method

Let us represent Eq .(4) in the form of a functional of the empirical distribution function $F_N(z)$:

$$\int \tilde{\Psi}(z, m(x_0)) dF_N(z) = 0, \tag{5}$$

$$\tilde{\Psi}(z, m(x_0)) = \frac{\partial}{\partial m} \ln g_2(y - m(x_0)) \cdot \left[g_2(y - m(x_0)\right] K_1 \left[\frac{x_0 - x}{h_{1N}}\right]^l.$$

Let us take advantage of the standard approach to investigation of the conditional M-estimates [1]-[3],[5], [6], [11], [15]. Omitting intermediate routine calculations, we can prove that:

$$\sqrt{Nh_N}(m_N(x_0) - m(x_0)) \Rightarrow N(b_N, V_N),$$

$$b_N = \frac{h_N^2}{2} \cdot \frac{d^2}{dx}\, m(x)|_{x=x_0} \cdot \int u^2 K(u)du,$$

$$V_N = \frac{\int K^2(u)du}{Nh_N g_1(x_0)} \frac{\int [\psi(y - m(x_0)]^2 dG_2(y/x_0)}{\left\{\int \frac{\partial}{\partial m}\psi(y - m(x_0) dG_2(y/x_0)\right\}^2}.$$

For local linear regression models, the matrix variant of the results presented above is used, for example, by analogy with [3], [5], [7].

1.3 Adaptation

In the present work, we consider two aspects of the adaptation of estimates: by the form of *a priori* distribution G_2 and by the form of the distribution of outlier fraction (p, H). The idea of global adaptation by the form of *a priori* distribution G_2 was suggested by Beran [16] in the middle 70s. It is reduced to application of the nonparametric estimate of the Rozenblatt-Parzen density. We note that from Eq. (4) at $l = 0$ we obtain LMLM estimates, at $l = 1$ maximum stability estimates (MSE), and at $l = 0.5$ radical estimates on the Hellinger distance [7]. This fact allows estimates robust to outliers in the given situations to be obtained using adaptation by the radical parameter l $(0 \leq l \leq 1)$. As a measure of robustness, it is reasonable to use the variation coefficient of the estimate V. The standard approach to construction of adaptive robust estimates is reduced to assignment of a supermodel for which a certain characteristic, for example, V is determined as a function of *a priori* distributions and (p, H). Then Hogg selectors are determined by which adaptation is carried out in the form of sample truncation operation [14]. Analogous procedures are performed for the Meshalkin exponentially weighted (EW) estimates and Shurygin stable estimates [7]. The given approach calls for complex theoretical research, but any critic has the right to declare that the ideal supermodel is close to reality, and construction of estimates already **robust** on supermodels is required. A nonparametric approach to the construction of an algorithm of adaptation by outliers is required. Such nonparametric approach - bootstrap procedures - has already been used for a long time. Taking advantage of the bootstrap method, local estimates $V(l)(0 \leq l \leq 1)$ can be obtained, and the optimal value of the radical parameter l^* can be found. Computational difficulties in implementation of the bootstrap procedures are not principal for modern PC.

2 Adaptive regression estimates by the weighted maximum likelihood method

2.1 Adaptive nonparametric estimates of correlation analysis

Classical algorithms of nonparametric regression were synthesized in the form of estimates of the location parameter θ for the conditional distribution $F(y/x_0)$: average, median, and mode [1], [2]. Let $F(y/x_0)$ be the distribution function symmetric about θ:

$$F(y/x_0) = 1 - F((2\theta - y)/x_0).$$

Estimates θ_N based on the functional $\int \varphi(1 - F_N(2\theta_N - t/x_0))dF_N(t/x_0) = 0$ lead to the nonparametric R-estimates of the regression [17]. The given classical estimates have been investigated sufficiently well. Let $F(y/x_0)$ be symmetric about θ. To determine θ_N, we use the WMLM. We obtained the following estimator for θ_N:

$$\sum_{i=1}^{N} \frac{d}{d\theta} f_N(y_i, \theta_N/x_0) \left[f_N(y_i, \theta_N, /x_0)\right]^{l-1} K_1\left(\frac{x_0 - x_i}{h_{1N}}\right) = 0,$$

where $f_N(y_i, \theta_N/x_0)$ is the Rozenblatt-Parzen symmetrized estimate.

2.2 Adaptive regression estimates by the local approximation method

In this section we consider examples of construction of local regression WMLM estimates when the distribution $G_2(\varepsilon)$ is set parametrically. Because the properties of the estimates depend on the degree of stretching of the distribution tail, four distributions are analyzed with short, intermediate, and long tails.

2.2.1 t-Student distribution

Let $G_2(\varepsilon)$ belong to the class of t-Student distributions with zero location parameter, ν degrees of freedom, and probability density

$$g_2(x, s, v) = \frac{B(v)}{s} \frac{1}{D^m(x/s)}, m = \frac{v+1}{2}, D(u) = 1 + \frac{u^2}{v}. \tag{6}$$

Substituting Eq. (6) into Eq. (4), we obtain the WMLM estimators:

$$\sum_{i=1}^{N} \frac{y_i - m_N(x_0)}{D^{lm+1}\left(y_i - m_N(x_0)\right)} K_1\left(\frac{x_0 - x_i}{h_{1N}}\right) = 0. \tag{7}$$

The random variable $\sqrt{Nh_N}(m_N(x_0) - m(x_0))$ has asymptotically normal distribution with mean b_N and variance V_N of the form:

$$b_N = \frac{h_N^2}{2} \frac{d^2}{dx}\, m(x)|_{x=x_0} \cdot \int u^2 K(u)du,$$

$$V_N = \frac{\int K^2(u)du}{Nh_N g_1(x_0)} \cdot \frac{\int \frac{(y-m(x_0))^2}{D^{2lm+2}(y-m(x_0))} dG_2(y/x_0)}{\left\{\int \left(1 - \frac{2(lm+1)}{v} \frac{(y-m(x_0))^2}{s^2 \cdot D(y-m(x_0))}\right) dG_2(y/x_0)\right\}^2}.$$

For the Cauchy distribution ($v = 1$), Eq. (7) can be represented in the following form:

$$\sum_{i=1}^{N} (y_i - m_N(x_0)) g_2^{l+1}(y_i - m_N(x_0)) K_1\left(\frac{x_0 - x_i}{h_{1N}}\right) = 0. \tag{8}$$

2.2.2 Generalized normal distribution

Let $G_2(\varepsilon)$ belong to the family of generalized normal distributions with zero location parameter and probability density of the form

$$g_2(x) = \frac{\beta}{2s\Gamma(1/\beta)} e^{-\left(\frac{|x|}{s}\right)^{\beta}}. \tag{9}$$

Substituting Eq. (9) into Eq. (4), we obtain the WMLM estimators

$$\sum_{i=1}^{N} Sign(y_i - m_N(x_0))|y_i - m_N(x_0)|^{\beta-1} g_2^l(y_i - m_N(x_0)) K_1\left(\frac{x_0 - x_i}{h_{1N}}\right) = 0. \tag{10}$$

The random variable $\sqrt{Nh_N}(m_N(x_0) - m(x_0))$ has asymptotically normal distribution with mean b_N and variance V_N of the form:

$$b_N = \frac{h_N^2}{2} \cdot \frac{d^2}{dx}\, m(x)|_{x=x_0} \cdot \int u^2 K(u)du,\ V_N = \frac{\int K^2(u)du}{Nh_N g_1(x_0)} \times$$

$$\times \frac{\int |y - m(x_0)|^{2\beta-2} g_2^{2l}(y - m(x_0)) dG_2(y/x_0)}{\left\{\int |y - m(x_0)|^{\beta-2}\left(\beta - 1 - l \cdot \beta\left(\frac{y-m(x_0)}{s}\right)^{\beta}\right) g_2^{l_2}(y - m(x_0)) dG_2(y/x_0) + 2|0|^{\beta-1} g_2^l(0) g_2(0/x_0)\right\}^2}$$

Let us study some special cases of generalized normal distributions.

1. Laplace distribution ($\beta = 1$). The WMLM estimators for the semi-nonparametric regression estimate from Eq. (10) is

$$\sum_{i=1}^{N} Sign(y_i - m_N(x_0)) g_2^l(y_i - m_N(x_0)) K_1\left(\frac{x_0 - x_i}{h_{1N}}\right) = 0. \tag{11}$$

2. Normal distribution ($\beta = 2$). The WMLM estimators for the semi-nonparametric regression estimate from Eq. (10) is

$$\sum_{i=1}^{N} (y_i - m_N(x_0)) g_2^l(y_i - m_N(x_0)) K_1\left(\frac{x_0 - x_i}{h_{1N}}\right) = 0. \tag{12}$$

3. Generalized normal distribution of the 4th degree distribution ($\beta = 4$).
   The WMLM estimators for the semi-nonparametric regression estimate from Eq. (10) is

$$\sum_{i=1}^{N} (y_i - m_N(x_0))^3 g_2^l(y_i - m_N(x_0)) K_1\left(\frac{x_0 - x_i}{h_{1N}}\right) = 0. \tag{13}$$

2.3 Adaptive nonparametric regression estimate

Let $G_2(\varepsilon)$ belong to the nonparametric class of unimodal distributions symmetric about zero. To estimate the unknown $g_2(\varepsilon)$, we take advantage of the symmetrized estimate of the Rozenblatt-Parzen density of the form

$$g_{2N}(\varepsilon) = \frac{1}{2Nh_{2N}} \sum_{i=1}^{N} \left(K_2\left(\frac{\varepsilon - \varepsilon_i}{h_{2N}}\right) + K_2\left(\frac{\varepsilon + \varepsilon_i}{h_{2N}}\right)\right), \tag{14}$$

where N is the sample length, h_{2N} is the bandwidth parameter, and K_2 is the kernel function. Substituting Eq. (14) into Eq. (4), we derive the WMLM estimator for the regression parameters of the following form:

$$\begin{gathered}\sum_{i=1}^{N} g_{2N}^{l_2-1}(y_i, m(x_0)) K_1\left(\frac{x-x_i}{h_{1N}}\right) \sum_{\substack{j=1 \\ i \neq j}}^{N} \left(\gamma\left(\frac{2m(x_0)-y_i-y_j}{h_{2N}}\right) K_2\left(\frac{2m(x_0)-y_i-y_j}{h_{2N}}\right)\right) = 0, \\ \gamma(x) = \frac{1}{K_2(x)} \frac{\partial}{\partial x} K_2(x). \end{gathered} \tag{15}$$

3 Modeling

3.1 Experiment description

The efficiency of regression estimates Eqs. (8), (11), (12), (13), and (15) was studied for the following class of symmetric noise distributions with different degrees of tail stretching: generalized normal distribution of the 4th degree (D4), normal distribution (ND), Laplace distribution (LD), Cauchy distribution (CD)) without outliers (WO) and with outliers for the Tukey model of symmetric outliers (SO) and asymmetric outliers (AO) along the y axis($p = 0.1$). The scaling parameter of each distribution from the examined class was determined so that the distribution quantile on a level of 0.95 coincided with the quantile on a level of 0.95 for standard normal distributions. The following regression function was considered:

$$m(x) = 2.5e^{-\frac{(x+1)^2}{4}} - 2.5e^{-\frac{(x-1)^2}{4}} + x.$$

The efficiency of estimates was calculated in the following form:

$$\xi = \frac{V_{\min}}{V},$$

where V is the conditional variation of the examined regression estimate and V_{min} is the minimum conditional variation among the examined estimates. To find the conditional variation of the estimate, the Monte Carlo method was used. For this purpose, we formed $M = 500$ two-dimensional samples (X, Y) from distributions

$F_1(x)$ and $F_2(\varepsilon)$ at the point in the vicinity of x_0 whose width was determined by the parameter h_{1N} using the regression function $m(x)$ and random number generators.

The adaptive nonparametric estimates (ANE) given by Eq. (15) were compared with the adaptive semi-nonparametric estimates (ASE: AECD Eq. (8), AELD Eq. (11), AEND Eq. (12), and AED4 Eq. (13)) with maximum likelihood (MLE), Nadaraya-Watson ((N-W)E), conditional median (ME), radical (RE), and maximum stability estimates (MSE).

3.2 Results

Table 1: Efficiency of estimates on the D4 distribution at the point $x_0 = 0$

Estimates	MLE	(N-W)E	ME	RE	MSE	ASE	ANE
WO	**1.00**	0.73	0.37	0.78	0.43	**1.00**	0.82
AO	0.00	0.01	0.11	0.78	0.36	**1.00**	0.86
SO	0.08	0.50	0.52	**1.00**	0.68	**1.00**	0.73

Table 2: Efficiency of estimates on the normal distribution at the point $x_0 = 1.5$

Estimates	MLE	(N-W)E	ME	RE	MSE	ASE	ANE
WO	1.**00**	1.00	0.58	0.73	0.50	**1.00**	0.81
AO	0.01	0.01	0.21	0.79	0.54	0.92	**1.00**
SO	0.62	0.62	0.79	0.97	0.73	**1.00**	0.80

Table 3: Efficiency of estimates on the Laplace distribution at the point $x_0 = 0.949$

Estimates	MLE	(N-W)E	ME	RE	MSE	ASE	ANE
WO	**1.00**	0.85	**1.00**	0.80	0.68	**1.00**	0.83
AO	0.30	0.01	0.30	0.71	0.56	0.79	**1.00**
SO	**1.00**	0.47	**1.00**	0.87	0.73	**1.00**	0.78

3.2.1 Efficiency of estimates on distribution classes

The average and minimal efficiency of the suggested estimates were investigated on classes of distributions of residues $S = \langle D4, ND, LD, CD \rangle$ with and without symmetric and asymmetric Tukey outliers. The following classes were considered: class S_1 of distributions of residues without outliers, class S_2 of distributions of residues with

Table 4: Efficiency of estimates on the Cauchy distribution at the point $x_0 = 0$

Estimates	MLE	(N-W)E	ME	RE	MSE	ASE	ANE
WO	0.83	0.01	0.88	0.77	0.66	0.83	**1.00**
AO	0.92	0.00	0.30	0.94	0.80	**1.00**	0.97
SO	0.81	0.01	0.81	0.81	0.71	0.84	**1.00**

asymmetric outliers, and class S_3 of distributions of residues with symmetric outliers. Adaptive nonparametric estimate (ANE) Eq. (15) was compared with the following estimates: maximum likelihood estimate (MLED4) provided that the *a priori* residues obeyed the distribution of the fourth degree; Nadaraya–Watson estimate ((N-W)E), radical estimate (REND), and maximum stability estimate (MSEND) provided that the *a priori* distribution of the residues was normal; and conditional median estimate (M). The average and minimal efficiencies of estimates were calculated for each class (Tables 5-7).

Table 5: Average and minimal efficiencies of estimates on class S_1

Estimates	MLED4	(N-W)E	ME	MLECD	REND	MSEND	ANE
Average efficiency	0.53	0.70	0.57	0.56	**0.80**	0.57	0.76
Minimal efficiency	0.00	0.06	0.38	0.28	0.47	0.30	**0.64**

Table 6: Average and minimal efficiencies of estimates on class S_2

Estimates	MLED4	(N-W)E	ME	MLECD	REND	MSEND	ANE
Average efficiency	0.00	0.01	0.20	0.65	0.85	0.59	**0.91**
Minimal efficiency	0.00	0.00	0.15	0.46	0.62	0.42	0.67

Conclusions

1. Based on the WMLM, new semi-nonparametric adaptive estimates of the local regression are synthesized for different levels of *a priori* information (Eqs. (7), (8), (10), (11), (12), (13), and (15));

2. Adaptive semi-nonparametric estimates (ASE) on the level of *a priori* information are effective;

3. MLE (MLED4, (N-W)E, and M) on the class of distributions have low or zero efficiency (Tables 1 - 4);

Table 7: Average and minimal efficiency of estimates on class S_3

Estimates	MLED4	(N-W)E	ME	MLECD	REND	MSEND	ANE
Average efficiency	0.07	0.45	0.70	0.72	**0.99**	0.75	0.75
Minimal efficiency	0.00	0.03	0.50	0.65	**0.95**	0.72	0.69

4. Classical robust estimates (M and MSEND) for particular distributions have low efficiency (Tables 1 - 4);

5. Adaptive nonparametric estimate (ANE) on nonparametric level of *a priori* information is effective.

References

[1] Medvedev F. V. (1983). *Nonparametric Adaptation Systems*. Nauka, Novosibirsk.

[2] Hardle V. (1993). *Applied Nonparametric Regression*. Mir, Moscow.

[3] Tsybakov A. B. (1986). Robust reconstruction of functions by the local approximation method. *Probl. Pered. Inform.*. Vol. **22**, pp. 69-84.

[4] Katkovnik V. Ya. (1985). *Nonparametric Identification and Smoothing of Data*. Nauka, Moscow.

[5] Vasil'ev V. A., Dobrovidov A. V , Koshkin G. M. (2004). *Nonparametric Estimate of Functionals of Distributions of Stationary Sequences*. Nauka, Moscow.

[6] Shurygin A. M . (2000). *Applied Statistics. Robustness. Estimate. Prediction.* Financy i Statistika, Moscow.

[7] Simakhin V. A. (2011). *Robust Nonparametric Estimates*. LAMBERT Academic Publishing, Germany.

[8] Simakhin V. A. (2006). Nonparametric Robust Regression Estimates. *Proceedings SPIE*. Vol. **6522**, pp. 130-139.

[9] Tamine J. (2001) *Smoothed Influence Function: Another View at Robust Nonparametric Regression*. Humboldt-Universität, Berlin.

[10] Tsypkin J. Z. (1984). *Principles of the Information Theory of Identification*. Nauka, Moscow.

[11] Lepski O. V., Levit B. Y. (1999). Adaptive Nonparametric Estimation of Smooth Multivariate Functions *Math. Methods Statist.*. Vol. **8**, pp. 344-370.

[12] Gallant A. R., Nychka D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica.* Vol. **55**, pp. 363-390.

[13] Simakhin V. A., Cherepanov O. S.(2014). Robust semi-parametric regression estimates. *Information Technologies and Mathematical Modeling Communications in Computer and Information Science.* Vol. **487**, pp. 397-405.

[14] Shulenin V. P. (2012). *Mathematical Statistics. Robust Statistics. Part 3.* Publishing House of Scientific and Technology Literature, Tomsk.

[15] Fan J. Q., Hu T. C., Truong Y. K. (1994). Robust nonparametric function estimate. *Scandinavian Journal of Statistics.* Vol. **21**. pp. 433-446.

[16] Beran R.(1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* Vol. **6**, pp. 292-313.

[17] Cheng P. T., Cheng K. F.(1990). Asymptotic normality for robust R-estimators of regression function. *Journal of Statistical Planning and Inference.* Vol. **24**, pp. 137-149.

On Adaptive Estimation Using a Prior Guess

Yuri G. Dmitriev and Peter F. Tarassenko
Department of Applied Mathematics and Cybernetics,
Tomsk State University, Tomsk, Russia
e-mail: dmit@mail.tsu.ru, ptara@mail.tsu.ru

Abstract

Statistical estimators of a linear functional of an unknown distribution are considering based on combined estimator in the form of weighted sum of nonparametric estimator and a prior guess about the value of this functional. The optimal (in terms of mean square error) weighting coefficient is subject of adaptive estimation itself. A series of k–adaptive estimators are constructed by using the prior guess recursively k times. Examples of combined estimators and results of numerical calculations are provided, that illustrates how the difference between prior guess and unknown value of functional affects the limit distributions of estimators and their probabilistic characteristics.

Keywords: linear functional, prior guess, a priori information, combined estimator, nonparametric estimator, k - adaptive combined estimator.

Introduction

The term 'prior guess' has been probably first introduced by Ferguson [11] and used later in various contexts. There are many papers in the literature devoted to the estimation of the probability characteristics with using additional information (prior guess). Combined statistical estimators adapting a prior guess and their properties have been considered in [2], [8], [9], [10], [17]. Estimators of the mean were proposed in [1], [3], [13], [18]. Estimators of the variance of finite samples have been considered in [4] and [19]. Estimators of conditional quantile have been developed in [19]. In [16] this problem was considered for dependent data. A new class of M–estimators with auxiliary information has been introduced in [14]. Missing data case presented in [7], censored data case has been considered in [15]. Problems of adaptive classification and optimization are considered in [5].

In this paper we consider the case when there exists an assumption on the value of estimated functional. The assumed value we will refer to as a prior guess. We propose k–adaptive combined estimators that use prior guess recursively k times. Asymptotic distributions of the estimators have been obtained, that allow to study the influence of a prior guess to the estimation accuracy.

The obtained asymptotic results extend the results presented in the paper [10].

1 Structure of estimator utilizing a prior guess

Let $X_1, ..., X_n$ be independent observations of size n over a random variable X with unknown distribution function F on R^1. Following to [10], consider the problem of statistical estimation of a linear functional on a certain class of distributions $\mathcal{F}$,

$$J(F) = M_F[\varphi(X)] = \int_{-\infty}^{\infty} \varphi(x)dF(x), \quad F \in \mathcal{F}, \tag{1}$$

using a prior guess J_a as a possible value of $J(F)$, specified by researcher. The real function φ is known. Nonparametric estimator of the functional is

$$\hat{J} = J(F_n) = n^{-1}\sum\nolimits_{i=1}^{n}\varphi(X_i),$$

where $F_n(x) = n^{-1}\sum_{i=1}^{n} c(x - X_i)$ is empirical distribution function, $c(t) = \{0 : t < 0; 1, t \geq 0\}$. Following to [8], [9], [10], [17], consider the combined estimator utilizing simultaneously $\hat{J}$ and prior guess J_a in the form

$$\hat{J}(\lambda) = (1 - \lambda)\hat{J} + \lambda J_a = \hat{J} - \lambda(\hat{J} - J_a), \tag{2}$$

where the weighting coefficient λ is selected from the minimum of mean square error (MSE) $S_F(\lambda) = M_F[\hat{J}(\lambda) - J]^2$. Optimal value of λ is given by

$$\lambda^* = \lambda^*(F) = (1 + n\Delta^2/\sigma^2)^{-1} = (1 + b_n^2(F))^{-1}, \tag{3}$$

where $\sigma^2 = \sigma^2(F) = D_F(\varphi(X))$ is the variance of $\varphi(X)$, $\Delta = \Delta(F) = J(F) - J_a$ is the value of displacement of the prior guess from the true value $J(F)$, and $b_n(F) = \sqrt{n}\Delta(F)/\sigma(F)$ is the normalized displacement.

The minimal value of MSE is given by the expression $nS_F(\lambda^*) = \sigma^2(1 - \lambda^*)$. The weighting factor λ^* varies between $0 < \lambda_n^* \leq 1$, and shows contribution of each estimator to the combined estimator (2) and their influence to the optimal MSE. Particularly, if $\Delta_F = 0$, we have $\lambda^* = 1$, and prior guess J_a should be taken as the estimator of the functional $J(F)$. When $\Delta_F \neq 0$, which usually happens in practice, $\lambda^* < 1$, and $\lambda^* \to 0$ with the growth of sample size ($n \to \infty$), so the influence of a prior guess and the advantage in the estimation accuracy decrease. More conclusions can be obtained if we assume that Δ decreases simultaneously with growth of n such that for each fixed $F \in \mathcal{F}$ there exists a limit $\lim b_n(F) = b$. Then $\lim nS_F(\lambda^*) = \sigma^2 b^2/(1 + b^2)$.

Practical usage of the combined estimator (2) is complicated because optimal coefficient λ^* is not possible to calculate due to distribution function F is unknown.

Construction of statistical estimators for λ^* leads to adaptive estimation of the functional (1). However, the weighting coefficient becomes non-optimal, and the question arises, under what conditions the adaptive estimator is more preferable by MSE as compared to the estimator $\hat{J}$. We consider this issue in the following sections.

2 Adaptive estimators and their asymptotic properties

We construct adaptive estimators by the method of substitution and consequent use of a prior guess. Let substitute unknown F with F_n in (3) and let use a prior guess

σ_a instead of σ. Then we do have the first estimator for λ^*:

$$\hat{\lambda}_1 = (1 + n\hat{\Delta}^2/\sigma_a^2)^{-1} = (1 + \hat{b}_n^2)^{-1},$$

where $\hat{\Delta} = \hat{J} - J_a$ is estimator of displacement Δ, $\hat{b}_n = \sqrt{n}\hat{\Delta}/\sigma_a$ is estimator of normalized displacement. By substitution λ with $\hat{\lambda}_1$ in (2), we obtain the first adaptive combined estimator $\hat{J}_1 = \hat{J} - \hat{\lambda}_1(\hat{J} - J_a)$. Using $\hat{J}_1$ in estimation of displacement Δ, we obtain $\hat{\Delta}_1 = \hat{J}_1 - J_a$ and $\hat{b}_{1,n} = \sqrt{n}\hat{\Delta}_1/\sigma_a$. Then the second estimator will be given by $\hat{\lambda}_2 = (1 + \hat{b}_{1,n}^2)^{-1}$ and $\hat{J}_2 = \hat{J} - \hat{\lambda}_2(\hat{J} - J_a)$. After repeating this procedure k times consecutively, we obtain the following expressions for the estimator

$$\begin{gathered} \hat{J}_k = \hat{J} - \hat{\lambda}_k(\hat{J} - J_a) = J_a + (1 - \hat{\lambda}_k)(\hat{J} - J_a), \\ \hat{\lambda}_k = \left(1 + n\hat{\Delta}_{k-1}^2/\sigma_a^2\right)^{-1} = \left(1 + \hat{b}_{k-1,n}^2\right)^{-1}, \end{gathered} \tag{4}$$

where $\hat{b}_{k,n} = \sqrt{n}\hat{\Delta}_k/\sigma_a$, $\hat{\Delta}_{k-1} = \hat{J}_{k-1} - J_a$, $\hat{\Delta}_0 = \hat{\Delta} = \hat{J} - J_a$, $\hat{b}_{0,n} = \hat{b}_n$.

Let us refer to $\hat{J}_k$ as k–adaptive estimator with parameter σ_a. We emphasize here that the prior guess J_a has been used at each step of estimation of Δ, but unknown value σ is replaced by the specified value σ_a. Let us note that in [10] the sample estimate $\hat{\sigma}^2$ was used instead of σ^2.

Consider asymptotic behavior of $\hat{J}_k$. Let

$$\hat{\xi}_k = \frac{\sqrt{n}(\hat{J}_k - J)}{\sigma}.$$

Denote

$$\eta_n = \frac{\sqrt{n}(\hat{J} - J)}{\sigma}, \qquad \tau = \frac{\sigma}{\sigma_a}.$$

Then we can write

$$\hat{b}_n = (\eta_n + b_n)\tau, \quad \hat{b}_{k,n} = q_k(\hat{b}_n) = q_k((\eta_n + b_n)\tau),$$

$$\hat{\lambda}_k = \left[1 + q_{k-1}^2((\eta_n + b_n)\tau)\right]^{-1},$$

$$\hat{\xi}_k = \frac{\sqrt{n}(\hat{J}_k - J)}{\sigma} = -b_n + q_k((\eta_n + b_n)\tau)/\tau,$$

where $q_k(x) = xq(q_{k-1}(x))$, $k \in \{1, 2, 3, \ldots\}$, $q(x) = x^2/(1 + x^2)$, $q_0(x) = x$.

Theorem 1. *Let $\sigma^2 < \infty$ for each $F \in \mathcal{F}$ and sequence b_n converges to non-random value b as $n \to \infty$. Then for each k the random sequence $\hat{\xi}_k$ converges in distribution to the random variable*

$$\xi_k = -b + q_k((\eta + b)\tau)/\tau \quad \mathit{if} \quad |b| < \infty, \quad 0 < \tau < \infty.$$

$$P\{\xi_k < x\} = \Phi(q_k^{-1}((x + b)\tau)\tau^{-1} - b), -\infty < x < \infty,$$

where η is the standard normal random variable with distribution function $\Phi(x)$, $q_k^{-1}(x)$ is inverse function.

Proof. Since functions $q_k(x)$ are continuous and monotonically increasing, then the statement of the theorem follows from convergency of η_n to η in distribution by the central limit theorem and the continuity theorem ([6], Chapter 6).

Corollary 1. *Under the conditions of the theorem 1, the following statements hold true.*

1. $\xi_k = \eta$ if $|b| = \infty$, $0 < \tau < \infty$.

2. $\xi_k \to \eta$ in distribution as $\tau \to \infty$, $|b| < \infty$.

3. If $\tau \to 0$ and $|b| < \infty$ then the distribution of ξ_k converges to degenerate distribution at point $-b$ (formally, $\xi_k \to -b$).

Proof. The first statement follows from the representation

$$\hat{\xi}_k = \eta_n - \frac{\eta_n + b_n}{1 + q^2_{k-1}((\eta_n + b_n)\tau)}, \tag{5}$$

where the second term converges weakly to zero as $|b_n| \to \infty$ due to the proposition 5 from lemma 1 [10]. The second and third statements of the corollary follows from the limit form of representation (5), convergency of $q^2_{k-1}(x)$ to infinity as $x \to \infty$, and convergency of $q_{k-1}(x)$ to zero as $x \to 0$.

3 Examples of k–adaptive combined estimators and numerical results

In this section we provide some examples of estimators, their asymptotic properties, and results of numeric calculations. Consider the k–adaptive combined estimators (4) $\hat{J}_k$ under $k \in \{1, 2, \ldots\}$.

$$\hat{J}_1 = \hat{J} - \left[1 + \hat{b}^2_n\right]^{-1} (\hat{J} - J_a),$$

$$\hat{J}_2 = \hat{J} - \left[1 + \frac{\hat{b}^3_n}{1 + \hat{b}^2_n}\right] (\hat{J} - J_a).$$

According to lemma 1 [10] where the expression for $q_\infty(x)$ is derived, the limit estimator (obtained after using the prior guess infinite number of times, $k = \infty$), can be written as

$$\hat{J}_\infty = \begin{cases} \hat{J} - \left[1 + \frac{\left(\hat{b}_n - \sqrt{\hat{b}^2_n - 4}\right)^2}{4}\right]^{-1} (\hat{J} - J_a), & \hat{b}_n \le -2, \\ J_a, & |\hat{b}_n| < 2, \\ \hat{J} - \left[1 + \frac{\left(\hat{b}_n + \sqrt{\hat{b}^2_n - 4}\right)^2}{4}\right]^{-1} (\hat{J} - J_a), & \hat{b}_n \ge 2. \end{cases}$$

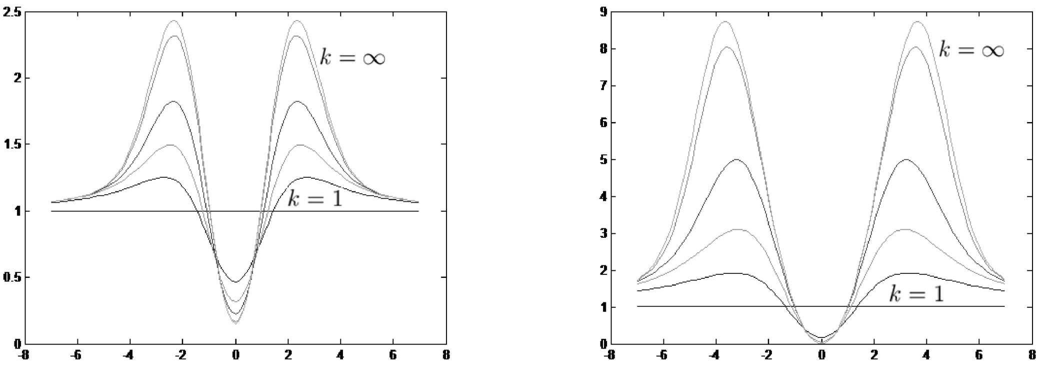


Figure 1: Dependence of the MSE $S\xi_k$ on normalized displacement b and $k \in \{1, 2, 4, 16, \infty\}$ for $\tau = 1.0$ (left plot) and $\tau = 0.5$ (right plot).

Using the theorem 1 we can compute moments of random variable ξ_k. Most interesting is the second moment, which due to (5) can be written in the form

$$M\xi_k^2 = S\xi_k = M\left[\eta - \frac{\eta + b}{1 + q_{k-1}^2((\eta + b)\tau)}\right]^2.$$

Figures 1 and 2 present the plots of $S\xi_k$ in dependence of k, b and τ. At the left plot of figure 1 the case of $\tau = 1$ is considered. It shows that there exist range of values of $|b|$ where $S\xi_k < 1$. Outside the range the combined estimators lose on MSE to regular empirical estimator represented by random variable ξ_0 with $S\xi_0 = 1$. The mentioned intervals and maximal loss are presented in the table 1 in numbers.

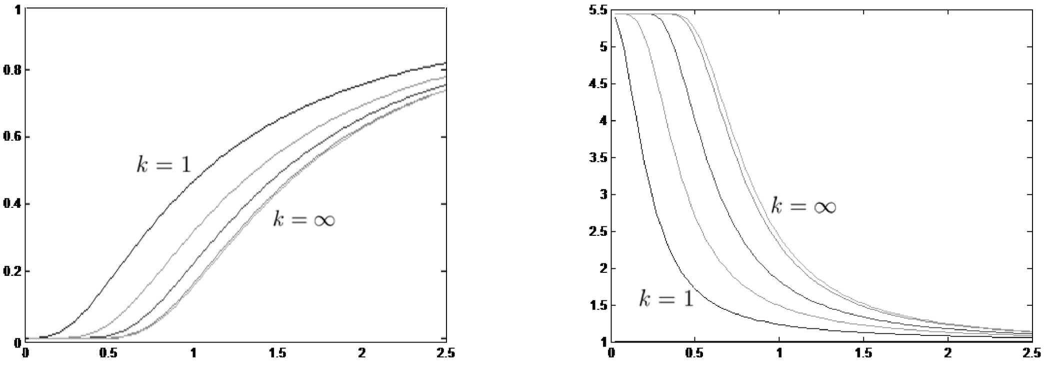


Figure 2: Dependence of the MSE $S\xi_k$ on τ and $k \in \{1, 2, 4, 16, \infty\}$ for normalized displacement $b = 0$ (left plot) and $b = 2.33$ (right plot).

When τ decreases, the maximum of $S\xi_k$ grows and minimum decreases down to zero (see for examples the right plot at the figure 1 and both plots at the figure 2). The inverse behavior is observing when τ increases, in that case $S\xi_k$ tends to

Table 1: Extremal points of $S\xi_k$ under $\tau = 1$ and points of its intersection with level one are presented with accuracy ±0.07.

k	1	2	4	16	∞
$\max_b S\xi_k$	1.25	1.49	1.82	2.31	2.43
$\arg\max_b S\xi_k$	± 2.66	± 2.52	± 2.38	± 2.38	± 2.24
$b:\ S\xi_k < 1$	$\|b\| < 1.40$	$\|b\| < 1.26$	$\|b\| < 1.12$	$\|b\| < 0.98$	$\|b\| < 0.98$

$S\xi_0 = 1$ for all b and τ. In the case of $b = 0$ (left plot at the figure 2) the value of $S\xi_k < S\xi_0 = 1$ for all $0 < \tau < \infty$, and this advantage is increasing with growth of k.

Acknowledgement

The authors thank Tomsk State University for organizational and financial assistance as the article is prepared within the Tomsk State University Competitiveness Improvement Program.

References

[1] Abu-Dayyeh W.A., Ahmed M.S., Ahmed R.A., Muttlak H.A. (2003) Some estimators of a finite population mean using auxiliary information. *Applied Mathematics and Computation.* Vol. 139, pp. 287–298.

[2] Albers C.J., Schaafsma W. (2003) Estimating a density by adapting an initial guess. *Computational Statistics and Data Analysis*, Vol. 42, pp. 27 – 36.

[3] Al-Omari Amer I. (2012) Ratio estimation of the population mean using auxiliary information in simple random sampling and median ranked set sampling. *Statistics and Probability Letters.* Vol. 82, pp. 1883–1890.

[4] Arcos A., Rueda, M., Martinez M.D., Gonzalez S., Roman Y. (2005) Incorporating the auxiliary information available in variance estimation. *Applied Mathematics and Computation.* Vol. 160, pp. 387–399.

[5] Baklizi A. (2005) Preliminary test estimation in the two parameter exponential distribution with time censored data. *Applied Mathematics and Computation.* Vol. 163, pp. 639–643.

[6] Borovkov A.A. (1998) *Mathematical statistics.* Gordon and Breach Science Publishers, Amsterdam.

[7] Bravo F. (2010) Efficient M-estimators with auxiliary information. *Journal of Statistical Planning and Inference.* Vol. 140, pp. 3326–3342.

[8] Dmitriev Yu.G., Skripin, S.V. (2012) On a combined assessment of the probability of failure-free operation for the full sample. *Tomsk State University Journal of Control and Computer Science*, Vol. 21, issue 4, pp. 32–38.

[9] Dmitriev Yu.G., Tarasenko P.F. (1992) The use of a priori information in the statistical processing of experimental data. *Russian Physics Journal*, September 1992, Vol. 35, Issue 9, pp. 888–893.

[10] Dmitriev Yu.G., Tarassenko P.F., Ustinov Y.K. (2014) On estimation of linear functional by utilizing a prior guess. *Communications in Computer and Information Science. A. Dudin et al. (Eds.): ITMM 2014.* Vol. 487, pp. 82–90.

[11] Ferguson T.S. (1973) A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, Vol. 1, Issue 2, pp. 209-230.

[12] Han F., Ling Q.-H. (2008) A new approach for function approximation incorporating adaptive particle swarm optimization and a priori information. *Applied Mathematics and Computation.* Vol. 205, pp. 792–798.

[13] Haq A., Shabbir J. (2014) An improved estimator of finite population mean when using two auxiliary attributes. *Applied Mathematics and Computation.* Vol. 241, pp. 14–24.

[14] Liang H.-Y., Jacobo de Una-Alvarez. (2011) Conditional quantile estimation with auxiliary information for left-truncated and dependent data. *Journal of Statistical Planning and Inference.* Vol. 141, pp. 3475–3488.

[15] Liu X., Liu P., Zhou Y. (2011) Distribution estimation with auxiliary information for missing data. *Journal of Statistical Planning and Inference.* Vol. 141, pp. 711–724.

[16] Qin Y.S., Wu Y. (2001) An estimator of a conditional quantile in the presence of auxiliary information. *Journal of Statistical Planning and Inference.* Vol. 99, pp. 59–70.

[17] Tarima S.S., Dmitriev Yu.G. (2009) Statistical estimation with possibly incorrect model assumptions. *Tomsk State University Journal of Control and Computer Science*, Vol. 8, issue 4, pp. 87–99.

[18] Vishwakarma G.K., Singh H.P. (2012) A general procedure for estimating the mean using double sampling for stratification and multi-auxiliary information. *Journal of Statistical Planning and Inference.* Vol. 142, pp. 1252–1261.

[19] Yadav S.K., Kadilar C. (2014) A two parameter variance estimator using auxiliary information. *Applied Mathematics and Computation.* Vol. 226, pp. 117–122.

Adaptive Estimation of Density Function Derivative

Dimitris N. Politis[1], Vyacheslav A. Vasiliev[2] and Peter F. Tarassenko[2]
[1] *Department of Mathematics, University of California, San Diego, USA*
[2] *Department of Applied Mathematics and Cybernetics,*
Tomsk State University, Tomsk, Russia
e-mail: dpolitis@ucsd.edu, vas@mail.tsu.ru, ptara@mail.tsu.ru

Abstract

The properties of non-parametric kernel estimators for the first order derivative of probability density function from special parameterized classes are investigated. In particular, in the case of known smooth classes parameter, rates of mean square convergency of density and its derivative estimators of smooth parameter estimators are found. Adaptive estimators of densities and their first derivatives from the given class with the unknown smooth parameter are constructed. Non-asymptotic and asymptotic properties of these estimators are established.

Keywords: Non-parametric kernel density estimators, smooth parameter estimation; adaptive density derivative estimators, mean square convergence, rate of convergence, smoothness class.

Introduction

Let $X_1, \dots, X_n$ be independent identically distributed random variables (i.i.d. r.v.'s) having a probability density function f. In the typical nonparametric set-up, nothing is assumed about f except that it possesses a certain degree of smoothness, e.g., that it has r continuous derivatives.

Estimating f via kernel smoothing is a sixty year old problem; M. Rosenblatt who was one of its originators discusses the subject's history and evolution in the monograph by [13]. For some point x, the kernel smoothed estimator of $f(x)$ is defined by

$$f_{n,h}(x) = \frac{1}{n} \sum\nolimits_{j=1}^{n} \frac{1}{h} K\left(\frac{x - X_j}{h} \right), \qquad (1)$$

where the kernel K is a bounded function satisfying $\int K(x)dx = 1$ and $\int K^2(x)dx < \infty$, and the positive bandwidth parameter h is a decreasing function of the sample size n. In this paper we will employ a particularly useful class of infinite order kernels namely the *flat-top* family; see [7] for a general definition.

It is a well-known fact that optimal bandwidth selection is perhaps the most crucial issue in such nonparametric smoothing problems; see [3] and the references therein. The goal typically is minimization of the large-sample Mean Squared Error (MSE) of $f_{n,h}(x)$. However, to do this minimization, the practitioner needs to know the degree of smoothness r. Using an infinite order kernel and focusing just on optimizing the order of magnitude of the large-sample MSE, it is apparent that the optimal bandwidth h must be asymptotically of order $n^{-1/(2r+1)}$ that yields a large-sample MSE of order $n^{-2r/(2r+1)}$ (see, e.g., [2]).

The problem of course is that, as previously mentioned, the underlying degree of smoothness r is typically unknown. In Section 3 of the paper at hand, we develop an estimator r_n of r and prove its strong consistency. In order to construct our estimator r_n, we operate under a class of functions that is slightly more general than, e.g., the Hölder class; this class of functions is formally defined in Section 1 via eq. (3) or (4). Under such a condition on the tails of the characteristic function we are able to show in Section 2 that the optimized MSE of $\hat{f}_n(x)$ is again of order $n^{-2r/(2r+1)}$ for possibly noninteger r.

Furthermore, in Section 4 we develop an *adaptive* estimator $\hat{f}_n(x)$ that achieves the optimal MSE rate of $n^{-2r/(2r+1)}$ within a logarithmic factor despite the fact that r is unknown, see Examples after Theorem 3. Similar effect arises in the adaptive estimation problem of the densities, in particular, from the Hölder class, see [1, 4, 5].

The estimaton problem of the density derivatives is actual as well; in particular for estimation of the logarithmic derivative. As the major theoretical result of our paper, we are able to prove a non-asymptotic upper bound for the MSE of the adaptive estimator of the density f and f'. The rate of convergency in the mean square sense satisfies (for the estimators of f in examples) the abovementioned optimal rate.

Section 5 contains some simulation results showing the performance of the estimator $\hat{f}_n(x)$ in practice.

Full investigation of the density function estimators will be presented in the paper [12].

1 Problem set-up and basic assumptions

Let $X_1, \ldots, X_n$ be i.i.d. having a probability density function f. Denote $\phi(s) = \int \mathrm{e}^{isx} f(x)dx$ the characteristic function of f and the sample characteristic function $\phi_n(s) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{e}^{isX_k}$. For some finite $r > 0$, define two families $\mathcal{F}_r^+$ and $\mathcal{F}_r$ of bounded, i.e.,

$$\exists\, 0 < \overline{f} < \infty : \ \sup_{y \in \mathcal{R}^1} f(y) \le \overline{f}, \tag{2}$$

and continuous functions f satisfying one of the following conditions respectively:

$$\int |s|^r |\phi(s)| ds < \infty, \quad \int |s|^{r+\varepsilon} |\phi(s)| ds = \infty, \quad \text{for all } \ \varepsilon > 0, \tag{3}$$

$$\int |s|^{r-\varepsilon} |\phi(s)| ds < \infty, \quad \int |s|^{r} |\phi(s)| ds = \infty, \quad \text{for all } \ 0 < \varepsilon < r. \tag{4}$$

It is easy to verify that the derivative f' satisfies the relations (3) and (4) if $f \in \mathcal{F}_{r+1}^+$ and $f \in \mathcal{F}_{r+1}$ respectively.

Define the family $\mathcal{F}_{r,m}^+$ (respectively $\mathcal{F}_{r,m}$) as the family of functions f from $\mathcal{F}_r^+$ (respectively $\mathcal{F}_r$) but with f being such that its characteristic function $|\phi(s)|$ has monotonously decreasing tails.

Consider the class Ξ of kernel smoothed estimators $f_{n,h}(x)$ of $f(x)$ as given in eq. (1). Note that we can alternatively express $f_{n,h}^{(l)}(x)$ in terms of the Fourier transform of kernel K, i.e.,

$$f_{n,h}^{(l)}(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h^{1+l}} K^{(l)}\left(\frac{x - X_j}{h}\right) = \frac{1}{2\pi}\int \lambda^{(l)}(s,h)\phi_n(s)\mathrm{e}^{-isx}ds, \quad l = 0;1, \tag{5}$$

where $\lambda^{(0)}(s,h) = \int K\left(\frac{x}{h}\right)e^{isx}dx$ and $\lambda^{(1)}(s,h) = \int K'\left(\frac{x}{h}\right)e^{isx}dx = -ish\lambda^{(0)}(s,h)$. In this paper, we will employ the family of flat-top infinite order kernels, i.e., we will let the function $\lambda^{(0)}(s,h)$ be of the form

$$\lambda_c(s,h) = \begin{cases} 1 \quad \text{if} \quad |s| \le 1/h, \\ g(s,h) \quad \text{if} \quad 1/h < |s| \le c/h, \\ 0 \quad \text{if} \quad |s| \ge c/h, \end{cases}$$

where c is a fixed number in $[1,\infty)$ chosen by the practitioner, and $g(s,h)$ is some properly chosen continuous, real-valued function satisfying $g(s,h) = g(-s,h)$, $g(s,1) = g(s/h,h)$, and $|g(s,h)| \le 1$, for any s, with $g(1/h,h) = 1$, and $g(c/h,h) = 0$; see [7]-[10] for more details on the above flat-top family of kernels.

Denote for every $0 \le \gamma < r$ the functions

$$\delta_\gamma(h) = \int\limits_{1/h<|s|<c/h} |s|^{r-\gamma}|\phi(s)|ds, \text{ when } h > 0, \quad \text{and } \delta_\gamma(0) = 0.$$

From (3) and (5) it follows that $\delta_\gamma(h) = o(1)$ as $h \to 0$ for $f \in \mathcal{F}_r^+$ and $\gamma = 0$, as well as for $f \in \mathcal{F}_r$ and $0 < \gamma < r$. In other cases $\delta_\gamma(h) = \infty$.

Define the following classes $\overline{\mathcal{F}}_r = \mathcal{F}_r^+ \cup \mathcal{F}_r$ and $\overline{\mathcal{F}}_{r,m} = \mathcal{F}_{r,m}^+ \cup \mathcal{F}_{r,m}$.

The main aim of the paper is adaptive estimation of densities and their first derivatives from the class $\overline{\mathcal{F}}_r$ with the unknown r.

2 Asymptotic mean square optimal estimation of f

According to [10, 11] the mean square error (MSE) $u_f^2(f_{n,h}) = E_f(f_{n,h}(x) - f(x))^2$ of the estimators $f_{n,h}(x) \in \Xi,\ f \in \overline{\mathcal{F}}_r$ has the following form:

$$u_f^2(f_{n,h}) = U_f^2(h,c) - \frac{1}{n}\left(\int K(v)f(x-hv)dv\right)^2, \tag{6}$$

where $U_f^2(h,c)$ is the principal term of the MSE,

$$U_f^2(h,c) = \frac{L_1 f(x)}{nh} + \left[\frac{1}{2\pi}\int\limits_{1/h<|s|<c/h} (1-g(s,h))\phi(s)\mathrm{e}^{-isx}ds\right]^2,$$

$L_1 = \int K^2(v)dv$. Thus, in particular, $\sup_{f\in\overline{\mathcal{F}}_r}\int K(v)f(x-hv)dv < \infty$.

The optimal (in the mean square sense) value $h^0 = h^0(n)$ is defined from minimization of the principal term $U_f^2(h,c)$.

Define the number $h_1^0 = h_1^0(n)$ from the equality

$$(h_1^0)^{2r+1-2\gamma}\delta_\gamma^2(h_1^0) = \frac{\pi^2 L_1 f(x)}{(c_0 + c_1(\gamma))n}. \tag{7}$$

In such a way we have proved the following theorem, which gives the rates of convergence of the random quantities $f_n^0(x) = f_{n,h^0}(x)$ and $f_{n,h_1^0}(x)$. We can loosely call $f_n^0(x)$ and $f_{n,h_1^0}(x)$ 'estimators' although it is clear that these functions can not be considered as estimators in the usual sense in view of the dependence of the bandwidths h^0 and h_1^0 on unknown parameters r and $f(x)$. Nevertheless, this theorem can be used for the construction of *bona fide* adaptive estimators with the optimal and suboptimal converges rates; see, e.g., Examples 1 and 2 in what follows.

Theorem 1. *Let $f(x) > 0$. Then for the asymptotically optimal (with respect to bandwidth h) in the MSE sense 'estimator' $f_n^0(x)$ of the function $f \in \overline{\mathcal{F}}_r$ and for the 'estimator' $f_{n,h_1^0}(x)$ of $f \in \overline{\mathcal{F}}_{r,m}$ the following limit relations, as $n \to \infty$, hold*

1°. $\sup\limits_{f\in\overline{\mathcal{F}}_r} \left| \inf\limits_h u_f^2(f_{n,h}) - U_f^2(h^0, c)\right| = O\left(\frac{1}{n}\right);$

2°. *for every $f \in \overline{\mathcal{F}}_{r,m}$ with $\gamma = 0$ if $f \in \mathcal{F}_{r,m}^+$ and every $0 < \gamma < r$ if $f \in \mathcal{F}_{r,m}$, as well as some constant C_γ, we have*

$$u_f^2(f_n^0) \le u_f^2(f_{n,h_1^0}) \le C_\gamma \cdot \frac{\delta_\gamma^{\frac{2}{2r+1-2\gamma}}(h_1^0)}{n^{\frac{2r-2\gamma}{2r+1-2\gamma}}}, \quad n \ge 1.$$

Remark 1. *The definition (7) of h_1^0 is essentially simpler than the definition of the optimal bandwidth h^0. From Theorem 1 it follows, that the (slightly) suboptimal 'estimator' f_{n,h_1^0} can be successfully used instead.*

Example 1. *Consider an estimation problem of the function $f \in \mathcal{F}_{r,m}^+$, satisfying the following additional condition*

$$|\phi(s)| \approx \frac{1}{|s|^{r+1}\ln^{1+\varphi}|s|} \quad \textit{as} \quad |s| \to \infty, \quad \varphi > 0. \tag{8}$$

We find the rates of convergence of the MSE $u_f^2(f_n^0)$ and $u_f^2(f_{n,h_1^0})$:

$$h_1^0 \approx \left(\frac{\ln^{2(1+\varphi)} n}{n}\right)^{\frac{1}{2r+1}} \quad \textit{and} \quad u_f^2(f_{n,h_1^0}) = O\left(\frac{1}{n^{2r}\ln^{2(1+\varphi)} n}\right)^{\frac{1}{2r+1}}$$

and using the piecewise linear flat-top kernel $\lambda_c^{LIN}(s,h)$, introduced by [9] (see [10] as well)

$$\lambda_c^{LIN}(s,h) = \frac{c}{c-1}\left(1 - \frac{h}{c}|s|\right)^+ - \frac{1}{c-1}\left(1 - h|s|\right)^+,$$

where $(x)^+ = \max(x, 0)$ is the positive part function, we find

$$h^0 \approx \left(\frac{\ln^{2(1+\varphi)} n}{n}\right)^{\frac{1}{2(r+1)}} \quad \textit{and} \quad u_f^2(f_n^0) = O\left(\frac{1}{n^{2r+1}\ln^{2(1+\varphi)} n}\right)^{\frac{1}{2(r+1)}} = o\left(u_f^2(f_{n,h_1^0})\right).$$

Example 2. *Consider an estimation problem of the function* $f \in \mathcal{F}_{r,m}$*, satisfying the following additional condition:*

$$|\phi(s)| \approx \frac{1}{|s|^{r+1}} \quad as \quad |s| \to \infty.$$

We find the rate of convergence of the MSE $u_f^2(f_n^0)$ *and* $u_f^2(f_{n,h_1^0})$*. From (7) we have*

$$h_1^0 \approx n^{-\frac{1}{2r+1}} \qquad and \qquad u_f^2(f_{n,h_1^0}) = O\left(n^{-\frac{2r}{2r+1}}\right), \quad as\ n \to \infty.$$

Similarly to Example 1, as $n \to \infty$*, for* $f \in \mathcal{F}_r$ *we find*

$$h^0 \approx n^{-\frac{1}{2(r+1)}} \qquad and \qquad u_f^2(f_n^0) = O\left(n^{-\frac{2r+1}{2(r+1)}}\right) = o\left(u_f^2(f_{n,h_1^0})\right).$$

Similar results can be obtained for the estimators of f'.

3 Estimation of the degree of smoothness r

Define the functions

$$\Phi_\alpha(A,B) = \int\limits_{A<|s|<B} |s|^\alpha |\phi(s)| ds, \qquad \Phi_{n,\alpha}(A,B) = \int\limits_{A<|s|<B} |s|^\alpha |\phi_n(s)| ds.$$

Let $(\delta_n)_{n\geq 1}$ and $(\rho_n)_{n\geq 1}$ be two given sequences of positive numbers chosen by the practitioner such that $\delta_n \to 0$ and $\rho_n \to \infty$ as $n \to \infty$. The sequence (δ_n) represents the 'grid'-size in our search of the correct exponent r, while (ρ_n) represents an upper bound that limits this search.

Define the following sets of non-random sequences

$$\mathcal{C}_+ = \{(A_n, B_n, \delta_n)_{n\geq 1} :\ A_n \to \infty,\ 0 < A_n < B_n \to \infty,\ \delta_n \to 0 \text{ as } n \to \infty;\ \text{ for some } m_0 \geq 2,$$
$$\sum_{n\geq 1} \frac{B_n^{2m_0(\varrho_n+1+\delta_n)}}{n^{m_0}} < \infty;\ \ \Phi_{r+\varepsilon}(A_n,B_n) \to \infty,\ \ \forall \varepsilon > 0;\ \ \Phi_{r+\delta_n}(A_n,B_n) \to \infty\},$$
$$\mathcal{C} = \{(A_n, B_n, \delta_n)_{n\geq 1} :\ A_n \to \infty,\ 0 < A_n < B_n \to \infty,\ \delta_n \to 0 \text{ as } n \to \infty;\ \text{ for some } m_0 \geq 2,$$
$$\sum_{n\geq 1} \frac{B_n^{2m_0(\varrho_n+1+\delta_n)}}{n^{m_0}} < \infty;\ \ \Phi_{r-\delta_n}(A_n,B_n) \to 0;\ \ \Phi_r(A_n,B_n) \to \infty\}$$

and for an arbitrary given $H > 0$ chosen by the practitioner, the estimators $(r_n^+)_{n\geq 1}$ and $(r_n)_{n\geq 1}$ of the parameter r in (3) and (4) respectively

$$r_n^+ = \min[\varrho_n, (\delta_n \cdot \inf\{k \geq 1 :\ \Phi_{n,(k+1)\delta_n}(A_n,B_n) \geq H,\ (A_n,B_n,\delta_n) \in \mathcal{C}_+\})], \tag{9}$$

$$r_n = \min[\varrho_n, (\delta_n \cdot \inf\{k \geq 1 :\ \Phi_{n,k\delta_n}(A_n,B_n) \geq H,\ (A_n,B_n,\delta_n) \in \mathcal{C}\})]. \tag{10}$$

Theorem 2. *The estimators* r_n^+ *and* r_n*, defined in (9) and (10) respectively have the following properties*

a) if $f \in \mathcal{F}_r^+$ *and for some* $\delta_n \to 0$ *the sequences* $(A_n, B_n, \delta_n) \in \mathcal{C}_+$*, then*

$$\lim_{n\to\infty} \delta_n^{-1}(r_n^+ - r) = 0 \quad P_f - a.s.$$

b) if $f \in \mathcal{F}_r$ *and for some* $\delta_n \to 0$ *the sequences* $(A_n, B_n, \delta_n) \in \mathcal{C}$*, then*

$$\lim_{n\to\infty} \delta_n^{-1}(r_n - r) = 0 \quad P_f - a.s.$$

4 Adaptive estimation of the functions $f, f' \in \overline{\mathcal{F}}_r$

The purpose of this section is the construction and investigation of an adaptive estimator of the functions $f, f' \in \overline{\mathcal{F}}_r$ with unknown r, which can either serve as the main estimator or can serve as a 'pilot' estimator for the construction of an adaptive optimal and suboptimal bandwidths $\hat{h}^0$ and $\hat{h}_1^0$.

We define an adaptive estimators of f and f' from $\overline{\mathcal{F}}_r$ as follows

$$\hat{f}_n^{(l)}(x) = \frac{1}{n}\sum_{j=1}^{n} \Lambda_{j-1}^{(l)}\left(x - X_j\right) = \frac{1}{2\pi n}\sum_{j=1}^{n}\int \lambda_{j-1}^{(l)}(s)\mathrm{e}^{-is(x-X_j)}ds, \tag{11}$$

where $\Lambda_{j-1}^{(l)}(z) = \frac{1}{\hat{h}_{j-1}^{1+l}}K^{(l)}\left(\frac{z}{\hat{h}_{j-1}}\right) = \frac{1}{2\pi}\int \lambda_{j-1}^{(l)}(s)\mathrm{e}^{-isz}ds$ is the smoothing kernel, and $\lambda_{j-1}^{(0)}(s) = \lambda_c(s, \hat{h}_{j-1})$, $l = 0; 1$. The required bandwidths are defined by

$$\hat{h}_j = (j+1)^{-\frac{1}{1+2(r(j)+l)}}, \quad j \geq 1,$$

where $r(j) = r_j^+$ if $f \in \mathcal{F}_r^+$ and $r(j) = r_j$ if $f \in \mathcal{F}_r$; recall that the estimators r_j^+ and r_j are defined in (9) and (10) respectively.

Below $C(\gamma, l)$ are some constants and $\Psi_{\gamma,l}(n)$ are concrete decreasing to zero functions. Main properties of constructed estimators are stated in the following theorem.

Theorem 3. *Let the sequences (A_n, B_n, δ_n) in the definition of the estimator r_n^+ belong to the set $\mathcal{C}_+$ and in the definition of the estimator r_n to the set $\mathcal{C}$. Let $\gamma = 0$ if $f \in \mathcal{F}_r^+$ and $\gamma \in (0, r)$ if $f \in \mathcal{F}_r$, as well as $r > 0$ if $l = 0$ and $r > 1$ if $l = 1$. Then for every $n \geq 1$ the estimators (11) has the following properties:*

$$\sup_{f \in \overline{\mathcal{F}}_r} u_f^2(\hat{f}_n^{(l)}) \leq \Psi_{\gamma,l}(n) + \frac{C(\gamma, l)}{n}, \qquad l = 0; 1.$$

Examples 1 and 2 revisited.

Under appropriate chosen $\delta > 0$ and sequences (A_n, B_n, δ_n) in the definition of sets $\mathcal{C}_+, \mathcal{C}$:

In Example 1 (case $(f \in \mathcal{F}_r^+)$)

$$\Psi_{0,0}(n) \approx (nh_n)^{-1} \cdot (\ln n)^{\frac{2\delta}{(1+2r)^2}} \approx n^{-\frac{2r}{1+2r}} \cdot (\ln n)^{\frac{2\delta}{(1+2r)^2}}.$$

Then, according to Theorem 3, in this case the rate of convergence of adaptive density estimators of $f \in \mathcal{F}_r^+$ differs from the rate of non-adaptive estimators in [10] on the extra log-factor only.

For the functions $f \in \mathcal{F}_r$ and $\gamma \in (0, \min(r, 1))$ from Example 2 it follows that

$$\Psi_\gamma(n) \approx n^{-\frac{2(r-\gamma)}{1+2(r-\gamma)}} \cdot (\ln n)^{\frac{\delta}{1+2(r-\gamma)}} \quad \text{as} \quad n \to \infty.$$

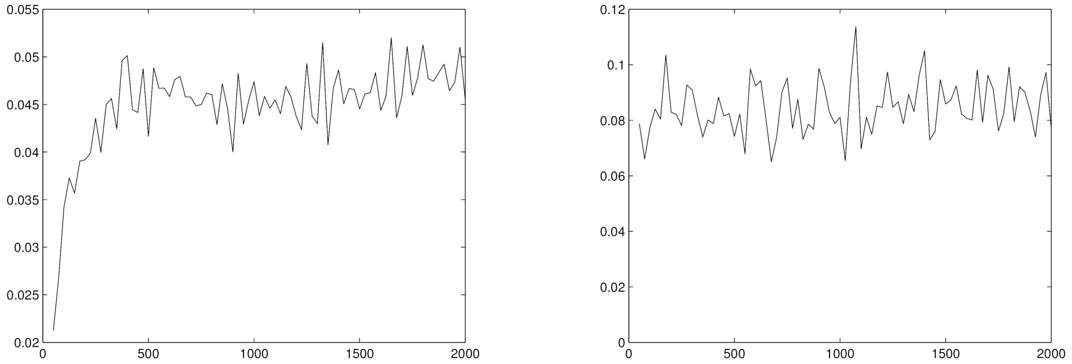


Figure 1: MSE of kernel estimators multiplied by $n^{3/4}$ versus $n \in \{25, 2000\}$. Left chart corresponds to the estimator with piece-wise linear kernel characteristic function. Right chart corresponds to the estimator with infinitely-differentiable flat-top kernel characteristic function.

5 Simulation results

In this section we provide brief results of simulation study of the estimators introduced in Section 2. We examine kernel estimators of triangular probability density function $f(x) = (a - |x|)/a^2, |x| \leq a$ belonging to the family $\mathcal{F}_1$ with characteristic function $\phi(s) = 2(1 - \cos(as))/(as)^2$. Also $\phi(s)$ meets requirements of the Example 2. Thus the bandwidth can be taken in the form $h = O(n^{-1/4})$ and expected convergence rate of the kernel estimator MSE is $n^{-3/4}$.

Two flat-top kernels have been used in the simulation. First one has the piece-wise linear kernel characteristic function introduced in [10]: $\lambda(s) = \{1, |s| \leq 1; (c-|s|)/(c-1), 1 < |s| < c; 0, |s| \geq c\}$. Second case refers to the infinitely-differentiable flat-top kernel characteristic function (see [6]) $\lambda(s) = \{1, |s| \leq c; exp\left[\frac{-b\exp\left[-b/(|s|-c)^2\right]}{(|s|-1)^2}\right]$, $c < |s| < 1; 0, |s| \geq 1\}$.

The main goal of the simulation study is investigation of the MSE behavior for the kernel estimator with the growth of sample size. We generate sequence of 150 samples for each sample size from 25 to 2000 with step 25, then calculate the estimator MSE multiplied by $n^{3/4}$ and expect visual stabilization of the sequence of resulting values with growth of n.

Two typical examples are presented at the Figure 1. Both cases refer to estimation of triangle density function $f(x)$ with unit variation (which support is bounded by ± 2.45, $a = 2.45$) at the point $x = 1.0$ by kernel estimators with flat-top kernels. The expected stabilization is observing in both cases.

References

[1] Brown L.D., Low M.G. (1992) Superefficiency and lack of adaptibility in functional estimation. Technical report, Cornell Univ.

[2] Dobrovidov, A.V., Koshkin, G.M., Vasiliev, V.A. (2012) Non-parametric state space models. Heber City, Utah: Kendrick Press.

[3] Jones M.C., Marron J.S., Sheather S.J. (1996) A brief survey of bandwidth selection for density estimation, J. Amer. Statist. Assoc., vol. 91, 401-407.

[4] Lepski O.V. (1990) One problem of adaptive estimation in Gaussian white noise. Theor. Probab. Appl., 35, pp. 459–470 (in Russian).

[5] Lepski O.V., Spokoiny V.G. (1997) Optimal pointwise adaptive methods in nonparametric estimation. Ann. Statist., 25 (6), pp. 2512–2546.

[6] McMurry T., Politis D.N. (2004) Nonparametric regression with infinite order flat-top kernels, J. Nonparam. Statist., vol. 16, no. 3-4, 549–562, 2004.

[7] Politis D.N. (2001) On nonparametric function estimation with infinite-order flat-top kernels, in Probability and Statistical Models with applications, Ch. Charalambides et al. (Eds.), Chapman and Hall/CRC: Boca Raton, pp. 469-483.

[8] Politis D.N. (2003) Adaptive bandwidth choice, J. Nonparam. Statist., vol. 15, no. 4-5, 517-533, 2003.

[9] Politis D.N., Romano J.P. (1993) On a Family of Smoothing Kernels of Infinite Order, in Computing Science and Statistics, Proceedings of the 25th Symposium on the Interface, San Diego, California, April 14-17, 1993, (M. Tarter and M. Lock, eds.), The Interface Foundation of North America, pp. 141-145.

[10] Politis D.N., Romano J.P. (1999) Multivariate density estimation with general flat-top kernels of infinite order. Journal of Multivariate Analysis, 1999, 68, 1-25.

[11] Politis D.N., Romano J.P., Wolf M. (1999) Subsampling. Springer, New York, 1999.

[12] Politis D.N., Vasiliev V.A., Tarassenko P.F. (2015) Estimating smoothness and optimal bandwidth for probability density functions. Metrika, Springer (submitted).

[13] Rosenblatt M. (1991) Stochastic curve estimation. NSF-CBMS Regional Conference Series, 1991, 3, Institute of Mathematical Statistics, Hayward.

Robust Estimation of Multivariate Regression Model in the Presence of Missing Data

Ekaterina M. Dolgovykh and Daniil V. Lisitsin
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: katdolgov@ya.ru, lisitsin@ami.nstu.ru

Abstract

We have applied our theoretically well-grounded method of robust parameter estimation from the multivariate nonhomogeneous incomplete data to multivariate normal regression in the presence of missing data with ignoring the missing-data mechanism. The theory is based on optimization of the weighted L_2-norm of Hampel's influence function. The estimators provide robustness against deviations from the assumed distribution. In this paper the form of estimators is given, questions of their calculation are discussed, Monte Carlo study is described.

Keywords: M-estimator, robustness, influence function, redescending estimator, multivariate regression, missing data, missing completely at random, ER algorithm.

Introduction

When studying complex objects their state is described by a vector of characteristics. However, the values of characteristics cannot always be fixed in observations. In this case, the data are incomplete and contain missing values [11]. This situation may arise in modeling multivariate data, which includes multivariate (multiresponse) regression model.

If a parametric model is assumed, parameters can be estimated by the maximum likelihood method. But such estimates may be unstable when there are deviations of the actual distribution from assumed distribution. To solve this problem robust procedures are used [5, 13, 14]. However, the theory of robustness was generally developed for modelling complete data.

In the presence of missing values in multivariate data in papers [1, 2, 4, 10, 12, 15] methods of robust estimation of shift and scale parameters of multivariate normal random variable or parameters of regression model are proposed. However these methods are based on heuristic arguments.

In this paper, the general theory of optimal estimation of the unknown parameters of the model from multivariate nonhomogeneous incomplete data [8] applies to multivariate normal regression in the presence of missing data and ignoring the missing-data mechanism. At the bottom of this theory we find synthesis of approach by F. Hampel [5] which is associated with the influence function and approach by A.M. Shurygin [14] which is associated with the Bayesian point-mass contamination model distribution. The resulting methods are robust against the deviation of the actual distribution of observations from an assumed one. Previously, these methods were applied to cases with nonhomogeneous quantitative (including count), qualitative and mixed data [3, 6, 9].

1 Elements of the theory of robust parameter estimation

Let n-dimensional independent random variables $\zeta_i = (\zeta_{i1}, \dots, \zeta_{in})^T$, $i = 1, \dots, N$, have an assumed (or ideal) pdf's $g_i(z_i|\phi)$, $z_i \in R^n$, with respect to a σ-finite measure μ and p-vector of parameters ϕ.

M-estimate $\hat{\phi}$ of vector of parameters ϕ is obtained from the observations $\tilde{\zeta}_i$, $i = 1, \dots, N$, of random variables ζ_i, $i = 1, \dots, N$, by means of a solution of the system of equations

$$\sum_{i=1}^{N} \psi_i(\tilde{\zeta}_i, \hat{\phi}) = 0,$$

where $\psi_i(\tilde{\zeta}_i, \hat{\phi})$ is p-dimensional score function satisfying further condition

$$\mathbf{E}\, \psi_i(z_i, \phi) = 0, \quad i = 1, \dots, N, \tag{1}$$

$\mathbf{E}$ is expectation under the assumed pdf.

In the robustness theory the estimates are of high quality not only in the assumed distribution, but in the case of a deviation from it. One of the major indicators of estimator's robustness is an influence function [5]. In our case, for M-estimator under certain regularity conditions, the influence function for the ith observation take the form [8]

$$\mathrm{IF}_i(z_i, \psi) = \mathrm{M}_1^{-1} \psi_i(z_i, \phi),$$

where $\psi = (\psi_1^T, \dots, \psi_N^T)^T$, $\mathrm{M}_1 = -\sum_{i=1}^{N} \frac{\partial}{\partial \tilde{\phi}^T} \mathbf{E}\, \psi_i(z_i, \tilde{\phi}) \Big|_{\tilde{\phi}=\phi} = \sum_{i=1}^{N} \int_{R^n} \psi_i(z_i, \phi) \frac{\partial g_i(z_i|\phi)}{\partial \phi^T}\, d\mu$ is non-singular $p \times p$ matrix.

Indicator of estimation badness can be written as square of the weighted L_2-norm of an influence function [8]

$$\Lambda_s(\psi) = \sum_{i=1}^{N} \int_{R^n} \mathrm{IF}_i^T(z_i, \psi)\, W\, \mathrm{IF}_i(z_i, \psi) s_i(z_i|\phi)\, d\mu,$$

where $s = (s_1, \dots, s_N)^T$, $s_i(z_i|\phi) > 0$ is weight function, $W = W(\phi)$ is some symmetric positive definite weight matrix of size $p \times p$ (under some conditions W can provide invariance of Λ_s to one-to-one differentiable parameter transformation [7, 8]).

Also, this indicator can be interpreted in accordance with the model of Bayesian point-mass contamination [14], when the first argument to the influence function is a random variable with pdf $s_i(z_i|\phi)$, $z_i \in R^n$, with respect to μ [8]. Then

$$\Lambda_s(\psi) = \sum_{i=1}^{N} \mathbf{E}_{s_i} \left[\mathrm{IF}_i^T(z_i, \psi)\, W\, \mathrm{IF}_i(z_i, \psi) \right],$$

where $\mathbf{E}_{s_i}$ is expectation under the pdf $s_i(z_i|\phi)$.

Optimal score function is a solution of minimization problem [8]:

$$\psi_s^* = \arg\min_{\psi} \Lambda_s(\psi)$$

under the constraints (1) and has the form

$$\psi_{s,i}^*(z_i, \phi) = C\left[\frac{\partial}{\partial\phi}\ln g_i(z_i|\phi) + \beta_i\right]\frac{g_i(z_i|\phi)}{s_i(z_i|\phi)},$$

where $C = C(\phi)$ is insignificant non-singular matrix, $\beta_i = \beta_i(\phi)$ is determined from the condition (1).

This general theory can be applied to the incomplete case [8].

Suppose that for each observation there is a set of *structures of missingness* — possible values of a vector of missing-data indicators [11]. Each structure of missingness shows the presence or absence of individual elements of observation. Enumerate these structures in some order and assume that the number of structure of missingness is random variable. Let ρ_i, $\tilde{\rho}_i$ denote such random variable and its observed value for the ith observation. By r_i denote the corresponding argument in the pdf's, estimators, etc.

For the r_ith structure of missingness we will introduce the vectors $\zeta_{i,obs}^{r_i}$ and $\zeta_{i,mis}^{r_i}$, corresponding to the observed and missing elements of the vector ζ_i and having the pdf's with respect to σ-finite measures $\mu_{i,obs}^{r_i}$ and $\mu_{i,mis}^{r_i}$, and the measure μ is a product of their. As a result, we have vectors $\tilde{\zeta}_{i,obs}$, $i = 1, \ldots, N$, of observations of random vectors $\zeta_{i,obs}^{r_i}$, $i = 1, \ldots, N$. In this way the sample is formed by the vectors $\left(\tilde{\zeta}_{i,obs}^{T}, \tilde{\rho}_i\right)^T$, $i = 1, \ldots, N$.

In general, the solution of problem of optimal estimation depends on the missing-data mechanism — the distribution of the random variable ρ_i. Often variable ρ_i is nuisance, therefore modeling this variable is not desirable. In [8] found conditions under which the missing-data mechanism can be ignored.

The first condition is MCAR — missing completely at random [11], when the random variable ρ_i is independent of the random vector ζ_i. MCAR condition with respect to the pdf's $g_i(z_i, r_i|\phi)$, $s_i(z_i, r_i|\phi)$ leads to $g_i(r_i|z_i) = g_i(r_i)$, $s_i(r_i|z_i) = s_i(r_i)$ under the assumption of independence of distribution of variable ρ_i from estimated parameters ϕ. The other conditions have the form $g_i(r_i) = s_i(r_i)$ and

$$\int\limits_{R^{n_{r_i}}} \psi_i(z_{i,obs}^{r_i}, r_i, \phi)\, g_i(z_{i,obs}^{r_i}|\phi)\, d\mu_{i,obs}^{r_i} = 0, \qquad (2)$$

where n_{r_i} is size of the vector $\zeta_{i,obs}^{r_i}$. The latter condition replaces (1).

As a result the optimal score function takes the form

$$\psi_{s,i}^*(z_{i,obs}^{r_i}, r_i, \phi) = C\left[\frac{\partial}{\partial\phi}\ln g_i(z_{i,obs}^{r_i}|\phi) + \beta_i^{r_i}\right]\frac{g_i(z_{i,obs}^{r_i}|\phi)}{s_i(z_{i,obs}^{r_i}|\phi)},$$

and is defined only by the marginal pdf's $g_i(z_{i,obs}^{r_i}|\phi)$, $s_i(z_{i,obs}^{r_i}|\phi)$; for each r_ith structure of missingness the vector $\beta_i^{r_i} = \beta_i^{r_i}(\phi)$ is determined from the condition (2).

2 Estimation of regression model in the presence of missing data

In the case of complete data the multivariate regression model has the form

$$y_i = F(x_i)\theta + e_i, \quad i = 1, \dots, N,$$

where y_i is the ith observation of the n-vector of quantitative responses, $F(x_i)$ is $n \times t$ matrix of regressors (functions of vector of deterministic input variables), x_i is vector of input variables of the ith observation, θ is t-vector of parameters, e_i is vector of errors having a multivariate normal distribution with zero mean vector and covariance matrix Σ. Actually observations of responses contain missing values.

Marginal distribution of the vector of observed responses is normal for the ith observation and the mth structure of missingness and has pdf

$$g(y^m_{i,obs}|x_i, \phi_m) =$$
$$(2\pi)^{-n_m/2} \left|\Sigma_{m,obs}\right|^{-1/2} \exp\left[-\frac{1}{2}\left(y^m_{i,obs} - F_{m,obs}(x_i)\theta\right)^T \Sigma^{-1}_{m,obs} \left(y^m_{i,obs} - F_{m,obs}(x_i)\theta\right)\right],$$

where $y^m_{i,obs}$ is n_m-vector of observed responses, ϕ_m is vector of parameters of marginal distribution, $\Sigma_{m,obs}$ is submatrix Σ, corresponding to the vector of observed responses, $F_{m,obs}(x_i)$ is $n_m \times t$ matrix, which consists of rows of a matrix $F(x_i)$, corresponding to the observed elements of a vector of responses.

One of the special cases of optimal estimators are generalized radical estimators with a pdf $s(y^m_{i,obs}|x_i, \phi_m) = \kappa_{\delta,m}\left[g(y^m_{i,obs}|x_i, \phi_m)\right]^{1-\delta}$, where $\kappa_{\delta,m}$ is normalizing factor, δ is parameter of robustness ($0 \leqslant \delta < 1$). Note that the case of $\delta = 0$ matches maximum likelihood estimator.

Thus, we have $g(y^m_{i,obs}|x_i, \phi_m)/s(y^m_{i,obs}|x_i, \phi_m) = \gamma_{\delta,m}\left[g(y^m_{i,obs}|x_i, \phi_m)\right]^{\delta}$, where $\gamma_{\delta,m} = (2\pi)^{\delta n_m/2}\left|\Sigma_{m,obs}\right|^{\delta/2}(1-\delta)^{-n_m/2}$. The factor $\gamma_{\delta,m}$ increases rapidly as $\delta \to 1$, so we use an equivalent factor $\tilde{\gamma}_{\delta,m} = (2\pi)^{\delta n_m/2}\left|\Sigma_{m,obs}\right|^{\delta/2}(1-\delta)^{(n-n_m)/2}$. As a result we define function

$$w(y^m_{i,obs}, x_i, \phi) = \tilde{\gamma}_{\delta,m}\left[g(y^m_{i,obs}|x_i, \phi_m)\right]^{\delta}.$$

Another special case is estimator analogous estimator of minimum variance sensitivity [13] corresponding $s(y^m_{i,obs}|x_i, \phi_m) = 1$ and $w(y^m_{i,obs}, x_i, \phi) = g(y^m_{i,obs}|x_i, \phi_m)$, it can be interpreted as some kind of generalized radical estimator with $\delta = 1$, $\gamma_{\delta,m} = 1$. Analogously, estimators with $\delta > 1$, $\gamma_{\delta,m} = 1$ can be used. They are useful under very unfavorable conditions.

Let's note that all presented estimators are redescenders [5, 13] and can be interpreted as some generalizations of Welsch's estimator, also known in the Russian publications as Meshalkin's estimator [14].

For the vector of regression parameters θ have the system of estimating equations

$$\left[\sum_{m=1}^{\widetilde{m}}\sum_{i\in J_m} w(y^m_{i,obs},x_i,\hat{\phi})F^T_{m,obs}(x_i)\hat{\Sigma}^{-1}_{m,obs}F_{m,obs}(x_i)\right]\hat{\theta}-$$
$$\sum_{m=1}^{\widetilde{m}}\sum_{i\in J_m} w(y^m_{i,obs},x_i,\hat{\phi})F^T_{m,obs}(x_i)\hat{\Sigma}^{-1}_{m,obs}y^m_{i,obs}=0,$$

where $\widetilde{m}$ is number of structures of missingness, J_m is set of numbers of observations corresponding to the mth structure of missingness.

Matrix Σ is symmetric positive definite matrix. So we can represent it as a Cholesky decomposition $\Sigma = SS^T$, where S is lower triangular matrix with real and positive diagonal entries. To ensure a positive diagonal, we parametrize the diagonal entries $S_{jj} = |\sigma_{jj}|$, where σ_{jj} is parameter. Although in the case of parameters transformation (this transformation must be one-to-one and differentiable [7, 8]) for maintaining optimality of estimator is necessary to impose a condition $\sigma_{jj} > 0$, in the calculations this condition is convenient to ignore.

The off-diagonal, nonzero entries of the matrix S denote by σ_{jk}, $j > k$. As a result, vector of parameters takes the form $\phi = (\theta^T, \sigma_{11}, \sigma_{21}, \ldots, \sigma_{nn})^T$.

Note that the matrix $\Sigma_{m,obs}$ represented as $\Sigma_{m,obs} = S_m\, S_m^T$, where S_m is $n_m \times n$ matrix, consisting of the rows of the matrix S, which correspond to the observed elements of a vector of responses for the mth structure of missingness.

Define a symmetric $n_m \times n_m$ matrix

$$\hat{B}_m = \hat{\Sigma}^{-1}_{m,obs}\sum_{i\in J_m} w(y^m_{i,obs},x_i,\hat{\phi})\,\hat{e}^m_{i,obs}\,(\hat{e}^m_{i,obs})^T\,\hat{\Sigma}^{-1}_{m,obs} - \frac{1}{1+\delta}\hat{\Sigma}^{-1}_{m,obs}\sum_{i\in J_m} w(y^m_{i,obs},x_i,\hat{\phi}),$$

where $\hat{e}^m_{i,obs} = y^m_{i,obs} - F_{m,obs}(x_i)\hat{\theta}$ is residual.

For the element σ_{jk}, $j \geqslant k$, have the estimating equation

$$\sum_{m\in J_j^{obs}}\sum_{v=1}^{n_m}(\hat{S}_m)_{vk}\,(\hat{B}_m)_{v,(j)} = 0, \tag{3}$$

where J_j^{obs} is set of numbers of structures of missingness, in which the jth entry of the vector of responses is observed, $(\hat{S}_m)_{vk}$ is entry in the vth row and kth column estimate of matrix S_m, $(\hat{B}_m)_{v,(j)}$ is entry of $\hat{B}_m$, and this entry is in the vth row and column corresponding to the jth entry of the vector of responses.

Component-wise procedure is used to calculate the estimates. According procedure the set of parameters is divided into a number of subsets. The solution is iterative. Each iteration involves several stages consisting in finding the next approximation of estimates for a subset of the parameters from the corresponding subsystem of estimating equations with fixed values of the parameters of the other subsets.

In our problem, forms two subsets — vector θ and set $\{\sigma_{jk}, j \geqslant k\}$. For θ the iteratively reweighted least squares algorithm is useful, for $\{\sigma_{jk}, j \geqslant k\}$ solve the appropriate subsystem of estimating equations by Broyden algorithm.

Instead of the equations of the form (3) the direct estimating equations for entries Σ_{jk}, $j \geqslant k$, of the covariance matrix can be used, then vector of parameters has the form $\phi = (\theta^T, \Sigma_{11}, \Sigma_{21}, \ldots, \Sigma_{nn})^T$. The estimating equation for the element Σ_{jk} has the form

$$\sum_{m \in J_{jk}^{obs}} (\hat{B}_m)_{(j),(k)} = 0, \tag{4}$$

where J_{jk}^{obs} is set of numbers of structures of missingness, in which the jth and kth entries of the vector of responses is simultaneously observed, $(\hat{B}_m)_{(j),(k)}$ is entry of $\hat{B}_m$, and this entry is in the row corresponding to the jth entry of the vector of responses and in the column corresponding to the kth entry of the vector of responses.

For the equations of the form (4) at appropriate stage of component-wise procedure is suitable the ER algorithm [15]. In our case ER algorithm is an iterative process such that at the $(u+1)$th iteration next approximation of estimate of the covariance matrix is computed, namely,

$$\Sigma^{(u+1)} = \frac{1+\delta}{\sum\limits_{i=1}^{N} w(y_{i,obs}^{m}, x_i, \phi^{(u)})} \sum_{i=1}^{N} w(y_{i,obs}^{m}, x_i, \phi^{(u)}) \left[e_i^{(u)} \, (e_i^{(u)})^T + \frac{1}{1+\delta} C_i^{(u)} \right],$$

where $\phi^{(u)}$ is approximation of the vector of parameters ϕ at the uth iteration, $e_i^{(u)} = y_i^{(u)} - F(x_i)\theta^{(u)}$, $y_i^{(u)} = \mathbf{E}(y_i | y_{i,obs}, \phi^{(u)})$, $C_i^{(u)} = \mathbf{cov}(y_i | y_{i,obs}, \phi^{(u)})$, **cov** is covariance under the assumed pdf. Note that $\Sigma^{(u)}$ is a symmetric positive definite matrix at every iterations of ER algorithm.

3 Monte Carlo results

A Monte Carlo study was conducted to verify the usefulness of the generalized radical estimates. The study is based on generating samples with the further computation of the parameter estimates.

The responses were generated by $y_{ji} = \theta_{j1} + \theta_{j2}\, x_i + e_{ji}$, where y_{ji} is value the jth response from the ith observation, $\theta_{j1} = \theta_{j2} = 1$, e_{ji} is the error value of the jth response from the ith observation, $n = 2$, input variable varies over the range $-1 \ldots 1$. Ideal distribution of errors was multivariate normal with zero mean vector and covariance matrix whose diagonal entries are equal to 0.01, off-diagonal entries are equal to 0.005. Actual distribution is received by contamination of ideal distribution [5]. Contaminating distribution was multivariate normal with mean $(0.1, -0.1)^T$ and covariance matrix twice greater than ideal. In study the results for different sample sizes, probabilities of missingness, fractions of contamination were obtained.

Measure of the badness of the estimates was $\sum\limits_{j=1}^{3} \sum\limits_{k=1}^{2} (\theta_{jk} - \hat{\theta}_{jk})^2$. Its average values increased on 10^3, for 2000 replications are shown in Figure 1. The abscissa is the parameter of robustness (its maximum value was 0.98). Following fractions of contamination are presented: 0 (solid lines), 0.1 (dashed lines), 0.2 (dotted lines). The left panel of Figure 1 represents the following case: $N = 200$; probability of missingness is 0 (circles) and 0.1 (triangles). The right panel of Figure 1 represents the following case: probability of missingness is 0.1; $N = 100$ (circles) and $N = 400$ (triangles).

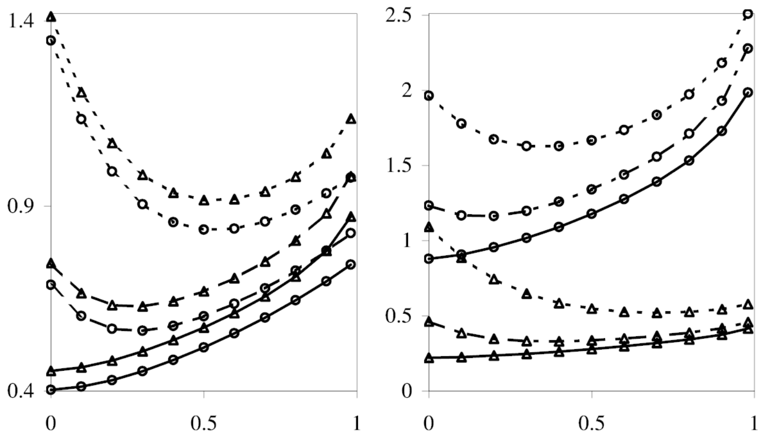


Figure 1: Measure of badness

The results show the advantage of robust estimates to maximum likelihood estimates in cases of contaminated distribution of errors. The best quality is obtained in different cases for the various values of robustness parameter. Estimation quality decreases with increasing the probability of missingness or fraction of contamination and with decreasing the size of sample.

Conclusion

The paper presents theoretically well-grounded robust estimates of multivariate normal regression model in the presence of missing data with ignoring the missing-data mechanism. The Monte Carlo study showed that robust estimates have better result than maximum likelihood estimates in cases of contaminated normal distribution of errors.

References

[1] Cheng T.-C., Victoria-Feser M.-P. (2002). High Breakdown Estimation of Multivariate Location and Scale with Missing Observations. *British Journal of Mathematical and Statistical Psychology*. V. **55**, pp. 317-335.

[2] Danilov M., Yohai V.J., Zamar R.H. (2012). Robust Estimation of Multivariate Location and Scatter in the Presence of Missing Data. *Journal of the American Statistical Association*. V. **107**, pp. 1178-1186.

[3] Dovgal S.Yu., Lisitsin D.V. (2011). Robust Estimation of Count Response Regression Models. *Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Simulations and Statistical Inference", Novosibirsk, Russia*, pp. 318-321.

[4] Frahma G., Jaekel U. (2010). A Generalization of Tyler's M-estimators to the Case of Incomplete Data. *Computational Statistics and Data Analysis*. V. **54**, pp. 374-393.

[5] Hampel F.R., Rouchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, Hoboken, NJ.

[6] Kalinin A.A., Lisitsin D.V. (2011). Robust Estimation of Qualitative Response Regression Models. *Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Simulations and Statistical Inference", Novosibirsk, Russia*, pp. 303-309.

[7] Lisitsin D.V. (2010). Invariance Properties under Estimating Model Parameters in Presence of Bayesian Dot Contamination. *Reports of Russian Higher Education Academy of Sciences*. No. **1(14)**, pp. 18-25 (in Russian).

[8] Lisitsin D.V. (2013). Robust Estimation of Model Parameters in Presence of Multivariate Nonhomogeneous Incomplete Data. *Science Bulletin of NSTU*. No. **1(50)**, pp. 17-30 (in Russian).

[9] Lisitsin D.V. (2013). Robust Estimation of Mixed Response Regression Models. *Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control", Novosibirsk, Russia*, pp. 139-144.

[10] Little R.J.A. (1988). Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values. *Applied Statistics*. V. **37**, pp. 23-38.

[11] Little R.J.A., Rubin D.B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

[12] Little R.J.A., Smith P.J. (1987). Editing and Imputing for Quantitative Survey Data. *Journal of the American Statistical Association*. V. **82**, pp. 58-68.

[13] Shevlyakov G., Morgenthaler S., Shurygin A. (2008). Redescending M-estimators. *Journal of Statistical Planning and Inference*. V. **138**, pp. 2906-2917.

[14] Shurygin A.M. (2000). *Applied Stochastics: Robustness, Estimation, Prediction*. Finances and Statistics, Moscow (in Russian).

[15] Yuan K.-H., Chan W., Tian Y. (2014). Expectation-Robust Algorithm and Estimating Equations for Means and Dispersion Matrix with Missing Data. *Annals of the Institute of Statistical Mathematics*. DOI: 10.1007/s10463-014-0498-1

Research of Stationary Random Processes with a Non-linear Regression. Part 1. Features and Nonlinear Transformations of Processes with Linear Regression

Vasiliy V. Gubarev, Olga K. Alsova, Nicolay V. Abalov,
Grigoriy A. Melnikov, Roman V. Terekhov and
Svetlana A. Pushkareva
Novosibirsk State Technical University, Novosibirsk, Russian Federation
e-mail: gubarev@vt.cs.nstu.ru, alsova@corp.nstu.ru,
nick.abalov@gmail.com, grmel89@hotmail.com,
stealth-709@yandex.ru, neliko3@rambler.ru

Abstract

Concerns about misinterpretation of research results of random processes (RP) with non-linear regression if their analysis tools are used traditional correlation and spectral analysis are states. A list of tasks that must be solved in order to determine the degree of fear and select targets for the development of alternative approaches. The features of regression functions (RF) random processes compared with RF random variables. In general terms, and specific examples investigate the property of proportional correlation of random processes with linear regression and nonlinear transformation of RP.

Keywords: Random process, correlation, correlation analysis, regression, regression analysis, nonlinear transformations.

Introduction. Problem Statement

Many physical signals are described by probabilistic models. These signals are called random. Among the most often considered a model of stationary random processes. At the same time to study the properties and characteristics of signals, as well as the solution of applied theoretical and practical problems using correlation, correlation-spectral[1], singular spectrum, wavelet and other derivative forms of analysis. They are directed, usually on the study of statistical relationships temporary signal samples, their frequency content, identifying the properties and characteristics of linear dynamic systems, signal flow through them, study differentiability signals (processes) in the mean square, as well as for the identification and prediction signals. Very rarely, in contrast to the random vectors, the study focused on stochastic processes and their regression and scedastic analyzes. This is all the more surprising that the correct application of the correlation and the associated correlation and spectral analysis for

[1] Under the spectral correlation is defined here as an analysis based on the use of spectral description of the power spectral density (PSD) obtained by Fourier transform of the correlation function (CF) or as a limit for an unlimited increase in the length T of trajectories (realizations) of random processes, ensemble averaged of PSD periodogram estimates[4, 5].

many applications is possible only if the corresponding private or cross regression function processes are linear. Examples of erroneous conclusions for random vectors when the function is non-linear regression, and we obtain (consciously or not) its approximation in the form of mean-square regression line are given in [2, 3, 4, 5]. However, neither the domestic or world literature specific individual studies related to the properties and characteristics of random processes with non-linear regression, its impact on the results of various applications, as well as a preliminary exploratory analysis of random signals belonging to the signals from the line (LR) or nonlinear (NR) regression, i.e. describing assignment of processes to LR or NR-classes haven't been conducted. Therefore, the aim of this paper are: first, the wording of the priorities of research, and secondly, to determine the characteristics of processes with linear regression, in particular those which would allow the RP to refer to a class of processes with linear or nonlinear regression, thirdly analysis of the impact of nonlinear transformations of LR-processes on auto and cross CF.

1 Priorities for research

Task 1 - Investigation of the features of the function (line) regression (LR) of random processes in comparison with the RF of random vectors variables.

Task 2 - Development and theoretical (task 2.1) and experimental (task 2.2) research of methods classification RP of classes LR or NR.

Task 3 - Theoretical (task 3.1) and experimental (task 3.2) investigation of the features of its auto and cross correlation, correlation-spectral and singular spectrum analysis RP with nonlinear regression.

Task 4 - Theoretical (task 4.1) and experimental (task 4.2) investigation of the effect of non-linearity of the RP regression function on the results of the applied problems solution.

Task 5 - Develop and theoretical (task 5.1) and experimental (task 5.2) study of the ratio characteristics and frequency of communication that are invariant to the form one to one nonlinearity of the RP regression function .

2 Accepted Designations

We denote by $X(t)$, $Y(t)$, $Z(t)$ studied in the stationary level in all considered characteristics of random processes. We shall consider the following characteristics (see., Eg, [1-4]):the mathematical expectations m_X, m_Y, m_Z, where

$$m_X = \mathbf{M}\{X(t)\} = m_X, \tag{1}$$

$\mathbf{M}\{\cdot\}$ - averaging operator to the probability measure corresponding dimension (the mathematical expectation operator); dispersion D_X, D_Y, D_Z and standard deviations σ_X, σ_Y, σ_Z, where

$$D_X(t) = \mathbf{M}\{[\overset{\circ}{X}(t)]^2\} = \mathbf{M}\{[X(t) - m_X(t)]^2\} = D_X = \sigma_X^2 = R_{XX}(0), \tag{2}$$

$\overset{\circ}{X}(t) = X(t) - m_X(t)$ - centric RP $X(t)$;
initial $m_{k,n}(t, t+\tau)$ and central $\mu_{k,n}(t, t+\tau)$ moments

$$m_{k,n;X,Y}(t, t+\tau) = \mathbf{M}\{X^k(t)Y^n(t+\tau)\} = m_{k,n;X,Y}(\tau), \tag{3}$$

$$\mu_{k,n;X,Y}(t, t+\tau) = \mathbf{M}\{[\overset{\circ}{X}(t)]^k[\overset{\circ}{Y}(t+\tau)]^n\} = \mu_{k,n;X,Y}(\tau), \tag{4}$$

in particular cross $B_{XY}(\tau)$, $R_{XY}(\tau)$, self(auto) $B_{XX}(\tau)$, $R_{XX}(\tau)$ covariation (B) and correlation (R) functions

$$B_{XY}(\tau) = m_{1,1;X,Y}(\tau), B_{XX}(\tau) = m_{1,1;X}(\tau); \tag{5}$$

$$R_{XY}(\tau) = \mu_{1,1;X,Y}(\tau), R_{XX}(\tau) = \mu_{1,1;X}(\tau); \tag{6}$$

the normalized correlation functions

$$\rho_{XY}(\tau) = \frac{R_{XY}(\tau)}{\sigma_x\sigma_y}, \rho_{XX}(\tau) = \frac{R_{XX}(\tau)}{\sigma_x^2} = \frac{R_{XX}(\tau)}{R_{XX}(0)}. \tag{7}$$

Consider also the eigen for $X(t)$, and cross for $X(t)$, $Y(t)$ regression functions

$$m_x(t+\tau; x, t) = \mathbf{M}\{X(t+\tau)|(X(t) = x)\} = m_X(x; -\tau), \tag{8}$$

$$m_x(t; x, t+\tau) = \mathbf{M}\{X(t)|(X(t+\tau) = x)\} = m_X(x; \tau), \tag{9}$$

$$m_y(t+\tau; x, t) = \mathbf{M}\{Y(t+\tau)|(X(t) = x)\} = m_y(x; -\tau), \tag{10}$$

$$m_y(t; x, t+\tau) = \mathbf{M}\{Y(t)|(X(t+\tau) = x)\} = m_y(x; \tau), \tag{11}$$

$$m_x(t+\tau; y, t) = \mathbf{M}\{X(t+\tau)|(Y(t) = y)\} = m_X(y; -\tau), \tag{12}$$

$$m_x(t; y, t+\tau) = \mathbf{M}\{X(t)|(Y(t+\tau) = y)\} = m_X(y; \tau), \tag{13}$$

where $\mathbf{M}\{Y|(X = x)\}$ - averaging operator for the conditional probability measure, ie operator finding the average value (conditional mathematical expectation) of the random value Y pair (X, Y), provided that this quantity X takes the value of $X = x$.

3 Theoretical study of features of random processes regression functions

3.1 Properties of random processes regression functions

Various properties of the regression function of the random variables X and Y-vector elements (X, Y) in the compact form shown in [3]. Important properties of RF RP, different from those of the RF random variables considered in [2]. Consider the supplement and important ones.

First of all, pay attention to two similar properties. *The first property is "linearity"*. It related to the fact that for the RP of LR class all regression function of the type (8) - (13) are linear. Wherein

$$m_X(x; \tau) = m_x + \rho_{XX}(\tau)(x - m_X) = m_X(x; -\tau), \tag{14}$$

$$m_Y(x;\tau) = m_Y + \frac{\sigma_Y}{\sigma_X}\rho_{XY}(\tau)(x-m_X) = m_Y + \frac{\sigma_Y}{\sigma_X}\rho_{YX}(-\tau)(x-m_X) = m_Y(x;-\tau), \tag{15}$$

$$m_X(y;\tau) = m_X + \frac{\sigma_Y}{\sigma_X}\rho_{XY}(\tau)(y-m_Y) = m_X + \frac{\sigma_Y}{\sigma_X}\rho_{YX}(-\tau)(y-m_Y) = m_X(y;-\tau). \tag{16}$$

And from (16)-(17) follows that since $\rho_{XX}(0) = 1$,

$$m_X(x;0) = x, \tag{17}$$

$$m_Y(x;0) = m_Y + \frac{\sigma_Y}{\sigma_X}\rho_{XY}(0)(x-m_X), \tag{18}$$

$$m_X(y;0) = m_X + \frac{\sigma_X}{\sigma_Y}\rho_{XY}(0)(y-m_Y). \tag{19}$$

The second property is a "proportionality" of RP $X(t)$, $Y(t)$ with the linear regression it reduced that all the center moments $\mu_{k,1}(\tau)$, $\mu_{1,n}(\tau)$ with $k, n \geq 1$ are proportional to $\rho_{XY}(\tau)$ [2].

The third property of LR RP $X(t)$, $Y(t)$ also applies to *"non-linear" proportionality* moments with respect to ρ [2]. Let $X(t)$ and $Y(t)$ belong to the LR class and $g(\cdot)$ - numerical (deterministic, unambiguous, instantaneous) function. Then

$$\begin{aligned}\mathbf{M}\{X(t)g[Y(t+\tau)]\} = m_x{\cdot}\mathbf{M}\{g([Y(t+\tau)]\} + \frac{\sigma_X}{\sigma_Y}\rho_{XY}\mathbf{M}\{\overset{\circ}{Y}(t+\tau)g[Y(t+\tau)]\} = \\ = m_X{\cdot}m_{g(Y)} + \frac{\sigma_X}{\sigma_Y}\rho_{XY}(\tau)\mathbf{M}\{\overset{\circ}{Y}(t)g[Y(t)]\}.\end{aligned} \tag{20}$$

Therefore

$$R_{Xg(Y)}(\tau) = \mathbf{M}\{\overset{\circ}{X}(t)\overset{\circ}{g}[Y(t+\tau)]\} = A_{Yg}\rho_{XY}(\tau) = A_{Yg}\rho_{XY}(-\tau); \tag{21}$$

where

$$A_{Yg} = \frac{\sigma_X}{\sigma_Y}\mathbf{M}\{\overset{\circ}{Y}(t)g[Y(t)]\}. \tag{22}$$

From (21) it follows that for fixed RP belongs to the LR class,

$$R_{Xg(X)}(\tau) = \mathbf{M}\{\overset{\circ}{X}(t)g[X(t)]\}\rho_{XX}(\tau) = A_X\rho_{XX}(\tau) = A_X\rho_{XX}(-\tau) = R_{Xg(X)}(-\tau). \tag{23}$$

The fourth property "reducible to a point". Assume that $m_X(x;\tau) = m_X + \psi_X(x;\tau)$ to $X(t)$, and $m_X(y;\tau) = m_X + \psi_X(y;\tau)$ to $X(t)$, $Y(t)$ where $\psi(\cdot)$ - function describing RF. Then, according to [2].

$$\mathbf{M}\{\overset{\circ}{X}(t)\psi_X[X(t+\tau)|X(t)=X]\} = \mathbf{M}_X\{\psi_X^2(X;\tau)\}, \psi_X(x;0) = x - m_X; \tag{24}$$

$$\begin{cases}\mathbf{M}\{\overset{\circ}{X}(t)\psi_X[Y(t+\tau)|X(t)=X]\} = \mathbf{M}_Y\{\psi_X^2(y;\tau)\}, \\ \psi_X(y;\tau_0) = \frac{\sigma_X}{\sigma_Y}(y-m_Y) sign\rho_{XY}(\tau_0) \ when \ |\rho_{XY}(\tau_0)| = 1,\end{cases} \tag{25}$$

where $sign(x)$ is signum function, i.e., $sign(x) = -1$ for $x < 0$, 0 at $x = 0$ and 1 when $x > 0$. The first relation in (24) and (25) are valid for random vectors and

LR-processes and the second in (24) are specific to RP with any RF and impose restrictions on the form of the function $\psi(\cdot)$. For RP with linear regression it follows from (18), (19).

Fifth property "reduction to the value of transfer function" refers to the RF $Y(t) = f[X(t)]$. If $f(\cdot)$ - valued deterministic numerical function, then

$$m_Y(x; 0) = f(x). \tag{26}$$

If the inverse of $f(x)$ function $x = \phi(y)$ is also unique, ie $f(\cdot)$ and $\phi(\cdot)$ are mutually-to-one function, then

$$m_X(y; 0) = \phi(y). \tag{27}$$

As for RF for other τ when $Y(t) = f[X(t)]$, in general terms $m_Y(x; \tau)$, $m_X(y; \tau)$ $m_Y(y; \tau)$ be imagined without the knowledge of the distribution law is impossible. However, (26) and (27) can be used as constraints, which must comply with the RF $X(t)$, $Y(t)$.

3.2 Special Cases

As an example, consider some special cases. Let $Y(t) = X^3(t) = [\overset{\circ}{X}(t)]^3$. For this dependence have

$$R_{YY}(\tau) = \mathbf{M}\{[\overset{\circ}{X}(t)]^3[\overset{\circ}{X}(t+\tau)]^3\} - [\mathbf{M}\{[\overset{\circ}{X}(t)]^3\}]^2, \tag{28}$$

$$R_{YY}(\tau) = \frac{R_{YY}(\tau)}{R_{YY}(0)} = \frac{\mu_{3,3}(\tau) - \mu_3^2}{\mu_6 - \mu_3^2}, \tag{29}$$

$$R_{XY}(\tau) = \mu_{1,3}(\tau), \tag{30}$$

$$R_{YX}(\tau) = \mu_{3,1}(\tau). \tag{31}$$

For LR - processes (28) - (31), taking into account (20)

$$R_{XY}(\tau) = R_{YX}(\tau) = \mu_{4;X}\rho_{XX}, \tag{32}$$

$$\rho_{XY}(\tau) = \frac{\mu_{4,X}\rho_{XX}(\tau)}{\sigma_X \sigma_{\overset{\circ}{X}^3}}. \tag{33}$$

Since $\sigma_X = \sqrt{\mathbf{M}\{[X^3(t) - \mathbf{M}X^3(t)]^2\}} = \sqrt{(\mu_6 - \mu_3^2)}$, from (33) we obtain for centered LR processes

$$\rho_{\overset{\circ}{X}\overset{\circ}{X}^3}(\tau) = \frac{\mu_{4,X}\rho_{XX}(\tau)}{\sigma_X\sqrt{\mu_6 - \mu_3^2}}. \tag{34}$$

Similarly, for centered LR process $X(t)$ we have

$$\rho_{XX^3}(\tau) = \frac{m_4 - m_1 m_3}{\sigma_X\sqrt{m_6 - m_3^2}}. \tag{35}$$

Consider special cases.

1. $X(t) = \mathring{X}(t)$ - centered normal process. This is LR - process. For him, we have [4] $\mu_{4X} = 3\mu_{2X}^2 = 3\sigma_X^4$, $\mu_6 = 15\sigma_X^6$; $\mu_3 = 0$, $\mu_2 = \sigma_X^2$ and so (see also [5]).

$$\rho_{XY}(\tau) = \rho_{X(\mathring{X})^3}(\tau) = \rho_{(\mathring{X})^3 X}(\tau) = \sqrt{\frac{3}{5}}\rho_{XX}(\tau) \approx 0,775\rho_{XX}(\tau). \tag{36}$$

Next to the case in question

$$\rho_{YY}(\tau) = \rho_{(\mathring{X})^3(\mathring{X})^3}(\tau) = \frac{1}{5}\rho_{XX}(\tau)[3 + 2\rho_{XX}^2(\tau)], \tag{37}$$

$$m_Y(y;\tau) = \mathbf{M}\{Y(t+\tau)|Y(t) = y\} = 3\sigma_X^2 y^{1/3}[\rho_{XX}(\tau) - \rho_{XX}^3(\tau)] + y\rho_{XX}^3(\tau), \tag{38}$$

$$m_Y(x;\tau) = \mathbf{M}\{Y(t+\tau)|X(t) = x\} = 3\sigma_X^2 x[\rho_{XX}(\tau) - \rho_{XX}^3(\tau)] + x^3\rho_{XX}^3(\tau). \tag{39}$$

2. $X(t)$ - a two-dimensional random process with the gamma density distribution[5]

$$W_X(x,y;\tau) = \frac{(z_1 z_2)^{\frac{a-1}{2}} \exp\{-\frac{z_1+z_2}{1-\rho(\tau)}\}}{[1-\rho(\tau)]\lambda^2\rho^{\frac{a-1}{2}}(\tau)\Gamma(a)} I_{a-1}(\frac{2\sqrt{z_1 z_2 \rho(\tau)}}{1-\rho}); \tag{40}$$

where $z_1 = (x-a)/\lambda$, $z_2 = (y-a)/\lambda$, $|a| < \infty$, $\lambda > 0$, $a > 0$, $\rho(\tau) \geq 0$; $\Gamma(a)$ - gamma function, $I_a(x)$ - Bessel function of imaginary argument [1]. RP with such a distribution has a linear regression [4]. A special case of the gamma distribution is an exponential distribution, obtained from (40) with $a = 1$. For our distribution points species $\mu_{1,1}(\tau)$, $\mu_{2,2}(\tau)$, $mu_{3,3}(\tau)$ are presented in the general form of the hypergeometric functions [4]. Therefore, we consider only the cross CF. Since, according to [4] for the distribution (40) $\mu_2 = \sigma_X^2 = a\lambda^2$, $\mu_3 = 2a\lambda^3$, $\mu_4 = 3\lambda^4(2a + a^2)$, $\mu_6 = 5\lambda^5(24a + 25a_2 + 3a^2)$ of (36), we have

$$\rho_{X(\mathring{X})^3}(\tau) = \frac{\sqrt{3}(2+a)\rho_{XX}(\tau)}{\sqrt{40 + 42a + 5a^2}}; \tag{41}$$

i.e.

$$\rho_{X(\mathring{X})^3}(\tau) = [\sqrt{\frac{3}{10}}; \sqrt{\frac{3}{5}}]\rho_{XX}(\tau) \approx [0,55; 0,755]\rho_{XX}(\tau). \tag{42}$$

For an exponential distribution, i.e. when $a = 1$, from (41) we obtain

$$\rho_{X(\mathring{X})^3}(\tau) = \sqrt{\frac{9}{29}}\rho_{XX}(\tau) \approx 0,56\rho_{XX}(\tau). \tag{43}$$

When $a \to \infty$, i.e. for the limiting case of one-dimensional gamma distribution - Gauss distribution, we obtain

$$\rho_{X(\mathring{X})^3}(\tau) = \frac{3}{\sqrt{15}}\rho_{XX}(\tau) \approx 0,755\rho_{XX}(\tau); \tag{44}$$

ie the same expression for the Gaussian (normal) distribution, but for the bivariate normal distribution (42) $\rho_{XX}(\tau) \in [-1,1]$ in contrast to (53), where $\rho_{XX}(\tau) \in [0,1]$. When $0 \le a \le 1/2$.

$$\rho_{(\mathring{X})^3(\mathring{X})^3}(\tau) \approx 0,55\rho_{XX}(\tau). \tag{45}$$

On the other hand, omitting the proof (see [4].), For the gamma distribution when $a = 1$ we have (see. (35))

$$\rho_{X(\mathring{X})^3}(\tau) = \frac{(m_4 - m_1 m_3)\rho_{XX}(\tau)}{\sigma_X\sqrt{m_6 - m_3^2}} = \frac{\sqrt{3(a+1)(a+2)}}{\sqrt{3a^2 + 15a + 20}}\rho_{XX}(\tau). \tag{46}$$

It follows that there $A_{XX^3} \in [\sqrt{0,3};1] \approx [0,55;1]$. When $a = 0$ for $Y(t) = X^3(t)$ we have

$$\begin{gathered} m_Y(y;\tau) = \mathbf{M}\{Y(t+\tau)|Y(t) = y\} = [1 - \rho_{XX}(\tau)]^3 a(a+1)(a+2)\lambda^{-3}\times \\ \times\exp\{-\frac{\lambda y^{1/3}\rho_{XX}(\tau)}{1-\rho_{XX}(\tau)}\}{}_1F_1(a+3,a;\frac{y^{1/3}\rho_{XX}(\tau)}{1-\rho_{XX}(\tau)}), \end{gathered} \tag{47}$$

where ${}_1F_1(a,\beta;x)$ is a degeneracy hypergeometric function.

3. $X(t)$ - *is a strictly stationary process*

$$X(t) = \lambda\sin(\nu t + \Phi); \tag{48}$$

where Φ is a random variable with a uniform on $(-\pi,\pi)$ or $(0,2\pi)$ distribution [4]. This process also applies to LP-class [5]. For them, [4]

$$\rho_{XX}(\tau) = \cos(\nu\tau); \tag{49}$$

$$Y(t) = [\mathring{X}(t)]^3 = \lambda^3\sin^3(\nu t + \Phi) = \frac{\lambda^3}{4}[3\sin(\nu t + \Phi) - \sin(3(\nu t + \Phi))]; \tag{50}$$

It follows directly from (49) or by (33), we find

$$\mathbf{M}\{\mathring{X}(t)[\mathring{X}(t+\tau)]\} = \mathbf{M}\{[\mathring{X}(t-\tau)][\mathring{X}(t)]^3\} = \frac{3\lambda_4}{8}\rho_{XX}(\tau); \tag{51}$$

$\mu_2^2 = \frac{\lambda^2}{2}, \mu_3 = 0, \mu_4 = \frac{3}{8}\lambda^4, \mu_6 = \frac{15}{48}\lambda^6$, i.e.

$$\rho_{XY}(\tau) = \rho_{XX^3}(\tau) = \frac{3}{\sqrt{10}} \approx 0,95\rho_{XX}(\tau); \tag{52}$$

Similarly, we find that

$$R_{YY}(\tau) = \mu_{3,3}(\tau) = \frac{\lambda^6}{32}[\cos(3\nu\tau) + 9\cos(\nu\tau)], \tag{53}$$

$$\rho_{YY}(\tau) = \rho_{X^3X^3}(\tau) = 0,9\cos(\nu\tau) + 0,1\cos(3\nu\tau). \tag{54}$$

Note that $\mu_{3,3}(\tau) = \rho_{X^3X^3}(\tau)$ can be found on the two-dimensional distribution of the arc sine $N = 22$ table 4.7 in [4], having a place to RP (43), using the expansion of the distribution of orthogonal polynomials. Anyone can do it yourself. For this RP

$$m_Y(y;\tau) = \mathbf{M}\{X^3(t+\tau)|X^3(t) = y\} = \frac{\lambda^2}{2}[\cos^3(\nu\tau + \arccos\frac{y^{1/3}}{\lambda})] + \cos^3(\nu\tau - \arccos\frac{y^{1/3}}{\lambda}). \tag{55}$$

Conclusions

This work is the first part in a series of papers devoted to the study of random processes with nonlinear regression. The tasks 1-5 set therein define directions for further research. In this paper we obtain only partial solutions of the first and the second task. Of the first results of such research presented in this paper and related to random processes with linear regression and nonlinear transformations, we can draw the following conclusions.

1. Special cases confirmed and specified the general properties of LR processes.
2. The nature and degree of nonlinearity of dependency of NCF $\rho_{YY}(\tau)$ SP $Y(t) = f[X(t)]$ from NCF $\rho_{XX}(\tau)$ and defining not only on the functions $f(\cdot)$, but also the type and parameters of the distribution of the RP $X(t)$. At the same time cross NCF $\rho_{XY}(\tau)$ for the SP processes with linear regression is always proportional to $\rho_{XX}(\tau)$. The type and parameters of the distribution of $X(t)$ affect only on the values of the coefficients proportionality A .
3. Upon the lack of proportionality $\rho_{Xf(X)}(\tau)$ and $\rho_{XX}(\tau)$ for various $f(\cdot)$ can be inferred non-linearity of the regression function $X(t)$. This is the basis for developing classification algorithms RP for linear and nonlinear autoregression them.
4. Subject to the study of the question of what conclusions can be drawn from the knowledge of the values of A the coefficient of proportionality $\rho_{XY}(\tau) = A\rho_{XX}(\tau)$ for various $f(X)$ and distribution parameters.
5. Confirmed known [5], the fact of a possible reduction in the cross-correlation between the input and output non-linear element as compared to linear. This fact, together with the fact that parity $\rho_{Xf(X)}(\tau)$ allows to develop an algorithm for determining the presence of only a fast-response nonlinearity in the system of the "black box" on its input and output signals, as well as inertial units.

References

[1] Gradstein I.S., Ridgik I.M. (1976). *Tables of integrals, series and products.* Phizmatgiz, Moscow.

[2] Gubarev V.V. (2005). *Algorithms for spectral analysis of random signals.* Izdatelstvo NGTU, Novosibirsk.

[3] Gubarev V.V. (2014). *Introduction to theoretical computer science, part 1.* Izdatelstvo NGTU, Novosibirsk.

[4] Gubarev V.V. (1992). *Probability models: Handbook in 2 parts.* NETI, Novosibirsk.

[5] Raibman N.S. and other(1981). *Dispersion identification.* Science, Moscow.

Tests for an Absence of Trend[1]

Irina V. Veretelnikova and Boris Yu. Lemeshko
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: lemeshko@ami.nstu.ru, ira-veterok@mail.ru

Abstract

The properties of various parametric and nonparametric tests are studied using methods of statistical simulation. Such tests are designed to test hypotheses for randomness or absence of a trend in dispersion characteristics. Statistics distributions and the test powers are studied with respect to various competitive laws. Advantages and disadvantages of the studied tests are noted.

The procedure of interactive simulation of distributions of the test statistics is proposed and implemented. Such procedure allows making valid conclusions when using the test in the case of violation of standard assumptions.

Keywords: trend, hypothesis of randomness, statistical simulation, test power.

Introduction

A variety of parametric and nonparametric tests has been proposed at different times to test the hypothesis for randomness or absence of a trend in the mathematical expectation and in the dispersion characteristics. However, available sources do not allow us to judge the benefits of a particular test and do not contain any distinct recommendations on the area of application and prerequisites providing correctness of statistical conclusions when using the tests under consideration.

As a rule, assumption of normal distribution law of noise is the main prerequisite for ensuring the correct application of parametric tests, but it is not always realized in practice. The usage of nonparametric tests is based on asymptotic distribution of statistics of such tests. For limited sample sizes, the distributions of statistics of parametric and non-parametric tests may differ significantly from the corresponding limit distributions of statistics used for testing the hypothesis. The common disadvantage of nonparametric tests is an apparent discreteness of the statistics distribution. In such situations, the usage of the limiting (asymptotic) distribution of the statistics instead of the actual distribution of such statistics to test the hypothesis may lead to wrong conclusion.

In this paper, the methods of statistical simulation are used to investigate the statistic distributions and the power of tests for an absence of trend in a mathematical expectation, as well as the dispersion characteristics of the observed random variables.

When testing the absence of a trend in the mathematical expectation, it is assumed that time series of values $x_1, x_2, ...x_n$ of mutually independent random variables with mathematical expectations $m_1, m_2, ...m_n$ and equal (but unknown) variances are

[1] This research has been supported by the Russian Ministry of Education and Science (project 2.541.2014K).

observed. The hypothesis $H_0 : m_i = m, i = 1, 2, ..., n$ is tested that all sample values belong to the same population with mean m, against a competitive hypothesis about the presence of a trend $H_j : |m_{i+1} - m_i| > 0, i = 1, 2, ..., n - 1$.

When testing the absence of a trend in dispersion characteristics, the hypothesis $H_0 : s_i = s, i = 1, 2, ..., n$ is tested that all sample values belong to the same population with standard deviation s, against a competitive hypothesis for the presence of a trend $H_l : |s_{i+1} - s_i| > 0, i = 1, 2, ..., n - 1$.

When testing the absence of a variance shift (in dispersion characteristics) the hypothesis $H_0 : s_1^2 = ... = s_n^2 = s_0^2$ (s_0^2 being unknown) is tested against a competitive hypothesis

$$H_l : s_1^2 = s_2^2 = ... = s_k^2 = s_0^2; s_{k+1}^2 = ... = s_n^2 = s_0^2 + d; (d > 0),$$

for variance value changes in some unknown point (k unknown $1 \leq k \leq n - 1$).

1 Tests for an absence of trend in mathematical expectation research results

We have carried out the research of statistics distribution and the powers of parametric tests, which are used for testing the hypotheses of a trend absence in mathematical expectation (Autocorrelation test [1], Autocorrelation test modification [1], Dufor-Roy test [2], Ljung-Box test [3], Moran test [4], Wald-Wolfowitz test [5]), as well as non-parametric tests used for the same purposes (Wald-Wolfowitz rank test [5], Dufor-Roy rank test [2], Bartels test [6], Foster-Stewart test [7], Cox-Stuart test [8], Hollin test [16], Wald-Wolfowitz series test [5], Inversion test [9], Cumulative sum test [10, 11], series Wald-Wolfowitz test [5], series Ramachandran-Ranganathan test [12] and number of sign series of the first-order differences [13]).

The results of such research are briefly summarized in Table 1. The tests studied are arranged in the order of power decreasing. Table 1 shows main advantages and disadvantages of tests, noted during the research.

2 Tests for an absence of trend in dispersion characteristics

Statistical distributions and powers of non-parametric tests (Foster-Stewart test [7], Cox-Stuart test [8], Savage test [14, 12], Klotz test [14, 12])and parametric test (Hsu test [15]) which are used to test an absence of trend in dispersion characteristics, are studied here in more detail.

2.1 Foster-Stuart test

This nonparametric test can be used to test hypotheses of absence of a trend in the mean values or in the variances (dispersion characteristics) depending on the used statistics type. The test for an absence of trend in distribution characteristics is given by [7]:

$$S = \sum_{i=2}^{n} S_i, \tag{1}$$

where $S_i = u_i + l_i$;
$u_i = 1$, if $x_i > x_{i-1}, x_{i-2}, ...x_1$, otherwise $u_i = 0$;
$l_i = 1$, if $x_i < x_{i-1}, x_{i-2}, ...x_1$, otherwise $l_i = 0$.
It is clear that $0 \leq S \leq n-1$.

In the absence of a trend the normalized statistics

$$\tilde{t} = \frac{S - \mu}{\hat{\sigma}_S}, \tag{2}$$

where

$$\mu = 2\sum_{i=2}^{n} \frac{1}{i}, \hat{\sigma}_S = \sqrt{\mu - 4\sum_{i=2}^{n} \frac{1}{i^2}} \approx \sqrt{2\ln n - 3.4253},$$

are approximately described by Student's distribution with $\nu = n$ degrees of freedom. The hypothesis of absence of a trend is rejected at large modulus values of statistics (2).

Actually, the area of discrete values is the range of definition of $\tilde{t}$ statistics. The analysis of statistics distributions shows that even with relatively large sample sizes (around $n = 100, 200$) the discrete distributions of test statistics are significantly different from the Student distribution with n degrees of freedom [17, 18]. It follows that the use of achieved significance level (p-value) for calculations instead of the actual (discrete) distributions of statistics of asymptotic Student t-distributions can lead to serious errors.

2.2 Cox-Stuart test

Cox-Stuart test [8] for the hypothesis of an absence of a trend in variance (in dispersion characteristics) is designed as follows.

Initial sample $x_1, x_2, ...x_n$ is divided into $[n/k]$ subsamples with k number of elements $x_1, ..., x_k; x_{k+1}...x_{2k}; x_{2k+1}...x_{3k}; ...; x_{n-k+1}...x_n$ (if n is not divided by k, then the required number of measurements in the center is dropped out). For every ith subsample the range w_i is found $((1 \leq i \leq r, r = [n/k]))$. Then, the resulting sequence of ranges is tested against the trend in the mean values using the test with statistics

$$S_1^* = \frac{S_1 - E[S_1]}{\sqrt{D}[S_1]}, \tag{3}$$

where

$$S_1 = \sum_{i=1}^{[n/2]} (n - 2i + 1)h_{i,n-i+1}, E[S_1] = \frac{n^2}{8}, D[S_1] = \frac{n(n^2-1)}{24},$$

where $h_{i,j} = 1$, if $x_i > x_j$ and $h_{i,j} = 0$, if $x_i \leq x_j (i < j)$. If the hypothesis for the absence of a trend is true, distribution (3) can be approximately described by the standard normal law.

It is recommended to choose the value of k in [8] according to the following correlations:

$$n \geq 90 \rightarrow k = 5; 64 \leq n < 90 \rightarrow k = 4;$$
$$48 \leq n < 64 \rightarrow k = 3; n < 48 \rightarrow k = 2.$$

The discreteness of the S_1^* statistics distribution upon detection of a trend in the variance is significantly higher than the discreteness of the Cox-Stuart statistics distribution for trend in mean. This is natural because the analyzed range sample contains only $[n/k]$ number of elements. When using the Cox-Stuart test for detection of a trend in the dispersion, the difference of statistics discrete distribution from the standard normal law can almost be neglected only for $n > 170$ [19].

2.3 Hsu test for an absence of variance shift and shift point detecting

Under this test the rejection of the hypothesis of randomness (for absence of a trend) can show the discovery of a variance shift. Hsu test statistics are given by [15]

$$H = \frac{\sum_{i=1}^{n} (i-1)(x_i - m_x)^2}{(n-1) \sum_{i=1}^{n} (x_i - m_x)^2}, 0 \leq H \leq 1, \qquad (4)$$

where m_x is median of variation series. Under the assumption that the mathematical expectation of a sequence of random variables has the same value, the hypothesis of a constant variances is tested. As a competitive hypothesis, the change in the dispersion of observed values at some (unknown) time (starting from some element of the sample) can be considered. The test is two-sided: the tested hypothesis of absence of a variance shift is rejected for small and large values of the statistics (4).

Usually the test is used in a normalized form

$$H* = \frac{H - 1/2}{\sqrt{D}[H]}, whereD[H] = \frac{n+1}{6(n-1)(n+2)}. \qquad (5)$$

Under the validity of the hypothesis of the absence of variance changes, statistic (5) obeys the standard normal law asymptotically.

The simulation results [17] show that for $n > 30$ statistics distribution agrees well with the standard normal law.

Statistics distribution (5) strongly depends on the law of distribution to which random variables belong. The greatest deviation from the standard normal law is observed in the case when random variables belong to the laws with heavy tails. Asymmetry of the law significantly affects the statis-tics distribution.

A test allowing to determine the change point of the variance (in the case when observations belong to the normal law) is proposed in [15] of this test are presented as follows. Let for $k = 1, 2, ..., n-1$

$$w_k = \sum_{i=1}^{k} (x_i - m_x)^2, W_k = \tfrac{w_n - w_k}{w_k} \tfrac{k}{n-k},$$

where k corresponds to the required variance change point. If x_i belongs to normal law, then values of $W_k, k = 1, 2, ..., n-1$, belong to corresponding $F_{n-k,k}(W)$ Fisher distributions with $n-k$ and k degrees of freedom.

Next, based on the corresponding distribution functions, we find $\gamma_k = F_{n-k,k}(W_k)$, where γ_k must obey to uniform law under the absence of variance shift.

G-test statistics are given by

$$G = \frac{1}{n-1} \sum_{k=1}^{n-1} \gamma_k, 0 \le G \le 1. \tag{6}$$

The hypothesis about absence of variance changes is rejected with significance level α, if $G < G_{\alpha/2}$ or $G > G_{1-\alpha/2}$. In this case value k corresponding to the maximum value $|\gamma_k - 1/2|$, evaluates the desired change point of the variance value in observed series. For $x_1 = m_x$ value $w_1 = 0$, thus $W_1 = \infty$ and $\gamma_1 = 1$.

The type of limit distribution of the statistics (6) is not given in the original material, only percentage points are given. Basing on the results of the statistical simulations we have shown that a good model of the limit distribution of the statistics (6) is a beta distribution of the 1st kind with the density of

$$f(x) = \tfrac{1}{\theta_2 B(\theta_0, \theta_1)} (\tfrac{x-\theta_3}{\theta_2})^{\theta_0 - 1} (1 - \tfrac{x-\theta_3}{\theta_2})^{\theta_1 - 1}$$

and parameter values $\theta_0 = 2.7663, \theta_1 = 2.7663, \theta_2 = 1, \theta_3 = 0$.

Based on this law we can find percentage points $G_{\alpha/2}$ and $G_{1-\alpha/2}$ or p-values.

G-test is also a parametric test. Thus its statistics distributions depend strongly on the type of the law under observation.

2.4 Klotz and Savage rank tests for an absence of variance shifts

Rank tests for detecting the change of the scale parameter (dispersion characteristic) in the unknown point are based on the usage of a family of rank statistics in form [20]

$$S_R = \sum_{i=1}^{n} i a_n(R_i), \tag{7}$$

where R_i are ranks of sampled values in an ordered series of measurements.

Tests differ by the used scores a_n . Their type determines the name of the test. The following scores are commonly used:

- Klotz scores $a_{1n}(i) = U^2_{i/(n+1)}$, where U_γ – is a γ-quantile of standard normal law;
- Savage scores $a_{2n}(i) = \sum_{j=1}^{i} \frac{1}{n-j+1}$.

If the tested hypothesis H_0 is true, then tests with statistics $S_{R,j} = \sum_{i=1}^{n} i a_{jn}(R_i)$, $j = 1, 2$ are free from the distribution and are symmetric with respect to $E[S_{R,j}] = \frac{n+1}{2} \sum_{i=1}^{n} a_{jn}(i)$.

Usually normalized tests with the following statistics are used

$$S_{R,j}^{*} = \frac{S_{R,j} - E[S_{R,j}]}{\sqrt{D}[S_{R,j}]}, \tag{8}$$

where

$$\begin{gathered} E[S_{R,1}] = \tfrac{n+1}{2} \textstyle\sum_{i=1}^{n} U_{i/(n+1)}^{2}, E[S_{R,2}] = \tfrac{n(n+1)}{2}; \\ D[S_{R,1}] = \tfrac{n(n+1)}{12} \textstyle\sum_{i=1}^{n} U_{i/(n+1)}^{4} - \tfrac{1}{3n+3}[E[S_{R,1}]]^{2}; \\ D[S_{R,2}] = \tfrac{n(n+1)}{12}(n - \textstyle\sum_{j=1}^{n} \tfrac{1}{j}. \end{gathered}$$

Statistics (8) are approximately obeying the standard law. The convergence of the statistics distributions to the standard law was studied in [16, 20].

Statistical simulation research of the distribution of statistics with Klotz scores has shown that for $n > 20$ distribution is well-approximated by the standard normal law. Distribution of the test statistics with Savage scores also matches well with the standard normal law, but only for $n > 30$.

3 Analysis of the test powers

During analysis of test powers for the tests against variance change in an unknown point hypotheses close to the H_0 (in case of normal distribution of random variables) were treated as competitive, when at some point the standard deviation was increased by $5, 10, 15\%$:

$$\begin{gathered} H_1 : \sigma_1^2 = ...\sigma_k^2 = 1; \sigma_1^{k+1} = ...\sigma_n^2 = 1.1025, \\ H_2 : \sigma_1^2 = ...\sigma_k^2 = 1; \sigma_1^{k+1} = ...\sigma_n^2 = 1.21, \\ H_3 : \sigma_1^2 = ...\sigma_k^2 = 1; \sigma_1^{k+1} = ...\sigma_n^2 = 1.3225, \end{gathered}$$

where $k = n/2$. One competitive hypothesis was considered as more distant:

$$H_4 : \sigma_1^2 = ...\sigma_k^2 = 1; \sigma_1^{k+1} = ...\sigma_n^2 = 4.$$

The presence of a linear trend in the dispersion characteristics of the observed series of random variables (change in scale parameter) in the interval $t \in [0, 1]$ can be simulated according to

$$x_i = \xi_i(1 + ct_i),$$

where $c \in (-1, \infty), t_i = (i - 1) \triangle t, \triangle = 1/n$. True tested hypothesis H_0 corresponds to parameter value $c = 0$.

In case of a periodic trend in the characteristics of dispersion, random values can be simulated, for example, in accordance with the following formula:

$$x_i = \xi_i(1 + d \sin(2k\pi t_i)$$

for $|d| < 1$. In case of a combined trend it can be simulated according to

$$x_i = \xi_i(1 + ct_i + d\sin(2k\pi t_i)$$

for $|d| < 1$, if $c \geq 0$, and for $|d| < 1 + c$, if $c \in (-1, 0)$. The absence of a periodic component of the trend corresponds to the parameter value $d = 0$, and the absence of a linear component corresponds to $c = 0$.

During the analysis of power with respect to linear, periodic, and combined trend in the dispersion characteristics (in variance) of a random variable in the interval $t \in [0, 1]$ the following competitive hypotheses were considered:

$$H_5 : x_i = \xi_i(1 + ct_i), c = 1;\ H_6 : x_i = \xi_i(1 + d\sin(2k\pi t_i), d = 0.8, k = 2;$$
$$H_7 : x_i = \xi_i(1 + ct_i + d\sin(2k\pi t_i), c = 1, d = 0.8, k = 2.$$

At that, $t_i = (i - 1) \triangle t, \triangle t = 1/n$, and random variables x_i have been simulated according to the normal law with parameters m and s.

In the course of work statistical simulation methods (for probabilities of errors of the first kind $\alpha = 0.15, 0.1, 0.05, 0.01$) provided estimations of the capacity of the investigated criteria with respect to the competitive hypotheses H_1, H_2, H_3 and H_4 (corresponding to the shift of the dispersion value), and with respect to the competitive hypotheses H_5, H_6, H_7,corresponding to the presence of a linear or nonlinear trend in the characteristics of the dispersion process.

In the columns of Table 2 tests are ordered by decreasing power $1 - \beta$ according to the power estimations with respect to studied competitive hypotheses with the significance level $\alpha = 0.1$ and sample volume $n = 100$.

For similar competitive hypotheses criteria Hsu tests with H and G statistics as well as Klotz test showed the highest power with respect to the analyzed sets of competitive hypotheses. They showed the ability to detect trend in the dispersion characteristics when it has a 10% increase. Hsu tests with $H-$ and $G-$statistics and Klotz test are also detecting the presence of a linear or periodic trend in the dispersion characteristics (H_0 is distinguished from the hypotheses H_5, H_6).At the same time Cox-Stuart, Savage and Foster-Stuart tests can not detect the presence of a periodic trend in the variance reliably (due to relatively low power against similar enough hypothesis H_6). Unfortunately, none of these tests has shown the ability to detect a mixed trend in the dispersion corresponding to the studied hypothesis H_7. The power with respect to such close hypothesis has been extremely low.

Considered criteria can be placed in order of preference in the following way [22]:

Trend in mathematical expectation

K-inversion, Reversed inversion $\succ$ Inversions$\succ$ Cox-Stewart$\succ$ Autocorrelation test modification$\succ$ Ramachandran-Ranganathan $\succ$ Wald-Wolfowitz, autocorrelation, Dufour-Roy, Moran, Ljung-Box$\succ$ Wald-Wolfowitz rank, Rank Dufour-Roy, Hollin$\succ$ Bartels $\succ$ CUSUM $\succ$ Series Wald-Wolfowitz test $\succ$ Foster-Stewart $\succ$ Number of sign series of the first-order differences.

Trend in variance

$$HsuH - test \succ Klotztest \succ HsuG - test \succ Cox - Stewart \succ$$
$$Foster - Stewarttest \succ Savagetest.$$

Conclusions

Thus, methods of statistical simulation have been used to study the statistics distribution of various parametric and nonparametric tests for randomness and the absence of a trend in the dispersion characteristics; within the framework of developing ISW software an interactive study mode of the distributions of the statistics has been implemented for the case of violation of standard assumptions. A comparative analysis of test powers against some competitive hypotheses has been carried out, and results of such analysis can be used to estimate the desirability of application of particular test. Disadvantages of individual criteria have been noted.

References

[1] Knoke J.D. (1975). Testing for randomness against autocorrelation: The parametric case. *Biometrica*. Vol. **62**, pp. 571-575.

[2] Dufor J.-M., Roy R. (1985). Some robust exact results on sample for randomness. *J. of Econometrics*. Vol. **29**, pp. 257-273.

[3] Ljung G. M., Box G. E. P. (1978). On a measure of lack of fit in time series models.*Biometrika*. Vol. **65**, pp. 297-303.

[4] Moran P. A. P. (1948). Some theoremson time series 2: The siginificance of th serial correlations coefficient.*Biometrika*. Vol. **35**, pp. 255-260.

[5] Wald A., Wolfowitz J. (1940). On a test whether two samples are from the same population.*Ann. Math Statist*. Vol. **11**, pp. 147-162.

[6] Bartels R. (1982). The rank version of von Neumann's ratio test for randomness.*JASA*. Vol. **77**, pp. 40-46.

[7] Foster F.G., Stuart A. (1954). Distribution-free tests in time series dated on the breaking of records. *JRSS*. Vol. **B16,No.1**, pp. 1-22.

[8] Cox D.R., Stuart A. (1955). Quick sign tests for trend in location and dispersion.*Biometrika*. Vol. **42.**, pp. 80-95.

[9] Himmelblau D. (1973). *Process Analysis by Statistical Methods*. M.: Mir.

[10] Mc Gielchrist C.A., Woodyer K.D. (1975). Note on a distribution-free CISIM technique.*Technometrics*. Vol. **17, No.3**, pp. 321-325.

[11] Woodward R.H., Goldsmith P.L. (1964). *Cumulative sum techniques. I.C.I. Monograph. No.3.* Oliver and Boyd.

[12] Kobzar A.I. (2006). *Applied mathematical statistics for engineers and academic researchers* . M. : Fizmatlit.

[13] Hald A. (1956). *Mathematical Statistics and Its Applications*. M.: ill.

[14] Hsieh H.K. (1956). Nonparametric tests for scale shift at a unknown time point. *Commun. Stat. - Theor. Meth.* Vol. **13. No.11**, pp. 1335-1355.

[15] Hsu D.A. (1977). Test for variance shift at an unknown time point.*Appl. Statist.*. Vol. **26, No.3**, pp. 279-284.

[16] Hollin M., Ingeubleek J.-F., Puri M.L. (1985). Linear serial rank tests for randomness against ARMA alternatives.*Appl. Statist.*. Vol. **13**, pp. 1156-1181.

[17] Lemeshko B.Yu, Komissarova A.S., Shcheglov A.Ye. (2010). Application of some tests for hypotheses about randomness or trend absence.*Metrologiya.* **No.12**, pp. 3-25.

[18] Veretelnikova I.V., Lemeshko B.Yu. (2014). Analytical review of tests for randomness and trend absence.*Materials of the XII international conference Actual problems of electronic instrument engineering.* Vol. **6**. Novosibirsk, pp. 16-23.

[19] Veretelnikova I.V., Lemeshko B.Yu. (2014). Application of tests for hypotheses about randomness or trend absence.*Cutting-edge technologies, fundamental researches and innovations: pressing of the XVII international scientific-practical conference High Technologies in Industry and Economics , 22-23th May 2014, St. Petersburg, Russia/ Science editors: A.P. Kudinov, M.A. Kudinov. – SPb: Publishing house of the Polytechnic University, 2014*, pp. 33-37.

[20] Gayek Ya., Shidak Z. (1971). *Theory of rank tests*. M.: Chief editorial board of physical and mathematical literature.

[21] Lemeshko B.Yu, Komissarova A.S., Shcheglov A.Ye. (2012). Properties and power of tests for trend detection and checking for randomness.*Scientific bulletin of NSTU.* **No.1(46)**, pp. 53-66.

[22] Veretelnikova I.V., Lemeshko B.Yu.. (2014). Application of tests for randomness or trend absence. *Materials of the Russian scientific-practical conference Data processing and mathematical simulation* . Novosibirsk, pp. 25-28.

Table 1: Main advantages and disadvantages of used tests for an absence of trend in mean

№	Test	Advantages	Disadvantages
1	Inversion	High power in respect to linear trend. For $n \geq 30$ discreteness of normalized statistics can be neglected.	The discreteness of normalized statistics must be considered for $n < 30$.
2	Reversed inversion		
3	K-inversion		
4	Cox-Stuart	Power is above the average. For $n \geq 40$ discreteness of normalized statistics can be neglected.	For $n < 40$ discreteness of normalized statistics must be considered.
5	Autocorrelation test modification	Relatively good power.	The difference of normalized statistics distribution from the standard normal law can be neglected only for $n \geq 200$
6	Ramachandran-Ranganathan	Relatively good power.	Statistics distribution have strong dependence on n. Usage of a table of critical values is necessary.
7	Dufour-Roy	The difference of normalized statistics discrete distribution from the standard normal law can be neglected for $n > 17$.	Low power.
8	Autocorrelation	The difference of normalized statistics distribution from the standard normal law can be neglected for $n > 30$.	Low power.
9	Moran		Low power. The difference of statistics distribution from the standard normal law can be neglected only for $n > 50$.
10	Ljung-Box		Low power. Statistics distribution converge very slowly to standard normal law.
11	Wald-Wolfowitz	The difference of normalized statistics distribution from the standard normal law can be neglected for sample sizes $n > 20$.	Low power.

12	Hollin	Average power.	Distribution of the statistics depends on n. The test is nonparametric, yet distribution of the statistics reacts to asymmetry of the observed law.
13	Rank Wald-Wolfowitz	Standard normal law can be used for $n > 10$ as distribution of the proposed modification of normalized statistics.	The power is slightly smaller than one of Dufour-Roy and Wald-Wolfowitz tests. Is equal to rank Dufour-Roy test.
14	Rank Dufour-Roy	For $n > 17$ distribution of the statistics is well-approximated by standard normal law. Discreteness of statistics distribution can be neglected for $n > 10$.	The power is slightly smaller than one of Dufour-Roy and Wald-Wolfowitz tests. Is equal to rank Wald-Wolfowitz test.
15	Bartels	The difference of normalized statistics discrete distribution from the standard normal law can be neglected for $n > 10$.	Low power.
16	Foster-Stuart		High discreteness of statistics distribution, persisting for hign values of n. Usage of assymptotic Student t_n-distribution for evaluation of p-value leads to serious errors. Power against linear trend is below the average. Power against nonlinear trend is low.
17	CUSUM	Good power against linear trend.	Statistics distribution is discrete and it is dependent on n. Very low power against nonlinear trend.
18	Series Wald-Wolfowitz		Normalized statistics distribution is discrete for a long time. Low power.
19	Number of sign series of the first-order differences		Normalized statistics distribution is discrete even for large sample sizes.Extremely low power.

Table 2: Comparative analysis of powers of all tests for randomness and tests for an absence of a trend in variances ($n = 100, \alpha = 0.1$)

№	Against H_1	$1-\beta$	Against H_2	$1-\beta$	Against H_3	$1-\beta$
1	Hsu H	0.156	Hsu H	0.304	Hsu H	0.500
2	Klotz	0.151	Klotz	0.287	Klotz	0.469
3	Hsu G	0.147	Hsu G	0.269	Hsu G	0.430
4	Cox-Stuart	0.123	Cox-Stuart	0.188	Cox-Stuart	0.284
5	Savage	0.110	Foster-Stuart	0.130	Foster-Stuart	0.165
6	Foster-Stuart	0.106	Savage	0.129	Savage	0.159

№	Against H_4	$1-\beta$	Against H_5	$1-\beta$	Against H_6	$1-\beta$
1	Hsu H	1	Hsu H	0.836	Hsu H	0.711
2	Klotz	1	Hsu G	0.818	Klotz	0.678
3	Cox-Stuart	0.997	Klotz	0.807	Hsu G	0.545
4	Hsu G	0.993	Cox-Stuart	0.489	Savage	0.196
5	Foster-Stuart	0.625	Foster-Stuart	0.346	Cox-Stuart	0.143
6	Savage	0.610	Savage	0.246	Foster-Stuart	0.048

№	Against H_7	$1-\beta$
1	Hsu H	0.162
2	Klotz	0.104
3	Savage	0.095
4	Foster-Stuart	0.082
5	Hsu G	0.057
6	Cox-Stuart	0.052

The Comparative Analysis of Tests in the Problem of Testing the Hypothesis of Uniformity[1]

Pavel Yu. Blinov and Boris Yu. Lemeshko
Novosibirsk State Technical University, Novosibirsk, Russian Federation
e-mail: blindizer@yandex.ru,lemeshko@ami.nstu.ru

Abstract

In the paper some statistical tests intended for testing of uniformity have been considered. Distributions of test statistics, the power of tests under different competing hypotheses have been studied. Considered tests have been ranked by the test power. Advantages and disadvantages of individual tests have been shown. Also, it has been shown that the large part of the tests traditionally used for testing uniformity has the bias under some kind of competing hypotheses. It is underlines that special uniformity tests haven't clear advantage over nonparametric goodness-of-fit tests used for testing uniformity in general.

Keywords: uniform distribution, hypothesis testing, test statistic, test power.

Introduction

The uniform distribution is one of common distributions in applied mathematics statistics and probability theory. It is often used to describe the measurement error of some instruments or measuring systems. Simulation of pseudorandom values according to different parametric laws relies on sensors of uniform pseudorandom values. Parametric laws are urgently needed in the systems of statistical simulation. Testing the uniformity actually represents goodness-of-fit testing the hypothesis of uniform distribution of the observed sample $x_1, ..., x_n$. In some papers, the authors states that testing composite hypothesis can be reduced to test simple hypothesis of uniformity on the interval $[0, 1]$, because if $x_1, ..., x_n$ belong law with probability distribution function $F(x)$, then random variable $y_i = F(x_i)$ is uniformly distributed on unit interval. All of these factors explain the increasing interest in the choice of simple and computationally efficient procedures for testing hypotheses about the uniform law of analyzed samples.

The various statistical tests used for testing hypothesis of uniformity can be divided into two subsets. These are common goodness-of-fit tests applicable for testing of uniformity and special tests oriented on testing hypothesis that sample $x_1, ..., x_n$ is uniform distributed.

The presence of numerous tests put not simple problem of choosing for specialists, because available information in papers doesn't allows to give preference to certain test, while every specialist is interested not only in correctness of using of tests, but else in reliability of statistical inferences.

[1] This work is supported by the Russian Ministry of Education and Science (project 2.541.2014K).

In this paper, a lot of considered tests are studied by the method of statistical simulations. The number of experiments carried out for statistical modeling is assumed equal to 1 660 000 in the study of the distributions of test statistics. On the one hand, such number of experiments allows tracing the qualitative picture of test statistic distributions in depend on various factors. On the other hand, this number of experiments provides acceptable accuracy of the power estimates and unknown probabilities.

1 The statement of testing uniformity

In the most of uniformity tests, ordered statistics of quantity X are used ($x_{(1)} < x_{(2)} < ... < x_{(n)}$ are elements $x_{(i)}$ of variation series of the sample). Further designation $U_i = x_{(i)}, i = \overline{1,n}$ will be used in expressions of statistical tests.

As usually tests are oriented on testing of simple hypothesis H_0 on interval $[0, 1]$. However, if hypothesis of uniformity is tested on interval $[a, b]$ then elements $x_{(i)}$ of variation series $a < x_{(1)} < x_{(2)} < ... < x_{(n)} < b$ are modified to corresponding (required in the tests) ordered statistics as: $U_i = \frac{x_{(i)}-a}{b-a}, i = \overline{1,n}$.

To test composite hypothesis of uniformity H_0: $F(x) = (x - a)/(b - a)$, $x \in [a, b]$, where a and b are non-known, we proceed as follows. Using the variation series $x_{(1)} < x_{(2)} < ... < x_{(n)}$ of sample $X_1, X_2, ..., X_n$ the parameter estimates are obtained as follows:

$$\hat{a} = x_{(1)} - \frac{x_{(n)} - x_{(1)}}{n-1}, \hat{b} = x_{(n)} + \frac{x_{(n)} - x_{(1)}}{n-1}. \tag{1}$$

It is obviously that testing of composite hypothesis of uniformity for sample $X_1, X_2, ..., X_n$ on interval $[\hat{a}, \hat{b}]$ equal to testing of simple hypothesis of uniformity for sample with sample size $n-2$ on interval $[x_{(1)}, x_{(n)}]$. The required values of order statistics for testing such hypothesis obtained by expressions: $U_{i-1} = \frac{x_{(i)}-x_{(1)}}{x_{(n)}-x_{(1)}}$, $i = \overline{2,(n-1)}$.

A number of considered tests can be divided into three groups. The first group has statistics based on interval between elements, in most of cases differences between neighbor elements denoted as:

$$D_i = U_i - U_{i-1}, \tag{2}$$

where $U_0 = 0, U_{n+1} = 1, n$ is the size of the sample. In the second group test statistics used difference between theoretical (expected) and empirical data. These tests also called as tests based on the empirical distribution function (EDF tests), and goodness-of-fit tests are contained in this group. The third group has statistics based on entropy estimator. The third group includes the tests based on the entropy estimator.

2 Alternative hypotheses

We compared the power of tests for relatively sample size $n = 10, 20, 30, 40, 50, 100, 150, 200, 300$. Empirical distributions of test statistics under either true null hypothesis or competing hypotheses were found based on 1 660 000 simulations also. The hypothesis under test H_0 was chosen as uniform law. Alternative hypothesis H_i was chosen as beta distribution with the density

$$f(x) = \frac{1}{\theta_2 B(\theta_0, \theta_1)} \left(\frac{x - \theta_3}{\theta_2} \right)^{\theta_0 - 1} \left(1 - \frac{x - \theta_3}{\theta_2} \right)^{\theta_1 - 1}, \tag{3}$$

where $B(\theta_0, \theta_1) = \Gamma(\theta_0)\Gamma(\theta_1)/\Gamma(\theta_0 + \theta_1)$ is beta-function, $\theta_0, \theta_1 \in (0, \infty)$ are parameters the of form, $\theta_2 \in (0, \infty)$ is shape parameter, $\theta_3 \in (-\infty, \infty)$ is bias parameter, $x \in [0, \infty]$. This distribution was chosen because the fact that the standard uniform distribution is a special case of the beta distribution with the parameters of form $\theta_0 = 1$ and $\theta_1 = 1$. We denote the function of beta distribution with values of parameters $B_I(\theta_0, \theta_1, \theta_2, \theta_3)$. So, three alternative hypotheses H_1, H_2, H_3, which are quite close to H_0, can be written by

$$H_1 : F(X) = B_I(1.5, 1.5, 1, 0), x \in [0, 1];$$
$$H_2 : F(X) = B_I(0.8, 1.0, 1, 0), x \in [0, 1];$$
$$H_3 : F(X) = B_I(1.1, 0.9, 1, 0), x \in [0, 1] \ .$$

The distribution functions and the density functions of these hypotheses are presented in Figure 1 and 2, respectively.

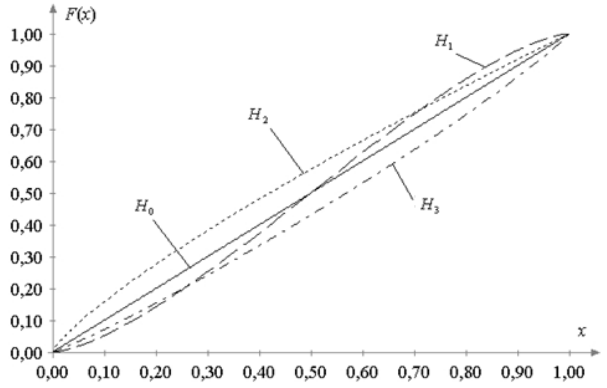


Figure 1: The distribution functions corresponding to the hypotheses

It is worth noting that the distribution function of alternative H_1 crossed the function of the uniform distribution, while the distribution functions of alternatives H_1 and H_3 are located above and below the function of uniform distribution, respectively. And abilities to distinguish hypothesis H_0 from H_1 and from H_2 and H_3 in tests are different. The comparative analysis shows that most of the considered tests have inability to distinguish hypothesis H_0 from H_1 under small sample size n , in other words these tests are biased in such cases.

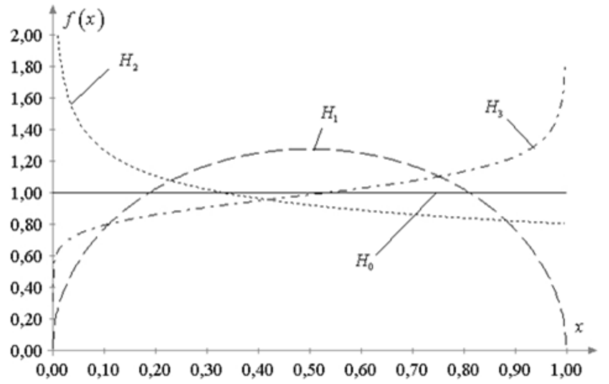


Figure 2: The density functions corresponding to the hypotheses

3 Simulation result

The expressions for statistics of special uniformity tests are presented in Table 1. The Table 2 contains considered tests ordered by decreasing of power (quantity $1 - \beta$) under alternatives H_1, H_2 and H_3 ($n = 100$ and $\alpha = 0.1$). The dark mark means that the test is biased under small sample size n, in other words that quantity α larger than $1 - \beta$. This bias take a place to a lesser extent in Neyman-Barton tests N_2 and N_3 [14]. This advantage isn't observed only for some tests: Kuper test [9], Watson test [19, 20], Dudewicz-Van Der Mulllen test [5], Cheng-Spiring test [3], Swartz test [18], second Cressie [4] test and chi-squared Pearson test.

Entropy procedure used different entropy estimator gives high power under alternative hypothesis H_1. Whereas their power is relatively worst under alternatives H_2 and H_3. It should be noted that only modifications of entropy test have bias under alternative H_2 for small sample size n. It is recognized that power of these tests and also Cressie tests and Pardo test [15] depends from choosing of parameter m called as window size also.

The Neyman-Barton test N_2 shows good power under H_1 and relatively good power under H_2 and H_3. The Hegazy-Green tests [7] and Frosini test demonstrate consistently good ability to distinguish alternative hypotheses from uniformity distribution. The low powers are shown by tests, the statistics of which use the differences (2) of successive values of order sample $U_i - U_{i-1}$ (Sherman test [17], Kimball test [8], Moran tests [12, 13], Greenwood test [6], Greenwood-Quesenberry-Miller test [16]). The Cheng-Spiring test, demonstrated quite high power under H_1, shows low power under H_2 and H_3. The lowest power is demonstrated by Yang test [22], under all considered alternative hypotheses. Among the non-parametric goodness-of-fit tests, the good powers are obtained by Zhang tests Z_A and Z_C [24], and Anderson-Darling tests [1].

Conclusions

Unfortunately, the distributions of most special uniformity tests depend on the sample size, therefore the researchers must rely on the tables of percent points. The similar issue occurs in using nonparametric goodness-of-fit Zhang tests.

It is found from comparative analysis of tests, which can be used for testing the hypothesis of uniformity, that using of single certain test can be incorrect in forming the reliable statistical inference. The applying more than one test based on different measure of deviation of empirical distribution from theoretical distribution improves the quality of statistical inference. It is better to use some series of tests, which have certain advantages for more objective inferences.

References

[1] Anderson T.W., Darling D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*. Vol. **49**, pp. 765-769.

[2] Blinov P.Yu., Lemeshko B.Yu. (2014). Entropy-based test of uniformity. *12th International Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*. Vol. **1**, pp. 540-548.

[3] Cheng S.W., Spiring F.A. (1987). A test to Identify the uniform distribution with applications. *IEEE Transactions on Reliability*. Vol. **36**, pp. 98-105.

[4] Cressie N. (1979). An optimal statistic based on higher order gaps. *Biometrika*. Vol. **66**, pp. 619-627.

[5] Dudewics E.J., van der Meulen E.C. (1981). A review of the properties of tests for uniformity. *Journal of the American Statistical Association*. Vol. **76**, pp. 967-974.

[6] Greenwood V. (1946). The statistical study of Infection disease. *Journal of the Royal Statistical Society: Series A*. Vol. **109**, pp. 257-261.

[7] Hegazy Y.A., Green J.R.S. (1975). Some new goodness-of-fit tests using order statistics. *Applied Statistics.*. Vol. **24**, pp. 299-308.

[8] Kimball B.F. (1947). Some basic theorems for developing tests of fit for the case of the non-parametric probability distribution function. *The Annals of Mathematical Statistics*. Vol. **18**, pp. 540-541.

[9] Kuiper N.H. (1960). Tests concerning random points on a circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen. Ser. A.*. Vol. **63**, pp. 38-47.

[10] Lemeshko B.Yu., Gorbunova A.A. (2013). Application and Power of the Nonparametric Kuiper, Watson, and Zhang Tests of Goodness-of-Fit. *Measurement Techniques*. Vol. **56**, pp. 465-475.

[11] Lemeshko B.Yu., Gorbunova A.A, Lemeshko S.B., Rogozhnikov A.P. (2013). Solving problems of using some nonparametric goodness-of-fit tests. *Optoelectronics, Instrumentation and Data Processing.*. Vol. **50**, pp. 21-35.

[12] Moran P.A.P. (1947). The random division of an intervals. *Journal of the Royal Statistical Society: Series B*. Vol. **9**, pp. 92-98.

[13] Moran P.A.P. (1951). The random division of an intervals, II. *Journal of the Royal Statistical Society: Series B*. Vol. **13**, pp. 147-150.

[14] Neyman J. (1937). "Smooth" tests for goodness-of-fit. *Scandinavisk Aktuarietidskrift*. Vol. **20**, pp. 149-199.

[15] Pardo M.C. (2003). A test for uniformity based on informational energy. *Statistical Papers*. Vol. **44**, pp. 521-534.

[16] Quesenberry C.P, Miller F.L. (1977). Power studies of some tests for uniformity. *Journal of Statistical Computation and Simulation*. Vol. **5**, pp. 169-191.

[17] Sherman B.A. (1950). A random variable related to the spacing of sample values. *The Annals of Mathematical Statistics*. Vol. **21**, pp. 339-361.

[18] Swartz T. (1992). Goodness-of-fit tests using Kullback–Leibler information. *Communications in Statistics – Theory and Methods*. Vol. **21**, pp. 711-729.

[19] Watson G.S. (1961). Goodness-of-fit tests on a circle. *Biometrika*. Vol. **48**, pp. 109-114.

[20] Watson G.S. (1962). Goodness-of-fit tests on a circle, II. *Biometrika*. Vol. **49**, pp. 57-63.

[21] Yousefzadeh A., Arghami N.R. (2008). Testing Exponentiality Based on Type II Censored Data and a New cdf Estimator. *Communications in Statistics - Simulation and Computation*. Vol. **37**, pp. 1479-1499.

[22] Young D.L. (1982). The linear nearest neighbour statistic. *Biometrika*. Vol. **69**, pp. 477-480.

[23] Zamanzade E. (2014). Testing uniformity based on new entropy estimators. *Journal of Statistical Computation and Simulation*. pp. 1-15, DOI: 10.1080/00949655.2014.958085.

[24] Zhang J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society: Series B*. Vol. **64**, pp. 281-294.

Table 1: Statistics of considered tests for uniformity

Number	Test	Test statistic
1	Sherman	$\omega_n = \frac{1}{2}\sum\limits_{i=1}^{n+1} \lvert D_i - \frac{1}{n+1} \rvert$
2	Kimball	$A = \sum\limits_{i=1}^{n+1} (D_i - \frac{1}{n+1})^2$
3	Moran 1	$B = \sum\limits_{i=1}^{n+1} (D_i)^2$
4	Moran 2	$M_n = -\sum\limits_{i=1}^{n+1} \ln\left[(n+1)D_i)\right]$
5	Yang	$M = \frac{1}{l}\sum\limits_{i=1}^{n} \min(D_i, D_{i+1});\ l = b - a$
6	Greenwood	$G = (n+1)\sum\limits_{i=1}^{n+1} (D_i)^2$
7	Greenwood-Qesenberry-Miller	$Q = \sum\limits_{i=1}^{n+1} (D_i)^2 + \sum\limits_{i=1}^{n} (D_{i+1}D_i)$
8	Swartz	$A_n^* = \frac{n}{2}\sum\limits_{i=1}^{n} \left(\frac{U_{i+1} - U_{i-1}}{2} - \frac{1}{n}\right)^2,$ where $U_0 = -U_1$, $U_{n+1} = 2 - U_n$
9	Cressie 1	$S_n^{(m)} = \sum\limits_{i=0}^{n+1-m} \left(U_{i+m} - U_i - \frac{m}{n+1}\right)^2,\ m < \frac{n}{2}$
10	Cressie 2	$L_n^{(m)} = \sum\limits_{i=0}^{n+1-m} \ln[\frac{n+1}{m}(U_{i+m} - U_i)],\ m < \frac{n}{2}$
11	Cheng-Spiring	$W_p = \left[(U_n - U_1)\frac{n+1}{n-1}\right]^2 / \sum\limits_{i=1}^{n} \left(U_i - \bar{U}\right)^2$
12	Hegazy-Green T_1	$T_1 = \frac{1}{n}\sum\limits_{i=1}^{n} \lvert U_i - \frac{i}{n+1} \rvert$
13	Hegazy-Green T_1^*	$T_1^* = \frac{1}{n}\sum\limits_{i=1}^{n} \lvert U_i - \frac{i-1}{n-1} \rvert$
14	Hegazy-Green T_2	$T_2 = \frac{1}{n}\sum\limits_{i=1}^{n} \left(U_i - \frac{i}{n+1}\right)^2$
15	Hegazy-Green T_2^*	$T_2^* = \frac{1}{n}\sum\limits_{i=1}^{n} \left(U_i - \frac{i-1}{n-1}\right)^2$
16	Frosini	$B_n = \frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n} \lvert U_i - \frac{i-0.5}{n} \rvert$
17	Neyman-Barton N_k; $k = 2, 3, 4$	$N_k = \sum\limits_{j=1}^{k} V_j^2$, where $V_j = \frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n} \pi_j(U_i - 0.5),$ $\pi_1(y) = 2\sqrt{3}y;\ \pi_2(y) = \sqrt{5}(6y^2 - 0.5);$ $\pi_3(y) = \sqrt{7}(20y^3 - 3y);$ $\pi_4(y) = 3(70y^4 - 15y^2 + 0.375)$

Table 1 (continued)

Number	Test	Test statistic
18	Dudewicz-Van Der Mulen	$H(m,n) = -\frac{1}{n}\sum\limits_{i=1}^{n}\ln\left[\frac{n}{2m}(U_{i+m}-U_{i-m})\right],$ where $m < \frac{n}{2}$; if $i+m \geq n$, then $U_{i+m} = U_n$, and if $i-m \leq 1$, then $U_{i-m} = U_1$
19	Pardo	$E_{m,n} = \frac{1}{n}\sum\limits_{i=1}^{n}\frac{2m}{n(U_{i+m}-U_{i-m})}$
20	The first modification of entropy test [23],	$HY_1 = -\frac{1}{n}\sum\limits_{i=1}^{n}\ln\left(\frac{U_{i+m}-U_{i-m}}{\hat{F}(U_{i+m})-\hat{F}(U_{i-m})}\right)$, where $\hat{F}(U_i) = \frac{n-1}{n(n+1)}\left(i+\frac{1}{n-1}+\frac{U_i-U_{i-1}}{U_{i+1}-U_{i-1}}\right)$, $i = \overline{2,(n-1)}$, $\hat{F}(U_1) = 1-\hat{F}(U_n) = \frac{1}{n+1}$
21	The first modification of entropy test [21],	$HY_2 = -\sum\limits_{i=1}^{n}\ln\left(\frac{U_{i+m}-U_{i-m}}{\hat{F}(U_{i+m})-\hat{F}(U_{i-m})}\right) * \left(\frac{\hat{F}(U_{i+m})-\hat{F}(U_{i-m})}{\sum\limits_{j=1}^{n}\left(\hat{F}(U_{j+m})-\hat{F}(U_{j-m})\right)}\right)$

Table 2: The tests ranked by power ($n = 100$,$\alpha = 0.1$)

	hypothesis H_1	$1-\beta$	hypothesis H_2	$1-\beta$	hypothesis H_3	$1-\beta$
1	The second modification of entropy test	0.883	Anderson–Darling	0.648	Anderson–Darling	0.526
2	Zhang Z_A	0.850	Hegazy-Green T_1	0.610	Hegazy-Green T_1	0.522
3	Neyman-Barton N_2	0.837	Zhang Z_C	0.606	Frosini	0.522
4	Cressie 2	0.820	Frosini	0.603	Hegazy-Green T_1^*	0.520
5	Zhang Z_C	0.819	Hegazy-Green T_2	0.602	Hegazy-Green T_2	0.508
6	Dudewicz-Van Der Mulen	0.790	Neyman-Barton N_2	0.597	Kramer-von-Misses-Smirnov	0.507
7	The first modification of entropy test	0.789	Kramer-von-Misses-Smirnov	0.595	Hegazy-Green T_2^*	0.506
8	Watson	0.779	Hegazy-Green T_1^*	0.595	Zhang Z_C	0.463
9	Neyman-Barton N_3	0.766	Zhang Z_K	0.590	Zhang Z_A	0.459
10	Neyman-Barton N_4	0.739	Hegazy-Green T_2^*	0.585	Kolmogorov	0.450

Table 2 (continued)

	hypothesis H_1	$1-\beta$	hypothesis H_2	$1-\beta$	hypothesis H_3	$1-\beta$
11	Kuper	0.736	Neyman-Barton N_3	0.577	Neyman-Barton N_2	0.447
12	Cheng-Spiring	0.722	Zhang Z_A	0.574	Zhang Z_K	0.438
13	Zhang Z_K	0.617	Neyman-Barton N_4	0.557	Neyman-Barton N_3	0.416
14	Pearson χ^2	0.593	Kolmogorov	0.542	Neyman-Barton N_4	0.381
15	Swartz	0.583	Pardo	0.463	Pearson χ^2	0.374
16	Anderson–Darling	0.505	Pearson χ^2	0.448	Pardo	0.291
17	Hegazy-Green T_1^*	0.443	Kuper	0.364	Dudewicz-Van Der Mulen	0.275
18	Hegazy-Green T_2^*	0.409	Watson	0.356	The first modification of entropy test	0.275
19	Pardo	0.408	The first modification of entropy test	0.328	The second modification of entropy test	0.267
20	Frosini	0.384	Dudewicz-Van Der Mulen	0.327	Watson	0.257
21	Kramer-von-Misses-Smirnov	0.358	Cressie 1	0.314	Kuper	0.254
22	Hegazy-Green T_1	0.322	The second modification of entropy test	0.266	Cressie 2	0.226
23	Kolmogorov	0.322	Greenwood-Qesenberry-Miller	0.244	Cressie 1	0.218
24	Hegazy-Green T_2	0.308	Swartz	0.226	Swartz	0.206
25	Greenwood-Qesenberry-Miller	0.290	Cressie 2	0.217	Greenwood-Qesenberry-Miller	0.186
26	Kimball	0.279	Sherman	0.204	Kimball	0.165
27	Moran 1	0.279	Kimball	0.201	Moran 1	0.165
28	Greenwood	0.279	Moran 1	0.201	Greenwood	0.165
29	Sherman	0.215	Greenwood	0.201	Sherman	0.154
30	Cressie 1	0.187	Moran 2	0.193	Moran 2	0.143
31	Moran 2	0.187	Cheng-Spiring	0.168	Cheng-Spiring	0.106
32	Yang	0.115	Yang	0.108	Yang	0.104

Comparisions by MC method the Mean and Mid-range as Estimators of Measurand for Samples from Uniform and Flatten-Gaussian Populations

Zygmunt L. Warsza[1] and Stefan Kubisa[2]

[1] *Industrial Research Institute for Automation and Measurements (PIAP), Poland*

[2] *West Pomeranian University of Technology, Faculty of Electrical Engineering, Poland*

e-mail: zlw@op.pl, kubisa@zut.edu.pl

Abstract

In this paper the statistical properties of mid range and mean as estimators of measured value, for the samples of varying number of observations taken from a population of uniform distribution, have been examined by the Monte Carlo simulation. The midrange of such samples has a smaller standard deviation than the mean value, which is recommended by the Guide GUM (fig. 1). A distribution similar to Student's t-distribution and an expanded uncertainty were also calculated for such samples. It was found for samples from the general population of Flatten-Gaussian distribution, that with increasing share of the normal distribution, the advantage of mid-range quickly decreases. Final conclusions are enclosed.

Introduction

The metrologically correct result of a measurement should contain the most probable value of a measurand together with an assessment of its accuracy. It should be determined in widely accepted uniform manner. Seven international organizations recommend the procedure described in the guide known by the acronym GUM [1]. It assumes that the observations are independent and can be treated as if they are taken from a normally distributed population and there are no outliers or been removed. In the most of laboratory measurements these assumptions are typically fulfilled and the uncertainties are determined for two or three significant digits. A description of the instrument accuracy by the worse case of limited errors is also used.

Notice: the statistical approach to unknown systematic errors and to calculations of the final result accuracy, nearly similar as in Guide GUM [1], was proposed 40 years earlier by S. Trzetrzewinski PhD work in 1951 at Gdansk Technical University [4]. In the GUM this approach is presented widely and using another terminology (e.g. the most probable final error - is the uncertainty) and GUM is now internationally sanctioned.

Measurements and processing of the measurement data carried on in science, industry and many other fields commonly use now electronic and computers. Some of them do not fulfill the assumptions of GUM. The distribution of measured values, or components of a random signal is often better modeled by Non-Gaussian distributions. There are also distortions (random, continuous or intermittent) – so called outliers. Sometimes there is a need to make statistical evaluations from the samples of low number of elements. Since the mid-twentieth century the new statistical tools were developed, such as robust and resampling methods, to analyze these issues.

The statistical properties of samples from a population of uniform distribution and few different flatten-Gaussian distribution in varying degree will be examined in detail by using Monte Carlo simulation.

1 Basic Equations

The classic approach of the measurement uncertainty calculation is in [1] and [2]. It is based on an assumption that the randomness of the observed N values of x is the source of their origin from the general population with normal distribution. After elimination of the known systematical errors from measurement data, the best estimate of the measured value is determined as the arithmetic mean of the empirical sample:

$$x_{av} = \frac{1}{N} \cdot \sum_{n=1}^{N} x_n = \frac{1}{\nu+1} \cdot \sum_{n=1}^{\nu+1} x_n \tag{1}$$

wherein $\nu = N - 1$ is the number of degrees of freedom.

The estimator of the standard deviation of average is

$$s_{cl} = \sqrt{\frac{\sum\limits_{n=1}^{N} (x_{av} - x_n)^2}{N \cdot (N-1)}} = \sqrt{\frac{\sum\limits_{n=1}^{\nu+1} (x_{av} - x_n)^2}{(\nu+1) \cdot \nu}}. \tag{2}$$

This deviation of the sample, determined by statistical method is named in [1] as a standard uncertainty $u_A(x)$. In addition, based on the knowledge of the observer, a standard uncertainty $u_B(x)$ is estimated. Then the combined standard uncertainty $u_C(x)$ is calculated. Considering the expansion coefficient k_p for the confidence level P, or using Monte Carlo method [2], the expanded uncertainty is $U(x) = k_P \cdot u_C(x)$.

In Supplement 1 of GUM [2] Monte Carlo method is recommended as the most universal, based on elementary mathematical relationships, possible to apply for the highly nonlinear measurement functions, as well as in cases of unusual, for example, asymmetrical distributions, as in [5]. If it is known that the observations come from different general population and the probability distribution is also given, it is better to use an approach called here: special.

Cramer, in his the excellent timeless monograph [3] for samples from a population with uniform distribution demonstrated analytically that a mid range is better estimator of a measurand than a mean due to having the smaller standard deviation. This is confirmed by numerical examples in [7] - [9] and the distributions of the three estimators of the samples with high cardinality N obtained by the MC method - Figure 1 [9]. Basic parameters of the sample are presented in the Table 1.

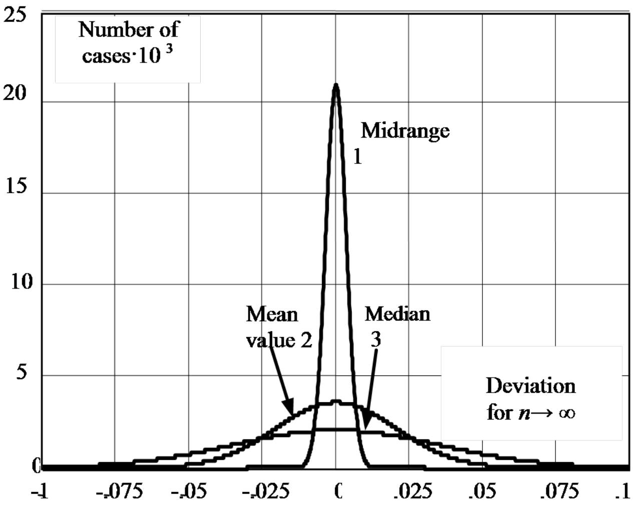


Figure 1: Histograms of estimators of measurand value for samples from population of rectangular distribution simulated by 200 × 2^{20} random numbers: 1 – midrange; 2 – mean value; 3 – median

Table 1: Statistical parameters of the sample of uniform pdf

Range of the sample	$V = x_{i\,\max} - x_{i\,\min}$
Midrange	x_V
Standard deviation of midrange x_V	s_V

For observations from a population with uniform distribution the best estimate of the measured value is mid range of the sample x_V:

$$x_V = \frac{x_{i\,\max} + x_{i\,\min}}{2}, \tag{3}$$

and the estimate of the standard deviation is its empirical deviation s_V:

$$s_V = \frac{V}{\sqrt{2}} \cdot \sqrt{\frac{N+1}{(N-1)^2\,(N+2)}} = \frac{V}{\sqrt{2}\cdot\nu} \cdot \sqrt{\frac{\nu+2}{\nu+3}}. \tag{4}$$

The paper presents the results of the properties of these estimators in function of degrees of freedom of sample $\nu = N - 1$, the simulation were carried out using the MC method. Observations were simulated with pseudo-random numbers from a population with a standard deviation
$\sigma = 1$ for the number $M = 2 \times 10^5$ of simulations and the numbers of degrees of freedom $\nu = 1, 2, 3, 4, 5, 7, 10, 16, 32, 63, 125, 250, 500, 1000$. The tests were carried out for the case where the population of random numbers (from which the observations come from) has clear and uniform distribution and for several cases where this population has uniform distribution contaminated by normal distribution.

2 Samples of Observations from a Population of Uniform Distribution

In general case estimator of the mean value x_{av} (1) in classic approach and estimator of midrange x_V (3) as special-estimator and their standard deviations respectively s_{Cl} and s_V have different values. A comparison of deviations from formulas (2) and (4) is shown in Figure 2.

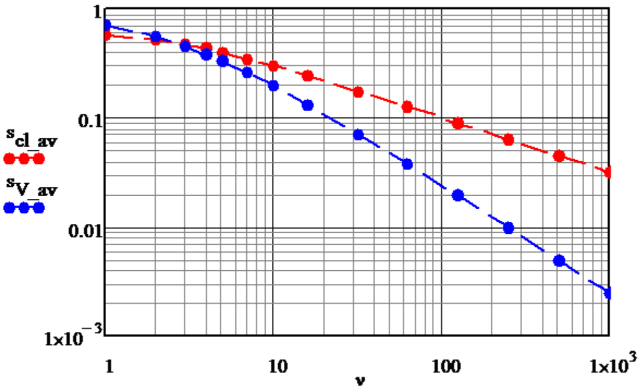


Figure 2: Standard deviations of mean and of midrange as functions of number ν of degree of freedom

On the basis of $N-1=\nu$ pure samples of uniform distribution with a standard deviation $\sigma=1$, in each simulation number $m(m=1,...M)$, for each value ν, the values s_{cl_m} and s_{V_m} were calculated. Then, for each value of ν from M simulations average values s_{cl_av} and s_{V_av} were calculated. Figure 2 suggests a superiority of special estimators (3), (4) over the conventional estimators (1), (2). For $\nu=10^3$ a value of s_{V_av} is approximately 13 times less than the s_{cl_av}.

Comparing the estimates only by empirical deviations is not fully reliable from metrological point of view, because the expanded uncertainties are important. Only the expanded uncertainties U_A associated with the random scatter of observations were taken into consideration here. With the classical approach the expanded uncertainty U_{Acl_m} is the product of deviations of the empirical s_{cl_m} and expansion coefficient k_{cl} calculated by Student's t-distribution for $\nu=N-1$ degrees of freedom and confidence level P:

$$U_{Acl\_m}=k_{cl}\cdot s_{cl\_m}. \tag{5}$$

In a special approach one can express the expanded uncertainty of type A:

$$U_{AV\_m}=k_V\cdot s_{V\_m}, \tag{6}$$

where: k_V is the coverage factor, specially adapted to estimators (3) and (4).

Classic Student's t-variable is defined as:

$$t\overset{def}{=}\frac{x_{av}-E\left(x\right)}{s_{cl}}=\frac{\Delta x_{av}}{s_{cl}}, \tag{7}$$

where $E(x)$ is the measured (expected) value, known in simulation experiments, and Δx_{av} - error of estimate x_{av}.

Similarly, the variable t_V of quasi-Student distribution is defined as

$$t_V \overset{def}{=} \frac{x_V - E(x)}{s_V} = \frac{\Delta x_V}{s_V}. \tag{8}$$

This is equivalent to the Student's t-variable for a population of observations of the uniform distribution with estimators (3) and (4). The probability distribution of the variable t_v can be called quasi-Student distribution.

For calculations of the expanded uncertainty from the sample data Dorozhovets [8] gives the following formula

$$U_P(x_V) = k_V \cdot x_V = \left(\frac{1}{\sqrt[N-1]{1-P}} - 1 \right) x_V. \tag{9}$$

As we did not find any mathematical proof of above formula we decided to calculate the appropriate coverage factor k_V by MC simulation. The graphs of the coverage factors k_{cl} and k_V for a confidence level $p = 95\%$ are shown in Figure 3.

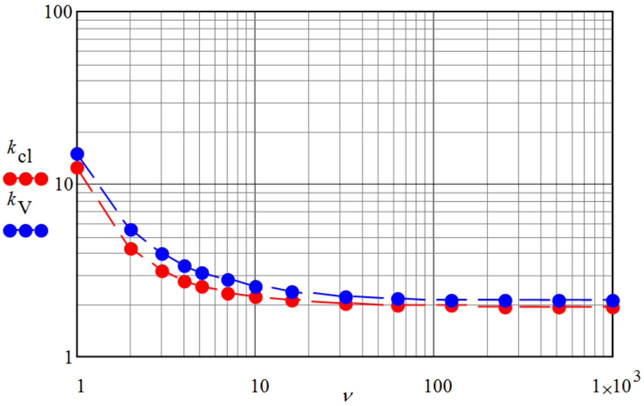


Figure 3: Coverage factors as function of number of ν degrees of freedom, $P = 0.95$

For following values ν_m the uncertainty $U_{Acl\_m}(5)$ and U_{AV_m} (6) M times have been calculated and also their average values U_{V_av} and U_{cl_av} in the data sets of size M are find. Results as function of ν are given in Figure 4.

These plots confirm the superiority of midrange estimators (3), (4) over the classical estimators of mean value (1), (2) for the numbers of degrees freedom ν greater than about 5. For $\nu = 10^3$ the value U_{AV_av} is about 11 times smaller than U_{Acl_av}.

It is desirable to verify the MC simulation results. It may be done by checking the empirical probability of an event which depend on verifying if estimate errors of the measured value (7), (8) are in the limits of the calculated uncertainty. For each of the $N = \nu + 1$ observations there is a need to calculate the number of successes and divide it by the number of M simulations. This quotient should have a value close to the postulated level of confidence $p = 95\%$. The results of this verification are in Figure 5.

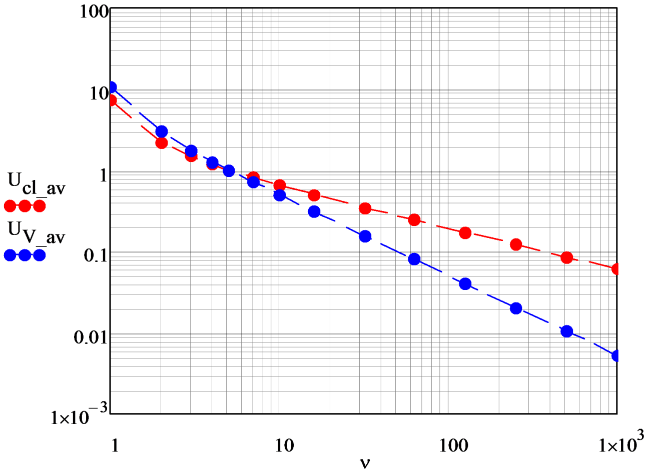


Figure 4: Average expanded uncertainties as function of number of ν degree of freedom

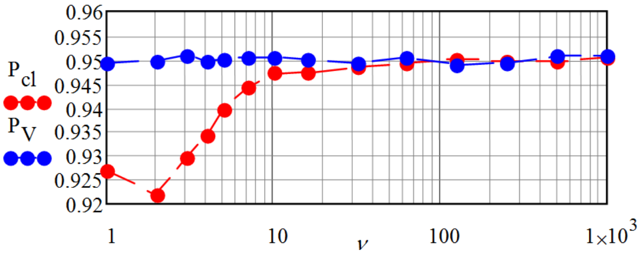


Figure 5: Empirical probabilities as a function of number of ν degree of freedom

They are fully satisfactory for a special approach - estimate (3) and (4). In contrast to the classical approach - estimate (1) and (2), they are unsatisfactory at small values of the number of observations $N = \nu + 1$.

3 Example

Let us find the best expected values of mean x_{av} and midrange x_V and their expanded uncertainties of the resistance R of the population of resistors with a nominal value $R_N = 100$ *ohm*. It can be assumed that the values of resistance R in the population have a uniform distribution based on the information that this population is the result of:

– selection in production from a population of resistors of the value of R of a continuous distribution of considerable width, or

– choosing the specially calibrated resistors taken from the population with a continuous, very wide distribution of values R, e.g. by *step by step* method [6].

To determine the expected value of resistance in the population and its uncertainty, the sample of size N is randomly collected and the resistance value of each of the downloaded items is measured. Table 2 shows the results of Monte Carlo simulations of such procedure.

Four variants are completed in the table: A - sample of size $N_A = 2(\nu_A = 1)$, B - sample of size $N_B = 4(\nu_B = 3)$, C - sample of size $N_C = 8(\nu_C = 7)$, D - sample of size $N = 17(\nu_D = 16)$. For each of the samples calculations of estimates of: the mesurand value (as expected value of resistance in the population), its standard deviation and the expanded uncertainty for a confidence level of $P = 95\%$ have been performed. Two methods of calculations are used: the classical method (estimate = average value) by formulas (1) and (2), and a special method (estimate = midrange) by formulas (3) and (4). It was also assumed that the uncertainty of the measuring equipment used for resistance is negligibly small.

Table 2: Example

No R_i	Sample						
	A	B	C	D			
1	99.925	100.006	100.089	1	100.015	No	R_i
2	100.057	99.929	100.016	2	100.025	10	99.971
3	–	100.046	99.951	3	99.910	11	100.049
4	–	99.906	99.970	4	99.952	12	100.048
5	–	–	100.059	5	100.079	13	99.940
6	–	–	99.914	6	100.046	14	100.036
7	–	–	100.018	7	99.979	15	99.974
8	–	–	99.940	8	100.081	16	99.922
9	–	–	–	9	99.978	17	99.941
x_{av}	99.991	99.972	99.995	99.997			
x_V	99.991	99.976	100.001	99.995			
s_{cl}	0.0660	0.0327	0.0216	0.0132			
s_V	0.0808	0.0301	0.0168	0.00736			
k_{cl}	12.71	3.18	2.36	2.12			
k_V	15.31	3.98	2.79	2.41			
U_{cl}	0.84	0.10	0.051	0.028			
U_V	1.2	0.12	0.047	0.018			

Results of calculations are presented in the lower rows of Table 2. Favorable results are given in bold numerals. As expected, the average value x_{av} has a lower value of the expanded uncertainty only for a very small sample sizes: $N_A = 2$ and $N_B = 4$ – variants A and B. In other cases, the midrange x_V is better as has the less uncertainty.

4 Statistical Properties of Samples from Population of Flatten-Normal Pdf

Value of the midrange x_V of samples from a population with uniform distribution depends only on the two external, minimal and maximal observations. Then the value of midrange and its uncertainty is strongly influenced by data outliers. It can be eliminated, similarly as for Gaussian samples according the Grabbs criteria, or by calculation of x_V for several external pairs of observations and discard the outlier score.

In the measuring system can also occur samples from a population, which is a convolution of uniform distribution with another distribution. If second one is, or can be approximated by a normal distribution, the flatten-normal distribution is obtain. The calculation of the mean value and its uncertainty of the sample from the flatten -normal distribution by use a number of conventional methods is described in [14]. By MC method will be checked now whether a midrange of the samples from this distribution has properties similar to those of the uniform distribution.

The distribution of a flatten-normal population can be characterized by the degree of participation λ of normal distribution. This means that when the population standard deviation $\sigma = 1$, the standard deviation of the normal distribution component is $\sigma_N = \lambda$, and for the main component of the uniform distribution $\sigma_J = \sqrt{1 - \sigma_N^2}$. Figure 6 shows the plots of the flatten-normal distribution in four different levels of λ. Thus, for $\lambda = 5\%$ the component with uniform distribution is characterized by standard deviation of $\sigma_J \approx 99.87\%$. Plots of the empirical deviations and expanded uncertainties with contribution of the normal distribution $\lambda = 5\%$ are not differ significantly from charts for a uniform distribution in Figure 2 and Figure 4.

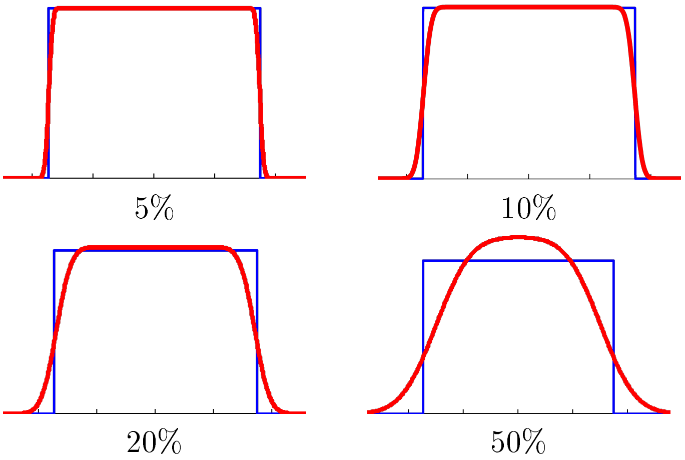


Figure 6: Flatten-Gaussian distributions of different λ and uniform pdf

In contrast, the empirical probability plots shown in Figure 7 differ significantly from those shown in Figure 4. Too small probabilities P_V for the larger numbers of observations $\nu = n + 1$ indicate the need to extend the coverage factors k_V for flatten-normal distribution.

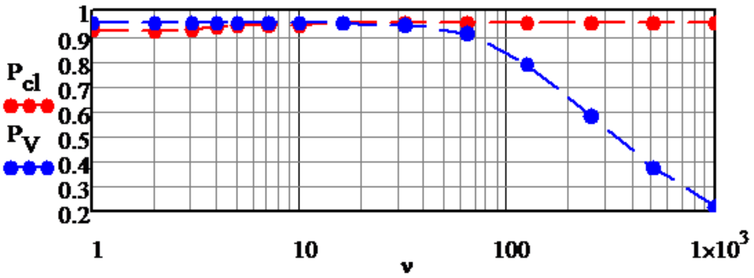


Figure 7: Empirical probability as function of number ν of degree of freedom

MC coefficients calculated by the factor k_V for a confidence level $p = 95\%$, and at the values of the degree of participation of the normal distribution $\lambda = 0\%, 5\%, 10\%, 20\%, 50\%$. The results of the calculations are presented in the Table 3 and in Figure 8.

Table 3: Coverage factor $k_V(p = 95\%)$ as function of number ν of degree of freedom for some values of λ

The degree of freedom ν	λ compactness of the normal distribution in %				
	$\lambda = 0\%$	$\lambda = 5\%$	$\lambda = 10\%$	$\lambda = 20\%$	$\lambda = 50\%$
1	15.31	15.54	15.01	14.40	11.94
2	5.51	5.47	5.42	5.31	4.66
3	3.98	3.95	3.96	3.91	3.67
4	3.42	3.41	3.41	3.40	3.34
5	3.09	3.09	3.10	3.12	3.23
7	2.79	2.79	2.81	2.88	3.19
10	2.57	2.59	2.62	2.74	3.36
16	2.41	2.42	2.52	2.83	3.96
32	2.24	2.39	2.71	3.55	5.63
63	2.18	2.61	3.48	5.13	8.68
125	2.14	3.45	5.23	8.20	14.23
250	2.13	5.31	8.60	13.93	24.33
500	2.14	8.79	14.79	24.30	42.57
1×10^3	2.13	15.39	26.44	43.61	76.29

With the increase of λ as degree of participation of the normal distribution in the flatten-normal population the efficiency of special approach (for the mid range) compared to classical approach is decreasing. This is illustrated on Figure 9 by graphs of the average expanded uncertainty and of the likelihood of verifying its legitimacy after-calculation.

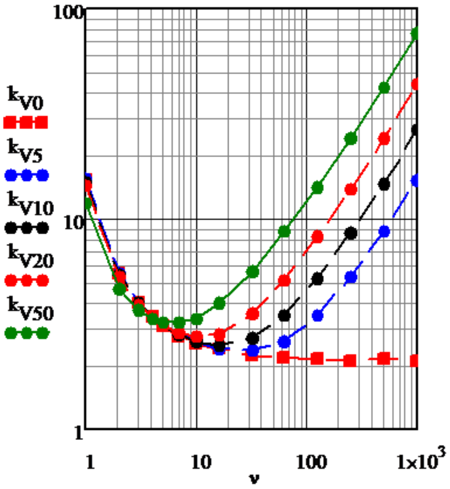


Figure 8: Coverage factors ($p = 0.95$) of Flatten-Gaussian pdf of normal pdf level $\lambda = 0\%, 5\%, 10\%, 20\%, 30\%, 50\%$ as function of number ν of degree of freedom

Remarks and Conclusions

The results of the MC simulation calculations are achieved with errors inversely proportional to the square root of M. In particular, the error of probability calculation is binomial (Bernoulli) with a standard deviation:

$$\sigma_P = \sqrt{\frac{P \cdot (1-P)}{M}}. \tag{10}$$

For $P = 0.95$ and $M = 2 \cdot 10^5$ one can receive $\sigma_P \approx 5 \cdot 10^{-4}$. For large values of M the error of probability calculation approaches the normal distribution and error limit can be estimated with the range 3σ as approximately 0.15%. This validates irregularities of plots in Figure 5 and Figure 9.

For the uniform distribution the use of special approach (3) and (4) is effective when the number of degrees of freedom ν is greater than 5 - see Figure 4. Then the average expanded uncertainty U_{V_av}, calculated according to a special approach using a special coverage factor, is less than the average uncertainty U_{cl_av} calculated classically by the GUM recommendations. U_{V_av} is less if the greater number ν.

For the observations from a population with convoluted uniform distribution even with a low content of another distribution, for example, $\lambda = 5\%$ of the normal distribution, there is a need to increase the coverage factor (Figure 8). It was assumed that this additional component has a normal distribution. Increasing the degree of participation of the normal distribution λ approach reduces the effectiveness of the special approach - see Figure 9 a,b,c. For example, for $\lambda = 20\%$ (Figure 9a) the effectiveness is only for the number of degrees of freedom ν from about 5 to about 100, and for $\lambda = 50\%$ (Figure 9c). The special approach is inferior to the classic approach in the whole range of numbers of degrees of freedom. The degree of participation of 50% does not mean that additional component of the standard deviation is 50%

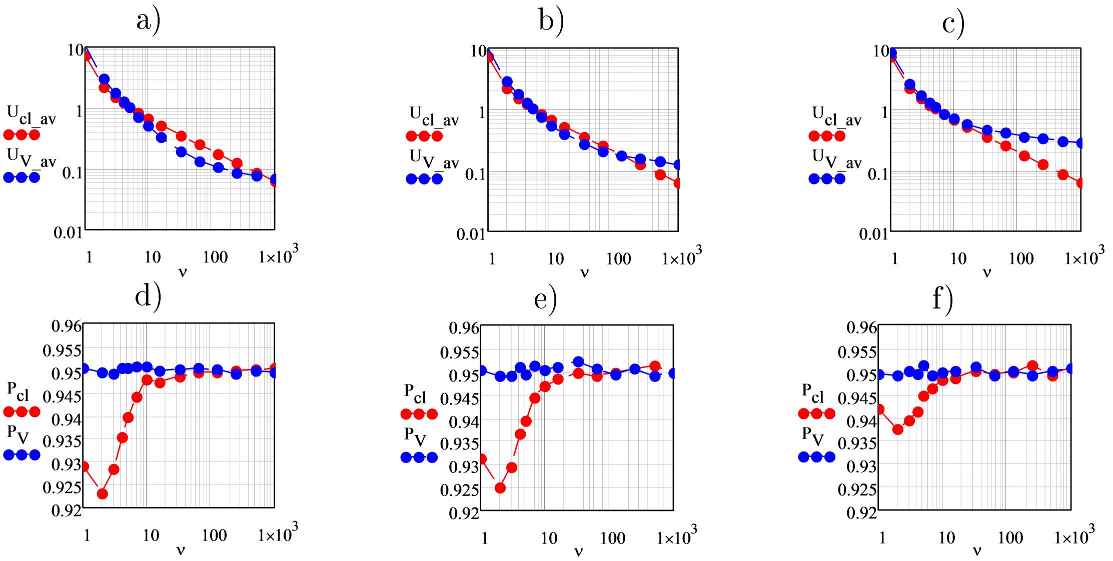


Figure 9: Uncertainties U_{clav}, U_{Vav} – a, b, c and control probabilities P_{cl} , P_V – d, e, f for levels of standard deviation of normal distribution $\lambda = 10\%, 20\%, 50\%$

of the population standard deviation of observation. The standard deviation of the main component is then $\sqrt{1 - 0,5^2} \approx 0,87 = 87\%$ of the standard deviation.

The classical approach is more efficient if there is a significant decrease of the uniform distribution in the plane-normal distribution (1), (2). However, the small numbers of degrees of freedom $\nu < 20$ gives a bit too low level of verifying probability P_{cl} – Figure 9 d,e,f.

In addition, it is worth mentioning that the other simple distributions also have the single component estimators better than the mean value. For U distribution (arc sin) mid range also is better, for the Laplace distribution (two-exponential) the best is median [7] - [9].

By MC method examined are also families of trapezoidal distributions, linear one - Trap as a convolution of two different uniform distributions and of concave shape CTrap [10] -[13]. For the ratio β of two bases of the trapezium in the range of 1 - 0.6315 the midrange is a better estimator than the mean value as it has a smaller standard deviation. For the linear Trap and concave CTrap trapezoidal distributions two-component estimator: 0.5 $\cdot$ (midrange + mean) is proposed [10] - [13]. It is more effective than any single-element estimator almost in the full range (0; 1) of the ratio β of trapezium basis.

References

[1] Guide to the Expression of Uncertainty in Measurement (GUM). ISO/IEC/OIML/BIPM, first edition, 1992, last ed. BIPM JCGM 100 (2008).

[2] Guide to the Expression of Uncertainty in Measurement (GUM), OIML ed. 2008 Supplement Propagation of distributions using a Monte Carlo method, G-101, (2007).

[3] H. Cramer, “Mathematical Methods of Statistics”, Stockholm Univ. (1946) Chapter 19.1.

[4] S. Trzetrzewiński ed., Drewnowski et all: “Pomiary Elektryczne” (Electrical Measurements), Chapter 2 of part I PWN Warszawa (1959) in Polish.

[5] S. Kubisa, S.Moskowicz, “A study on transitivity of Monte Carlo based evaluation of the confidence interval for a measurement result”. PAK (Pomiary Automatyka Kontrola) no 6 (2007) pp. 3-7 (in Polish).

[6] Kubisa S., Error distribution of a set of measuring instrument and an influence of “step by step” calibration procedure on the distribution. Metrologia i Systemy Pomiarowe. Tom V no 4 (1998), PWN, Warszawa p. 291-302 (in Polish).

[7] M. Dorozhovets, Z.L.Warsza, “Upgreading calculating methods of the uncertainty in measurements”. Przegląd Elektrotechniczny – Electrical Review no 1, (2007), pp. 1-13 (in Polish).

[8] M. Dorozhovets, “Opracovania rezultatov vimirovan” (Calculation of measurement results) Publisher: National University of Ukraina – Lviv Politechnic, Chapter 7.10.2, (2007), (in Ukrainian).

[9] Z. L. Warsza, M. Dorozhovets, “Type A uncertainty evaluation of autocorrelated observations and choosing the best estimators of data distribution”. Proceedings of 18th National Symposium Metrology and Metrology Assurance. Sept. (2008), Sozopol Bulgaria, pp. 70-78.

[10] Warsza Z. L., Galovska M., About the best measurand estimators of trapezoidal probability distributions. Przegląd Elektrotechniczny – Electrical Review 5 (2009), p. 86-91.

[11] Warsza Z. L., Galovska M., The best measurand estimators of trapezoidal PDF. Proceedings of IMEKO World Congress “Fundamental and Applied Metrology” (2009), Lisbon, CD pp. 2405-10.

[12] Galovska M., Warsza Z. L., The ways of effective estimation of measurand. PAKgoś (Pomiary Automatyka Komputery w gospodarce i ochronie środowiska) no 1 (2010) pp. 33-41.

[13] Warsza Z. L., Effective Measurand Estimators for Samples of Trapezoidal PDFs. JAMRIS (Journal of Automation, Mobile Robotics and Intelligent Systems) vol. 6, no 1, (2012) pp. 35-41.

[14] P. Fotowicz, Method of calculating the coverage interval based on Flatten-Gaussian distribution. Measurement, vol 55(2014) p. 272-275.

Evaluation of the Precision of Interlaboratory Measurements by Robust Algorithm S

Eugenij T. Volodarsky[1], Zygmunt L. Warsza[2], Larysa A. Kosheva[3], Adam Idzkowski[4]

[1] *National Technical University "KPI", Dep. of Experimental Studies Automation, Kiev, Ukraine*

[2] *Industrial Research Institute of Automation and Measurement (PIAP), Warszawa, Poland*

[3] *National Aviation University, Dep. of Biocybernetics and Aerospace Medicine, Kiev, Ukraine*

[4] *Bialystok University of Technology, Faculty of Electrical Engineering, Bialystok, Poland*

e-mail: vet-1@ukr.net, zlw1936@gmail.com, l.kosh@ukr.net, a.idzkowski@pb.edu.pl

Abstract

The application of robust statistical methods to assess the precision (uncertainty) of the results of interlaboratory comparison test with outliers is presented. An usual rejection of outlier data reduces the reliability of evaluation, especially for small samples. And the robust statistical methods take into consideration all data of sample (also outliers). In this paper the use of robust method "Algorithm S" for evaluation of the precision of interlaboratory measurements is presented and discussed in detail on the numerical example.

Keywords: robust statistics, algorithm S, outlier, interlaboratory comparisons, precision, uncertainty.

Introduction

Increasing demands of users and intensifying competition caused by producers and globalization of the World market led to the need for a comprehensive study of product parameters. Simultaneously, the interests of customers and suppliers, often conflicting, appeared during such studies. The principles of mutual recognition of product quality evaluation results were developed (to assess their compliance with requirements).

Their use is not possible without a well-functioning independent research laboratories with a high professional level. Mutual recognition of test results can be based only on the basis of proven technical competence of such laboratories [1], that is obtained in the process of accreditation. A particularly important role is played by the accuracy and reliability of test results that allow comparability. To this end, the laboratory must achieve the appropriate test conditions and maintain them as immutable. It also must apply a comprehensive modern statistical methods in processing the measurement results. Even with a limited set of data, statistical approach contributes to a better understanding of the course and causes of variability factors affecting the accuracy and reliability of the results [2] and allow comparability and compatibility studies [3] which are carried out in different laboratories.

A solution is to carry out the same measurements of homogenous objects in several accredited laboratories and to calculate the mean precision of all results. This interlaboratory comparisons are, in fact, an experimental implementation of a physical model by specific test procedures in certain conditions. This model is created on the basis of the measurement results obtained in the laboratories of a similar essential level of competence, which are specialized in a particular type of testing.

1 Test procedure

Quality testing procedures and its implementation affect the quality of the results. When assessing the suitability of the procedure, it is verified the possibility of its use for the test items that may occur with factors affecting them. When applying the procedure, the obtained results are controlled including ones received on the basis of participation in joint interlaboratory experiments. Previously, to assess the parameters of the results it was enough to do an experiment only in one laboratory. The latest certification rules require transition to other forms such an assessment, in particular the implementation of the joint experiment in order to more objectively determine the accuracy.

Depending on the purpose of research, the relevant statistical model analysis of variance and various types of indicators of accuracy. It is assumed that each measurement y is the sum of three components (in the regulations on testing laboratories there are other names than in GUM [4]), i.e.:

$$y = M_{\overline{\overline{y}}} + B + e, \tag{1}$$

where:$M_{\overline{\overline{y}}} = \mu + \delta$ - mean value of the measurement results from all laboratories; δ - component of the correctness of the result, i.e. moving average value (bias) due to the imperfections of the test procedure; B - validation results component (under reproducibility conditions); e - random measurement error component (under repeatability conditions).

The organization of interlaboratory experiment is presented in Fig. 1.

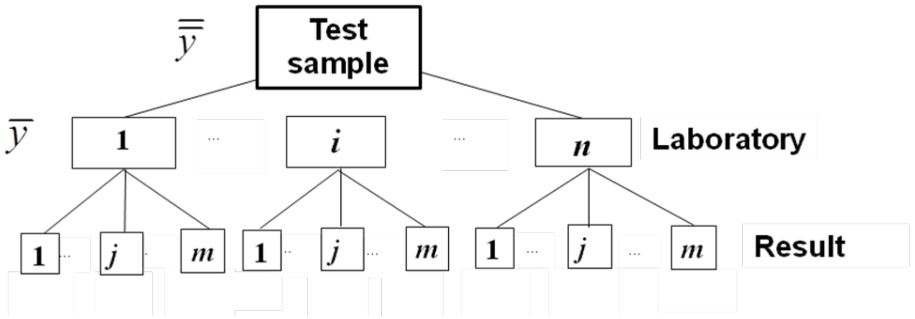


Figure 1: Organization diagram of interlaboratory experiment

Within-lab variance $s_{w_i}^2$ is separately calculated for the results in each laboratory

$$s_{w_i}^2 = \frac{1}{m-1}\sum_{j=1}^{m} (y_{ij}-\bar{y}_i)\ ^2. \tag{2}$$

The estimate of repeatability variance s_r^2 is a component which expresses repeatability of the scattering of results

$$s_r^2 = \frac{1}{n}\sum_{i=1}^{n} s_{wi}^2. \tag{3}$$

Component σ_L^2 is the estimate of the between-lab variance which describes a dispersion of the measurement results for the homogeneous objects in individual laboratories when the same measuring procedure is used

$$s_L^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\bar{y}_i-\bar{\bar{y}})\ ^2. \tag{4}$$

The estimate of reproducibility variance σ_R^2 represents the results of the interlaboratory test, conducted with the controlled measuring method, according to certified procedures. It is the sum of variances which defines repeatability, for $n \to \infty$

$$\sigma_R^2 = \sigma_r^2 + \sigma_L^2. \tag{5}$$

The relationships between these components are given in Fig. 2.

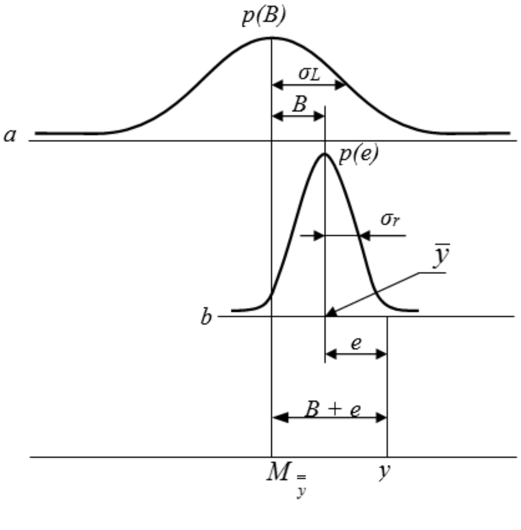


Figure 2: Basic statistical model of the measurement result

Usually, the statistical data analysis is based on the assumption that the scattering of data is normally distributed. It is also the basis for making a decision in statistical inference. Significant percentage of measurement results in practice can include data outliers. In particular this concerns the datasets with a small number of samples. The reason of outlier values in datasets are: failure of measuring instruments, non-compliance with the principles of an experiment, errors in the estimation of results, the impact of external factors. Rejecting these data in the calculations can significantly affect accuracy and reliability of the statistical evaluation of research precision.

The classic parametric evaluation of the experimental results based on normal distribution as well as the theory of statistical inference are firmly settled in practice. Cancellation of this approach would have been inadequate. Thus, a need of adaptation of the "old" model to the new challenges emerged. It can be realized by developing such methods of estimation which, under certain conditions, include "data outliers" or allow sufficiently to assess the parameters of results on the basis of acquired data. Several methods, named as robust, were developed by Tukey, Huber and others [5-7]. Some of them are applied in accredited laboratory practice and interlaboratory comparisons, [11], [14-part 5], [15].

As the example in the robust method proposed by Tukey [7], the basic model used is not a single normal distribution, but it is mixed from two of them. Tukey assumed that there are a large number n of measurement data, as accidentally mixed "good" and "bad" observation x_i from a population with a mean value μ, respectively, with probability (1-ε), where ε is a low number. Both types of observations x_i have different normal distributions, i.e. the first - N (μ, σ^2) and the second - N (μ, $9\sigma^2$), but with the same mean value μ – Fig. 3.

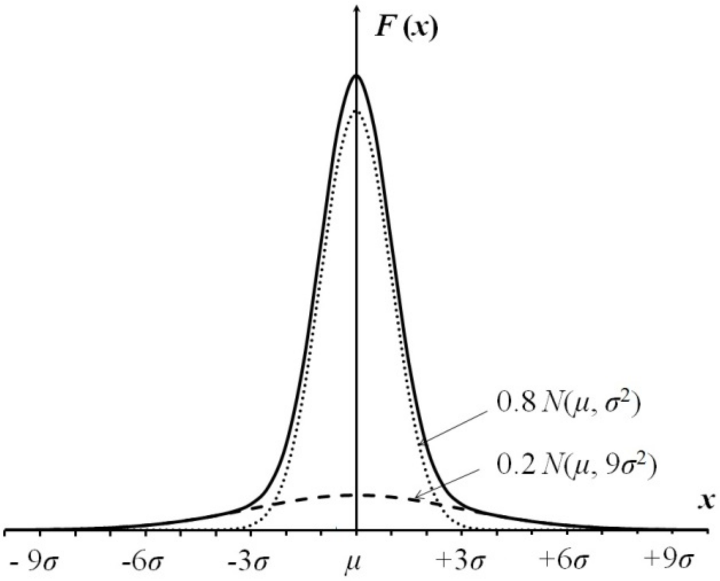


Figure 3: Joint distribution$F(x)$=(1- ε)N (μ, σ^2)+εN (μ, $9\sigma^2$) for ε= 0.2

The standard deviation of the "bad" is 3 times higher than "good". Assuming that all values x_i are independent, the following joint distribution can be expressed as

$$F\left(x\right)=\left(1{-}\varepsilon\right)\cdot\Phi\left(\frac{x{-}\mu}{\sigma}\right)+\varepsilon\cdot\Phi\left(\frac{x{-}\mu}{3\sigma}\right). \qquad (6)$$

Among robust methods and algorithms the approach of Huber is widely spread [5] and is also currently regarded as classical. Huber introduced k value which depended on the degree of "contamination" of the general population. It defines the boundaries of the central area of the measurement data histogram, i.e. difference between the upper and lower quartiles modeled by the normal distribution – Fig. 4 [7], [8], [12], [13]. Observations are less common in the lateral areas and in one of the criteria they can be considered as outliers. In the method IRLS (iteratively reweighted least squares) extreme observations are subject to winsorizing, i.e. pulling them on the borders of the central area. It follows a change in the mean value and standard deviation of the new set of observations, and constriction of the central area. Therefore customizing the extreme data should be repeated. This process is iterated until changes become negligible.

The application of this robust method (to assess: the result obtained with a measurement method, proficiency testing for laboratory using small samples of data and the occurrence of outliers) was presented in [12], [13]. The difference between the average values designated in the interlaboratory study is utilized to assess the reproducibility of the result. The basis of applied robust algorithms in these works is high stability of interquartile range (Fig. 4) with the "pollution" reaching up to 50

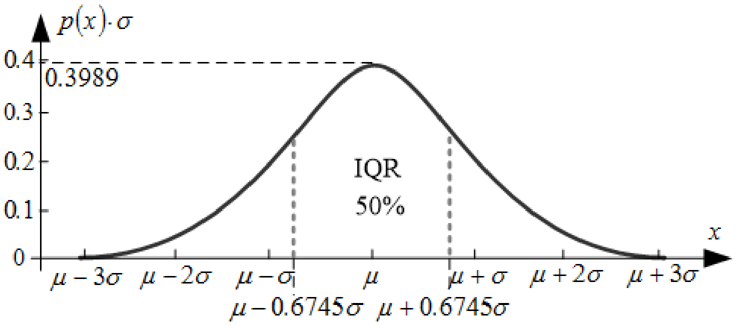


Figure 4: The inter-quartile range (IQR) and probability density function of a normal distribution N (μ,σ^2).

Some other robust methods are also applied in accredited laboratory practice and in interlaboratory comparisons and in quality assessment [5-12]. One of them - Algorithm S, recommended by ISO [14], [15] for the estimation of precision of the common result in interlaboratory measurements, is analyzed below in detail.

2 The robust method "Algorithm-S"

The implementation condition of this algorithm is that the bias estimate of robust standard deviation of results from laboratories should be equal to zero. For real experimental data at each j-th step of iteration, this assessment is closer to the standard deviation σ of the normal distribution. Adjustment factor ξ is introduced to estimate a variance shift. The condition should be provided

$$E\left\{\ (\xi\, s^*)^2\right\} = \sigma^2, \tag{7}$$

where:σ - standard deviation of "pure" normal distribution population, ξ - adjustment factor.

Robust standard deviation s^* should be stable with some probability (1-α) - Fig. 5.

$$P\left\{s^* > \eta\sigma\right\} = \alpha, \tag{8}$$

where:η – limit factor, $\eta\ \sigma$ - upper $\alpha\%$ point of distribution s^*.

The values of adjustment factor ξ and limit factor η are usually determined for $\alpha = 0.1$. It is made by intersecting of cumulative curves of one-modal distributions near the point where the probability equals 0.9. This approach should be examined analytically and its effectiveness should be assessed. Factor η corresponds to the upper value (1-α) 100% of distribution describing the scattering of robust standard deviation s^*. Standard deviation of this distribution may be used to assess the scattering.

A value $\chi^2_{\nu,\,=1-\alpha}$ can be found from Pearson distribution tables [8], [9] and then limit factor η for which the condition (8) occurs

$$\eta^2 = \frac{\chi^2_{\nu,\,P=0{,}1}}{\nu}. \tag{9}$$

Starting from the relation $P\left(\chi^2_\omega \leq \nu\cdot\eta^2\right) = 1-\alpha$, for the main part of the distribution z value corresponding to the value of probability P can be found from the tables. Adjustment factor ξ for the selected limit factor η, which assures that robust estimate will not be shifted

$$\xi = \frac{1}{\sqrt{z+0.1\eta^2}}. \tag{10}$$

Robust standard deviation s^*_j is calculated for the j-th step of iteration. In the iterative calculation the value s^*_j is updated as follows

$$\psi_j = \eta\ s^*_j, \tag{11}$$

In the ordered series of variances of results from laboratories participating in the experiment, a median is selected as an initial assessment of the standard deviation of the predicted normal population

$$s_0^{*2} = Me(s_i^{*2}), \tag{12}$$

where $i = 1\ ..\ n$ - number in an ordered series of laboratories.

Then the laboratory standard deviations are changed according to formula

$$s^*_{ij} = \begin{cases} \psi_j \text{ when } s_i > \psi_j \\ s_i \text{ –in other cases} \end{cases} \quad j = 0,\ 1,\ ... \tag{13}$$

On the basis of the value ψ_j which is found in the current step, the values of deviations s^*_{ij} in the dataset are modified and the new values are calculated from

$$s^*_{j+1} = \xi \sqrt{\sum_{i=1}^{n} \frac{(s^*_{ij})^2}{n}}, \tag{14}$$

where - s^*_{ij} robust standard deviation in the j-th step of iteration, for the i-th from n laboratories participating in the experiment.

Robust estimate s^*_{j+1} is used to establish a new limit ψ_{j+1}. Iterative procedure is continued until all standard deviations of the laboratories involved in this experiment converge within the ranges of current limit. Within-lab variances (3) $s^2_{w_1,...} s^2_{w_i}, ...s^2_{w_n}$are used in Cochran's C test. The C test evaluates the ratio

$$G_p = \frac{s^2_{w_i \max}}{\sum_{i=1}^{n} s^2_{w_i}} \leq G_{kr} \left(\alpha, m, n\right). \tag{15}$$

The estimate of repeatability variance is $s_r^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} s^2_{w_i}$.

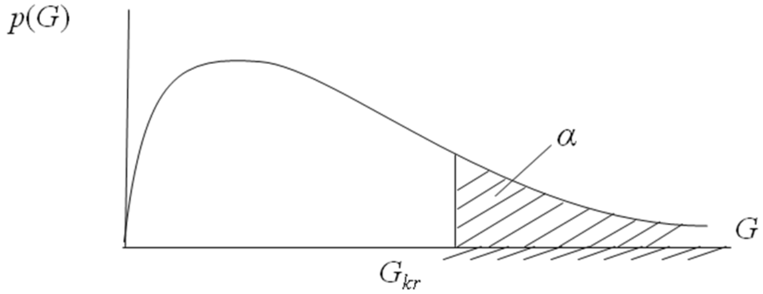


Figure 5: Tail probability computed from the F-distribution, G_{kr} - critical value, α - significance level, n - number of considered data series, m - number of data points per data series

Application of Cochran's criteria suggests that the experimental data belongs to the general normal distribution. However, experience has shown that this assumption is not always fulfilled. In addition, a decrease in the sample size causes that the statistical reliability of the decision hypothesis is reduced. The type of distribution of ratio (10), as shown in Fig. 6, suggests that the violation of normally distributed sample should be considered. If it not considered, the checked hypothesis of equality of dispersions (uniformity of selective dispersions) can be rejected with a high probability at violation of normally distributed sample. The standard [11] suggests the use of Cochran's test for small sample sizes: it must be kept in mind that in such a situation it can be reliably to distinguish only far enough alternatives when the variances are significantly different.

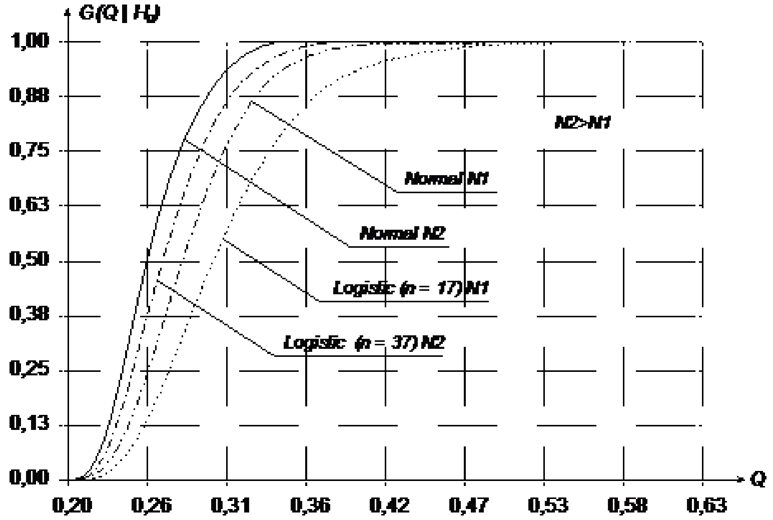


Figure 6: The distribution functions of (10) when the sample observations belong to different distributions: normal, logistic [17]

3 The application of algorithm-S to evaluate differences between laboratories

Nine laboratories with extensive experience in this type of research were selected for the experiment. Standard deviation (range) values w_i for all laboratories are as follows:

$w_1 = 0.28$; $w_2 = 0.49$; $w_3 = 0.40$; $w_4 = 0.00$; $w_5 = 0.35$; w_6=1.98; $w_7 = 0.80$; $w_8 = 0.32$; $w_9 = 0.95$.

The variance of the difference of m=2 results from the i-th laboratory is $s_i^2 = \frac{w_i^2}{2}$.

The hypothesis of a statistical outlier in a 6^{th} laboratory (value $w_6 = 1.98$) is tested using the Cochran's C test [8], [16]

$$G_p = \frac{1.98^2}{6.1663} = 0.636, G_{kr}\,(5\%) = 0.638, G_{kr}\,(10\%) = 0.754.$$

$G_p(w_6) < G_{kr}\,(5\%)$ and w_5 can be considered as outlier.

For $m = 2$ adjustment factor ξ and limit factor η are equal to

$$\xi = 1.097, \eta = 1.645.$$

Initial data $w_i$are ordered by valuesand put in column 0 of Table 1 as w_{i0}^*.

In the first step of iteration from (12) is: $\psi_1 = \eta\, w_{50}^* \approx 0.66$. Values of $(w_{10}^*, ...w_{60}^*) < \psi_1$. Three values $w_{70}^* > \psi_1$, $w_{80}^* > \psi_1$, $w_{90}^* > \psi_1$ need to be modified to value ψ_1, as it is in column 1 of Table 1.

The new value of standard deviation is $w_1^* = \xi \sqrt{\frac{1}{9} \sum_{i=1}^{9} \left(w_{i1}^*\right)^2} = 0.52$.

In the second step of iteration $\psi_2 = 1.645 \cdot w_1^* \approx 0.86$. New values are: $w_{72}{}^* = w_{70}{}^* = 0.80$, $w_{82}{}^* = w_{92}{}^* = \psi_2 = 0.86$ and $w_2^* = 0.56$.

In the third step: $\psi_3 = 1.645{\cdot}w_2^* = 1.00$, $w_{83}{}^*{=}w_{80}{}^*$and $w_{93}{}^* = 1.0$ as $w_{90}{}^*$modified to value ψ_3 and $w_3^* = \xi\sqrt{\frac{1}{9}\sum_{i=1}^{9}\left(w_{i3}^*\right)^2} = 0.60$.

Next values are: $\psi_4 = 1.645 \cdot w_3^* = 1.09$,$w_{94}^* = 1.09$, and computed is $w_4^* = 0.62$, which is higher than w_3^* only about 3%. Then finally another changes of w_j^* can be neglected.

Table 1: Example of calculations using algorithm S

Iteration j	0	1	2	3	4	5
ψ_j	-	0.66	0.86	1.00	1.09	1.12
$w_1{}^*$	0.00					
$w_2{}^*$	0.28					
$w_3{}^*$	0.32					
$w_4{}^*$	0.35					
$w_5{}^*$	0.40					
$w_6{}^*$	0.49					
$w_7{}^*$	0.80	0.66	**0.80**			
$w_8{}^*$	0.95	0.66	0.86	**0.95**		
$w_9{}^*$	1.98	0.66	0.86	1.00	1.09	**1.12**
St.Dev. $w_j{}^*$	0.83	0.47	0.56	0.60	**0.62**	

Final processed values of general standard deviations of interlaboratory experiment are:

1. for all initial values w_i w_0^* $(n = 9)$ $= 0.83$,
2. with rejection w_9 as outlier w^* $(n = 8) = 0.53$,
3. by robust method w_{rob}^* $(n = 9) = 0.68$.

Referring to Tab. 1 it is supposed that w_9^* can take values greater than 1.98. Studies have shown that in this case for α =0.1 the robust deviation w_{rob}^*would change its value only in the third decimal place.

Conclusions

The method of determining precision of a measurement method is briefly presented. If the full model is not known then tests are conducted on homogeneous objects by the same procedure in several laboratories with similar competencies. It can be assumed that the scattering is modeled by random variable with normal distribution. On the basis of the results of this research a statistical model is created and its accuracy is determined. In practice the outliers in results may occur. Rejection of them from further processing, when there is a small number of experi-mentally acquired data, diminishes the credibility of the assessment. Thus a robust statistical method should be applied.

The evaluation of precision of results, using the same method for homogeneous objects in nine laboratories, was presented. By traditional calculations, the estimate of the standard deviation was achieved 1.5 times higher without rejection of outlier in comparison to one with rejection of outlier. In turn, after using of robust method "Algorithm S" a value close to the lower of them was received, and with greater reliability.

References

[1] ISO/IEC 17025:2005 General requirements for the competence of testing and calibration laboratories

[2] ISO/TR 10017:2003 Guidance on statistical techniques for ISO 9001:2000

[3] ISO 10012:2004 - Measurement management systems - Requirements for measurement processes and measuring equipment

[4] Guide to the Expression of Uncertainty in Measurement.GUM. First ed. 1993 ISO Switzerland, last corrected ed. JCGM BIPM, 2008.

[5] Tukey J. W.: Exploratory Data Analysis, Addison-Wesley, 1978.

[6] Willinik R., What is robustness in data analysis. Metrologia 45, pp. 442-447, 2008.

[7] Huber P. J., Ronchetti E. M., Robust Statistics 2nd edition. Wiley, 2011.

[8] Farrant T.J.: Practical statistics for the analytical scientist: A bench guide, Royal Society of Chemistry, 1997

[9] Wilrich P.T.: The determination of precision of qualitative measurement methods by interlaboratory experiments, Accreditation and Quality Assurance, vol. 15, issue 8, pp. 439-444, 2010, doi: 10.1007/s00769-010-0661-1.

[10] Coucke W., China B., Delattre I., Lenga Y., Van Blerk M, Van Campenhout C., Van de Walle P., Vernelen K., Albert A.: Comparison of different approaches to evaluate External Quality Assessment Data, Clinica Chimica Acta, vol. 413, no 5-6, pp. 582, 2012, doi:10.1016/j.cca.2011.11.030.

[11] Rosario P., Martínez J.L., Silván J.M.: Comparison of different statistical methods for evaluation of proficiency test data, Accreditation and Quality Assurance, vol. 13, issue 9, pp. 493-499, 2008, doi: 10.1007/s00769-008-0413-7.

[12] Volodarsky E., Warsza Z.: Zastosowanie statystyki odpornościowej na przykładzie badań międzylaboratoryjnych (Applications of the robust statistic estimation on the example of inter-laboratory measurements), Przeglad Elektrotechniczny - Electrical Review 11, pp. 260 –267, 2013 (in Polish)

[13] Volodarsky E. T., Warsza Z. L.: Application of two robust methods on the example of inter-laboratory comparison, Monograph "Advanced Mathematical and Computational Tools in Metrology and Testing X" (Editors: F. Pavese et all), vol.10, Series on Advances in Mathematics for Applied Sciences vol. 86, World Scientific, New Jersey, London, Singapore, pp. 385 -391, 2015

[14] ISO 5725-2, -5: 2002 Accuracy (trueness and precision) of measurement methods and results – Part 2: basic method for the determination of repeatability and reproducibility of a standard measurement method - Part 5: Alternative methods for the determination of the precision of standard measurement methods, International Standardization Organization, Geneva, Switzerland.

[15] ISO 13528:2005 Statistical methods for use in proficiency testing by interlaboratory comparisons (IDT), attachment C2.

[16] Zieliński R., "Tablice statystyczne (Statistical Tables)" PWN Warszawa, (1972) (or tables in internet).

[17] Lemeshko B. Yu., Lemeshko S. B., Gorbunova A. A.: Application and power of criteria for testing the homogeneity of variances. Part I. Parametric criteria, Measurement Techniques, vol. 53, issue 3, pp. 237-246, 2010, doi: 10.1007/s11018-010-9489-7.

Investigation of Maximum Likelihood Estimates and Goodness-of-Fit Tests for Data with Known Measurement Error[1]

Stanislav S. Vozhov and Ekaterina V. Chimitova
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: chimitova@corp.nstu.ru

Abstract

In many practical situations, we only know the upper bound Δ on the measurement error. It means, that the precise measurement is located on the interval $(x-\Delta, x+\Delta)$. In other words, the data can be represented as a sample of interval observations. This paper is devoted to the problems of estimation of distribution parameters and testing goodness-of-fit with interval data. At first, we have compared the properties of maximum likelihood estimates (MLEs) with complete and interval data. Then, the modifications of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling statistics for testing goodness-of-fit have been proposed for interval data. The power of these tests has been investigated for complete and interval data.

Keywords: interval data, maximum likelihood estimates, Monte Carlo method, Kolmogorov test, Cramer-von Mises-Smirnov test, Anderson-Darling test.

Introduction

The development of statistical methods for the analysis of interval data is a promising area of research in the field of applied mathematical statistics. The nature of interval data is various. For example, obtained observations can be considered as intervals of fixed length due to the measurement errors. In marketing research, observations obtained from the survey of a target group of consumers are usually interval. In reliability and survival analysis, lifetime data are often interval-censored.

The basis of interval data analysis was initially laid by the measurement theory in metrology, where an interval uncertainty is introduced naturally. It is expected, that every observation is a value measured by an instrument with absolute error Δ. Thus, if the precise value of an observed response is $\dot{x}$, measurement error is $e \in [-\Delta, \Delta]$, then the measurement is equal to $x = \dot{x} + e$. In this case, we deal with a usual complete sample $\mathbb{X}_n = \{X_1, ..., X_n\}$. On the other hand, the measurement can be represented as an interval $(x - \Delta, x + \Delta) = (L, R)$. In this case, for the sample of observations we obtain an interval sample of the form

$$\mathbb{I}_n = \{(L_1, R_1), ..., (L_n, R_n)\}.$$

[1] This research has been supported by the Russian Ministry of Education and Science (project 2.541.2014K).

Interval observations are considered in many publications, see for example [4] - [10], [12]. In these papers, the reasonability of constructing new mathematical and statistical models, according to which observations are not numbers but intervals, was shown.

On the one hand, the transformation of complete observations (measurements) to intervals is associated with the loss of information and, as a result, with the decrease of accuracy of distribution parameters estimation. On the other hand, when the distribution of measurement errors $F_e(t)$ is unknown, the properties of parameter estimates by interval data, in theory, should not be sensitive to this distribution. This should distinguish this case from complete data, for which the estimates of distribution parameters can be biased in the case of asymmetric distribution $F_e(t)$.

In this paper, we investigate the statistical properties of MLEs of distribution parameters for interval samples. New goodness-of-fit tests for simple and composite hypotheses for interval data are proposed. The investigation of statistics distributions and the power of proposed tests are carried out using statistical simulations. All results for interval samples are compared with similar results for complete samples.

1 Maximum-likelihood method

Let us assume, that random variables $\dot{x}_1, ..., \dot{x}_n$ have the parametric distribution $F(t;\theta)$. The maximum likelihood method was chosen to estimate parameter θ. This method is based on maximization of the likelihood function

$$L(\mathbb{I}_n;\theta) = \prod_{i=1}^{n} \left(F(R_i;\theta) - F(L_i;\theta)\right).$$

Then, the MLE of distribution parameter is

$$\hat{\theta} = \arg \max_{\theta\in\Theta} \ln L(\mathbb{I}_n;\theta).$$

To compare the accuracy of the parameter estimates by complete and interval samples, we simulated complete and interval samples in accordance with the given distributions $F(t)$ and $F_e(t)$. Interval samples, in which each element is an interval of length 2Δ, were generated according to the following algorithm:

1. Generate an observation $\dot{x}$ from the distribution $F(t)$.

2. Generate a measurement error e from the distribution $F_e(t)$.

3. Obtain the interval observation:

$$(L, R) = (\dot{x} + e - \Delta, \dot{x} + e + \Delta).$$

4. Repeat 1-3 n times, thus forming an interval sample $(L_1, R_1), ..., (L_n, R_n)$.

Table 1: The properties of MLEs of distribution parameters for complete and interval data

		$F_e(t)$-Uniform		$F_e(t)$-Exponential	
	n	$\mathbb{X}_n$	$\mathbb{I}_n$	$\mathbb{X}_n$	$\mathbb{I}_n$
$M\hat{\theta}_1$	50	4.949	4.949	4.932	4.932
	100	4.949	4.949	4.934	4.934
	200	4.949	4.949	4.934	4.934
	300	4.951	4.951	4.934	4.934
	500	4.950	4.951	4.934	4.934
$M\hat{\theta}_2$	50	0.986	0.984	0.986	0.984
	100	0.993	0.991	0.993	0.992
	200	0.996	0.995	0.997	0.996
	300	0.998	0.996	0.996	0.996
	500	0.999	0.998	0.999	0.997
det Cov $\left(\hat{\theta}_1, \hat{\theta}_2\right)$	50	2.3E-04	2.3E-04	2.5E-04	2.5E-04
	100	6.4E-05	6.5E-05	7.4E-05	7.5E-05
	200	1.9E-05	1.9E-05	2.3E-05	2.38E-05
	300	9.7E-06	9.7E-06	1.3E-05	1.3E-05
	500	4.5E-06	4.5E-06	6.4E-06	6.5E-06

The normal distribution with the density function

$$f(t) = \frac{1}{\sqrt{2\pi}\theta_2} e^{\frac{(t-\theta_1)^2}{\theta_2^2}}$$

and parameters $\theta_1 = 5$, $\theta_2 = 1$ was considered as the distribution $F(t)$. As the distribution of measurement errors, we used the uniform distribution on the interval [-0.1, 0.1] and the right truncated at point 0.1 exponential distribution with the density function

$$f_e(t) = \begin{cases} \dfrac{1}{0.0333} e^{-\frac{t}{0.0333}}, t \geq -0.1, \\ 0, t < -0.1. \end{cases}$$

Unknown parameters of the normal distribution were estimated by maximum likelihood method. We simulated $N = 20000$ samples of size $n = 50, 100, 200, 300, 500$. Table 1 represents the mean values for MLEs of parameters θ_1 and θ_2 as well as the determinant of covariance matrix *det Cov* $\left(\hat{\theta}_1, \hat{\theta}_2\right)$ for samples $\mathbb{X}_n$ and $\mathbb{I}_n$.

As can be seen from Table 1, the properties of MLEs of normal distribution parameters by complete and interval samples are almost identical, but the mean values of estimates by complete samples are closer to the true parameter values than the corresponding estimates for interval samples. As it was expected, in the case of exponential distribution of measurement errors the MLEs of parameters by complete samples are biased. However, the assumption, that the MLEs of parameters by interval data are not sensitive to the distribution of measurement errors, was not confirmed.

2 Goodness-of-fit tests for interval data

The problem of testing a composite goodness-of-fit hypothesis

$$H_0 : F(t) \in \{F_0(t;\theta), \theta \in \Theta\}$$

with an interval sample can be solved similarly to the case of complete data. The difference is in the calculation of nonparametric estimate of the distribution function $F(t)$. In this case, the test statistic of the Kolmogorov type is calculated as follows:

$$D_n = \sup_{0<t<\tau_m} \left| \hat{F}_n(t) - F_0(t,\hat{\theta}) \right|,$$

the statistic of Cramer-von Mises-Smirnov type test is defined as:

$$S_{\omega^2} = \int\limits_0^{\tau_m} \left(\hat{F}_n(t) - F_0(t,\hat{\theta}) \right)^2 dF_0(t,\hat{\theta}),$$

and the statistic of Anderson-Darling type test has the form:

$$S_{\Omega^2} = \int\limits_0^{\tau_m} \left(\hat{F}_n(t) - F_0(t,\hat{\theta}) \right)^2 \frac{dF_0(t,\hat{\theta})}{F_0(t,\hat{\theta}) \left(1 - F_0(t,\hat{\theta}) \right)},$$

where $\hat{F}_n(t)$ is the nonparametric estimate of the distribution function by interval data, which is calculated using the ICM-algorithm [1, 2, 3, 11], $0 = \tau_0 < \tau_1 < ... < \tau_m$ is a sequence of points, which consists of all non-recurring ordered boundary points L_i and R_i, $i = \overline{1,n}$. The hypothesis H_0 is rejected for large values of these statistics. The analytical form of the distributions of considered statistics under the true null hypothesis is unknown. However, the significance level achieved (p-value) can be estimated using Monte-Carlo simulations. To investigate the power of the proposed goodness-of-fit tests for complete and interval samples, we consider a pair of close competing hypotheses:

H_0 : the normal distribution against

H_1 : the logistic distribution with the density function

$$f(t) = e^{-\frac{(t-\theta_1)}{\theta_2}} \Bigg/ \theta_2 \left(1 + e^{-\frac{(t-\theta_1)}{\theta_2}} \right)^2, \quad \theta_1 = 5, \theta_2 = \frac{\sqrt{3}}{\pi}.$$

The estimates of the power calculated for the significance level $\alpha = 0.1$ are given in Table 2 in the case of uniform distribution of measurement errors and in Table 3 in the case of exponential distribution of measurement errors. We simulated $N = 20000$ samples of size $n = 50, 100, 200, 300, 500$.

Table 2: The power of tests in the case of uniform distribution of measurement errors

	D_n		S_{ω^2}		S_{Ω^2}	
n	$\mathbb{X}_n$	$\mathbb{I}_n$	$\mathbb{X}_n$	$\mathbb{I}_n$	$\mathbb{X}_n$	$\mathbb{I}_n$
50	0.181	0.175	0.210	0.213	0.239	0.235
100	0.234	0.215	0.286	0.293	0.326	0.318
200	0.356	0.306	0.453	0.459	0.498	0.496
300	0.454	0.363	0.578	0.578	0.675	0.621
500	0.647	0.504	0.780	0.779	0.887	0.819

Table 3: The power of tests in the case of exponential distribution of measurement errors

	D_n		S_{ω^2}		S_{Ω^2}	
n	$\mathbb{X}_n$	$\mathbb{I}_n$	$\mathbb{X}_n$	$\mathbb{I}_n$	$\mathbb{X}_n$	$\mathbb{I}_n$
50	0.176	0.169	0.207	0.209	0.231	0.225
100	0.239	0.224	0.285	0.291	0.326	0.318
200	0.352	0.292	0.448	0.457	0.502	0.497
300	0.463	0.373	0.586	0.587	0.640	0.631
500	0.651	0.515	0.779	0.779	0.896	0.821

As can be seen from Tables 2 and 3, the power of all considered tests is higher for complete samples. The most powerful test among considered tests is the Anderson-Darling type test for both complete and interval samples. It is interesting to note, that the power of the Cramer-von Mises-Smirnov type test for interval data practically do not yield to the power of this test for complete data.

Conclusion

In the case, when the measurement error is known, the data obtained at the experiment can be considered as a sample of interval observations. In this paper, we have compared the properties of MLEs of distribution parameters for complete and interval data. It has been shown, that the precision of estimates by complete and interval data is almost the same.

In this paper, we have proposed the modifications of the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling type goodness-of-fit tests for interval data. The application of these tests in practice is based on the usage of Monte-Carlo simulations. The power of the proposed goodness-of-fit tests for interval data for all considered distributions of measurement errors is insignificantly less than the power of similar tests for complete data.

So, it is possible to conclude that under such problem definition, the consideration of data with measurement errors as an interval sample is not advisable.

References

[1] Groeneboom P. Asymptotics for interval censored observations // Technical Report 87-18. Department of Mathematics, University of Amsterdam, 1987. – 69p.

[2] Groeneboom P. Nonparametric maximum likelihood estimation for interval censored data // Technical Report, Statistics Department, Stanford University, 1991. – 87 p.

[3] Groeneboom P., Wellner J.A. Information Bounds and Nonparametric Maximum Likelihood Estimation. – Basel: Birkhauser Verlag, 1992. – 126 p.

[4] Kreinovich V. Interval computations and interval-related statistical techniques: estimating uncertainty of the results of data processing and indirect measurements // Advanced Mathematical and Computational Tools in Metrology and Testing X. - Singapore : World Scientific, 2015. - P. 38-49. - (Book series: Advances in Mathematics for Applied Sciences; vol. 86).

[5] Lemeshko B. Yu., Postovalov S.N. On estimation of distribution parameters by interval observations (in Russian) // Computing technology. 1998. Vol.3. - No.2. - P. 31-38.

[6] Lemeshko B. Yu., Postovalov S.N. On solving the problems of statistical analysis of interval data (in Russian) // Computing technology. - 1997. - Vol.2. - No.1. - P. 28-36.

[7] Lemeshko B. Yu., Postovalov S.N. Statistical analysis of interval observations (in Russian) // Science Bulletin of the Novosibirsk State Technical University. - 1996. - No.1. - P. 3-12.

[8] Orlov A.I. Basic ideas of interval data statistics (in Russian) // Science Bulletin KubSAU, no. 94(10), 2013. P. 1-26.

[9] Voshchinin A.P. A method for data analysis with interval errors in problems of hypothesis testing and parameter estimation of fussy linear parameterized functions (in Russian) // Industrial laboratory. 2000. Vol.66. no. 3. P.51 – 64.

[10] Voshchinin A.P. Interval data analysis: development and prospects (in Russian) // Industrial laboratory. 2002. Vol.68. no. 1. P. 118-126.

[11] Vozhov S.S. Investigation of the properties of nonparametric estimate for distribution function with interval data (in Russian) // Science Bulletin of the Novosibirsk State Technical University. – 2015. – no. 1(79). – P. 33-44.

[12] Zenkova Zh.N., Krakovetskaya I.V. Nonparametric Turnbull estimator for intervalcensored data in the marketing research of the demand of bio-energy drinks (in Russian) // Tomsk state university journal of control and computer science, 2013, no. 3 (24), pp. 64–69.

Investigation of L-distance between Probability Densities and their Estimates

Oleg A. Makhotkin
The Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia,
e-mail: `oam@osmf.sscc`

Abstract

In this paper, the usefulness of L-distance between the investigated probability densities and their sample approximates is demonstrated.

Keywords: L-distance, random sample, probability density estimation, Monte Carlo simulation.

1 Introduction

A well-known method to test the hypothesis that the elements of the random sample $\mathcal{S}_N = \{\eta_1, \ldots, \eta_N\}$ have the probability distribution with p.d.f.[1] $p(x),\ a < x < b$ needs the calculation of the criterion

$$\chi^2 = \sum_{i=1}^{m} \frac{(N \cdot P_i - m_i)^2}{N \cdot P_i}. \tag{1}$$

The sum in (1) can be represented in the other form. For this objective, on the grid $\mathcal{X}_m = \{x_0 = 0 < x_1 < \ldots < x_{m-1} < x_m = b\}$ the variables $h_i = x_i - x_{i-1}$ and the sets $\Delta_i = \{x_{i-1} < x < x_i\}$ are defined. The piecewise constant p.d.d.f.($\chi(x|Y) = 1$ if $x \in Y$, and $= 0$ otherwise)

$$\hat{p}_g(x) = \sum_{i=1}^{m} \frac{P_i}{h_i} \chi(x|\Delta_i), \tag{2}$$

where $P_i = \int_a^b p(x)\chi(x|\Delta_i)dx =< p(x)\chi(x|\Delta_i) >$ gives the Galerkin approximation of the p.d.d.f. $p(x)$. This approximation has the form $\hat{p}(x) = \sum_{i=1}^{m} c_i \chi(x|\Delta_i)$, where the coefficients $\{c_i\}$ are obtained from the equations

$$< \ [p(x) - \sum_j c_j \chi(x|\Delta_j)] \cdot \chi(x|\Delta_i) \ >= 0, \ i = 1, \ldots, m.$$

These equations have the solution $c_i = b_i/h_i = P_i/h_i$. The values of $b_i =< \chi(x|\Delta_i)p(x) >$ can also be estimated over a random sample $\mathcal{S}_N$ as $\bar{b}_i = \sum_{j=1}^{N} \chi(\eta_j|\Delta_i)/N$. This gives a random piecewise constant p.d.d.f.

$$p_{PC} = \sum_{i=1}^{m} \frac{m_i}{N h_i} \chi(x|\Delta_i) = \sum_{i=1}^{m} \bar{c}_i \chi(x|\Delta_i). \tag{3}$$

[1] p.d.d.f. - probability distribution density function(s),

Criterion (1) can be written as L_2-distance between densities (2) and (3) with the piecewise constant weight function $W_{PC}(x) = \sum_{i=1}^{m} W_i \chi(x|\Delta_i)$

$$L_2 = \int_a^b W_{PC}(x)(\hat{p}_g(x) - p_{PC}(x))^2 = \sum_{i=1}^{m} h_i W_i (\frac{P_i}{h_i} - \frac{m_i}{Nh_i})^2.$$

The choice of the weight values $W_i = Nh_i/P_i$ transforms L_2 to criterion (1).

2 L - distance between densities

There is another measure of the distance between the densities $p(x)$ and $\hat{p}_g(x) = \sum_{i=1}^{m} c_i \chi(x|\Delta_i)$

$$L_g = \int_a^b |\hat{p}(x) - p(x)| dx = \sum_{i=1}^{m} h_i \int_0^1 |c_i - p(x_{i-1} + h_i t)| dt = \sum_{i=1}^{m} h_i I_i.$$

The use of L-distance for estimation of the probability density was investigated in [1]. The value of L-distance can be calculated using the the compound Gauss quadrature. For its application every t-integral $I_i = \int_0^1 g(t)dt$ is represented as the sum of n subintegrals ($h_t = 1/n,\ t_{l-1} = h_t(l-1),\ t = t_{l-1} + h_t \cdot s$)

$$I_i = h_t \sum_{l=1}^{n} \int_0^1 g(x_{i-1} + h_i(t_{l-1} + h_t s))ds \approx \frac{h_t}{2} \sum_{l=1}^{n} \sum_{k=1}^{2} g(x_{i-1} + h_i(t_{l-1} + h_t X_k^{(2)})).$$

Here $X_{1,2}^{(2)} = (1 \mp \sqrt{3}/3)/2$ are the knots of the two-point Gauss quadrature.

The computer experiments have shown that L_g is very near to a minimum value $L_{min} = \min_d || \sum_{i=1}^{m} d_i \chi(x|\Delta_i) - p(x)||$. This means that a predominant part of random L-distances $L_{est} = ||p_{PC}(x) - p(x)||$ will be greater then a L_g. If standard methods for the generation of random variates with p.d.d.f. $p(x)$(see, for example, [3]) do not give an effective simulation algorithm, it is possible to use the approximation of $p(x)$ [4]. Suppose we have constructed the piecewise constant approximation of $p(x)$ on the uniform grid $\mathcal{X}_M$, $M > m$ with the error $L_{apr} \leq \delta \cdot L_g$, where $\delta \ll 1$. Then the use of this p.d.d.f. for the generation of the new test samples $\bar{\mathcal{S}}_N$ gives the values of the random distances $L_{est}(\bar{\mathcal{S}}_N) \leq (1+\delta) L_{est}(\mathcal{S}_N)$.

For the approximation of the parametric p.d.d.f. $p(x|\theta)$, $\theta = (\theta_1, \dots, \theta_K)$ the random estimates $\bar{\theta}(\mathcal{S}_N)$ are calculated over the initial sample $\mathcal{S}_N$. Then the approximation of the p.d.d.f. $p(x|\bar{\theta})$ by $\hat{p}_M(x|\bar{\theta}))^2$ is used for the generation of the new artificial samples $\mathcal{S}_N^{(1)}, \dots, \mathcal{S}_N^{(K)}$. Using K random distances $L_k = ||\hat{p}_M(x|\bar{\theta}(\mathcal{S}_N)) - p(x|\bar{\theta}(\mathcal{S}_N^{(k)}))||$ it appears possible to approximation p.d.d.f. $p(L)$ by the histogram with n intervals $p(L)$. Then for the given $\alpha(= 0.1, 0.2)$ the quantile L_α can be calculated. The approximation $\hat{p}_M$ is adopted if $L_{apr} = ||\hat{p}_M(x|\bar{\theta})) - p(x|\bar{\theta})|| \leq \delta \cdot L_\alpha$. Otherwise the

[2] For the piecewise constant or piecewise linear approximation M is the number of the grid intervals.

procedures repeated with increasing M until success has been achieved.

For the kernel estimation we need to simulate a random variate with the sample p.d.d.f. $\bar{p}(x) = (\sum_{j=1}^{N} K(x, \eta_j)/N$. It is possible to choose the kernel density with an effective simulation algorithm. For example, the triangle kernel $K(x, y) = 1 - (x - y)/h, \text{for} |x - y| \leq h, \text{otherwise} = 0$ has the simulation algorithm $\xi := y + h(U_1 - U_2)$, where U_1, U_2 are independent random variates uniformly distributed on $(0, 1)$.

3 Computer experiments

3.1 Non-parametric density estimation

1. The test p.d.d.f. is $p(x) = 6x(1 - x)$, $0 \leq x \leq 1$. The distribution function is $F(x) = x^2(3 - 2x)$.
The Galerkin approximation on the uniform grid with $m = 10$ has $L_g = 7.51e - 2$. The random variate $\eta \sim p(x)$[3] was generated by the inversion method: the equation $F(\eta) = U$ was solved by the Newton iteration method with the relative accuracy 1.0e-6. Hundred random samples $\mathcal{S}_{100}$ were generated. For every sample, the function $p_{PC}(x)$ was constructed on the uniform grid $\mathcal{X}_{10}$ and the distance $L = ||p_{PC}(x) - p(x)||$ was calculated. Then for the sample $L_1, \ldots, L_{100}$ the histogram with 10 intervals was constructed. It is presented in the Figure 1. The following extremum values were obtained for L-sample: $L_{min} = 1.23e - 1, L_{max} = 4.25e - 1$. The estimates of the quantiles are equal to $L_{0.1} = 0.17, L_{0.2} = 0.20$.

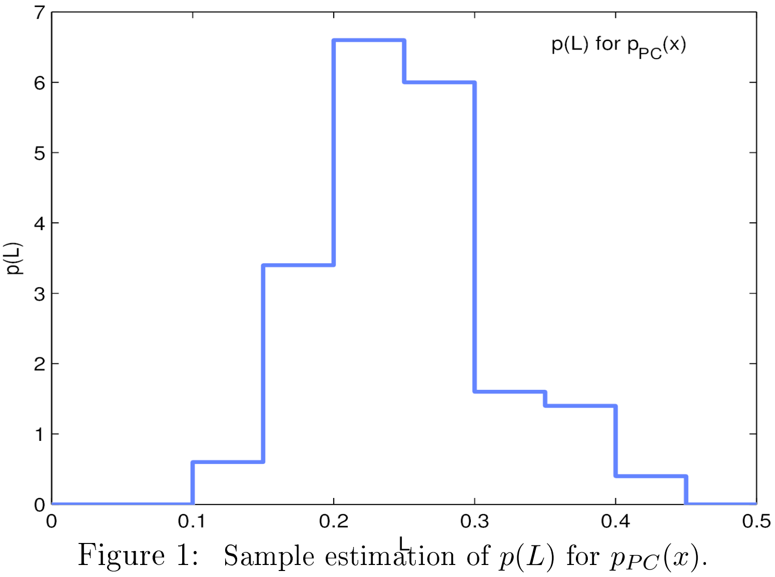


Figure 1: Sample estimation of $p(L)$ for $p_{PC}(x)$.

The same experiments were carried out for the piecewise linear approximations of the test p.d.d.f. which have the form $\hat{p}(x) = \sum_{j=0}^{m} c_j \phi_j(x)$. Here $\{\phi_j(x)\}_{j=0}^{m}$ are the basic functions of the piecewise linear approximation [4]. As for the piecewise

[3] The symbol "$\sim$" means "has p.d.d.f.".

constant case, the coefficients $\{c_j\}_{j=0}^m$ are defined from the orthogonality conditions: for $i = 0, \dots, m < [\sum_{j=0}^m c_j\phi_j(x) - p(x)] \cdot \phi_i(x) >= 0$. This gives the system of linear algebraic equations $A \cdot c = b$, where the matrix $A = \{a_{i,j} =< \phi_i(x)\phi_j(x) >\}_{i,j=0}^m$ and $b = \{b_i =< p(x)\phi_i(x) >$. If the source vector b is calculated by the quadrature formula, the solution of the system gives the non-stochastic Galerkin approximation $p_g^{(PL)}(x)$. It was calculated for $m = 10$ and L-distance $L_g^{(PL)} = 3.84e{-}3$ was obtained. The components of the source vector were also estimated over a random sample with $N = 1000$ as $\bar{b}_i = \frac{1}{N}\sum_{k=1}^N \phi_i(\eta_k)$, $\eta_k \sim p(x) = 6x(1-x)$. The L-distance between $p(x)$ and its random estimate $p_{PL}(x)$ equals $L_{est} = 3.15e - 2$.

Hundred realizations of the random L-distance were obtained by the simulation of samples $\mathcal{S}_{100}$ from the test distribution with p.d.d.f. $p(x) = 6x(1-x)$. Then the histogram with 10 intervals was calculated. It is shown in the Figure 2. The limiting values are $L_{min} = 0.12$, $L_{max} = 0.41$. The estimates of the quantiles are equal to $L_{0.1} = 0.16, L_{0.2} = 0.19$.

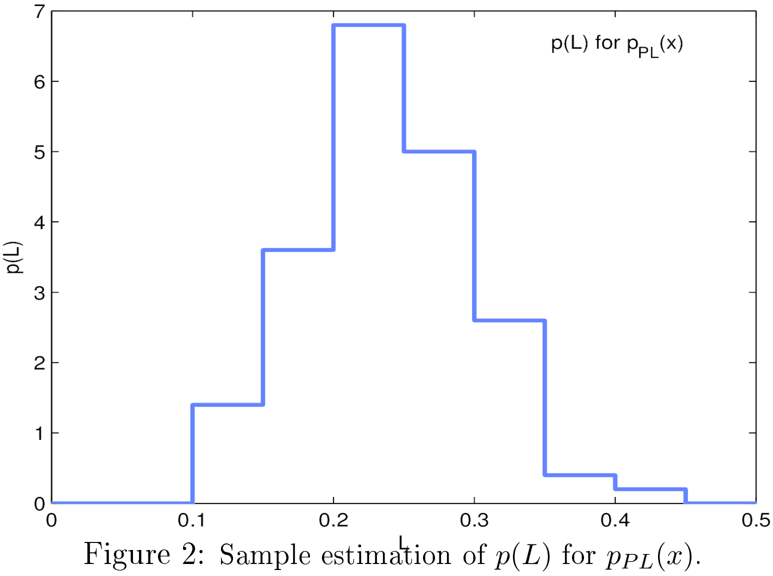


Figure 2: Sample estimation of $p(L)$ for $p_{PL}(x)$.

The test p.d.d.f. $p(x) = 6x(1-x)$ has the effective direct simulation algorithm. Suppose, that one has to use the approximation of the test probability density with the relative error $\delta = 1.0e - 2$. To this end, it is needed to use the approximations of the test p.d.d.f. with L-distance $L \leq L_{err} = \delta \cdot L_g$. The Galerkin piecewise approximation with 10 intervals has $L_g = 7.5e - 2$, and therefore $L_{err} = 7.5e - 4$. This error can be obtained for the piecewise constant approximation with $m = 1000$ intervals ($L_g = 7.5e - 4$) or for the piecewise linear approximation with $m = 40$ intervals ($L_g = 2.4e - 4$).

3.2 Parametric density estimation

In the case of the parametric density estimation the investigated p.d.f has the form $p = p(x|\theta_1, \dots, \theta_K)$. Using the sample $\mathcal{S}_N$ the parameter estimates $\bar{\theta}_1(\mathcal{S}_N), \dots, \bar{\theta}_K(\mathcal{S}_N)$

are obtained. The p.d.d.f. $p(x|\bar{\theta}_1,\dots,\bar{\theta}_K)$ is then used in the statistical simulation. The error of this approximation is $L_S = ||p(x|\bar{\theta}_1,\dots,\bar{\theta}_K) - p(x|\theta_1,\dots,\theta_K)||$. The fast algorithm for the simulation of the random variates with p.d.d.f. $p(x|\bar{\theta}_1,\dots,\bar{\theta}_K)$ can be obtained by its approximation (see [4]). If $\hat{p}(x)$ is the approximation of $p(x|\bar{\theta}_1,\dots,\bar{\theta}_K)$ with L-error $L_{apr} = ||\hat{p}(x) - p(x|\bar{\theta}_1,\dots,\bar{\theta}_K)|| \le \delta \cdot L_S$, then $L = ||\hat{p}(x) - p(x|\theta_1,\dots,\theta_K) \le (1+\delta)L_S$.

A well-known example gives the normal p.d.d.f.

$$\mathcal{N}(x|\mu,\sigma^2) = \exp(-(x-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}, \ -\infty < x < +\infty.$$

The sample estimates $\bar{x} = \sum_{j=1}^{N} \eta_j$ and $s^2 = \sum_{j=1}^{N} (\eta_j - \bar{x})^2/(N-1)$ are independent random variables. They are distributed as $\bar{x} \sim \mathcal{N}(x|\mu,\sigma^2/N)$ and $s^2 \sim \sigma^2\chi^2_{N-1}/(N-1)$. For the sample $\mathcal{S}_{40}$ of the normal random variates with $\mu = 1, \ \sigma = 2$ the following estimates were obtained: $\bar{x} = 0.931, s = 1.93$. The L-distance between densities is equal to $L_S(\bar{x}, s) = 4.20e-2$. Then 100 random values of L_S for $\mathcal{S}_{40}$ were generated, using the following simulation algorithm:

1. Obtain two standard normal variates by transformation $\xi_1 := \sqrt{-2\ln(U_1)}\cos(2\pi U_2), \quad \xi_2 := \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$,
2. Calculate $\bar{x} := \mu + \frac{\sigma}{\sqrt{N}}\xi_1, \ s := \frac{\sigma}{\sqrt{2(N-1)}}[\sqrt{2N-3} + \xi_2]$,
3. Calculate $L_S = ||\mathcal{N}(\bar{x}, s^2) - \mathcal{N}(\mu,\sigma^2)||$.

The histogram with ten intervals for the obtained sample of random L-values is shown in the Figure 3.

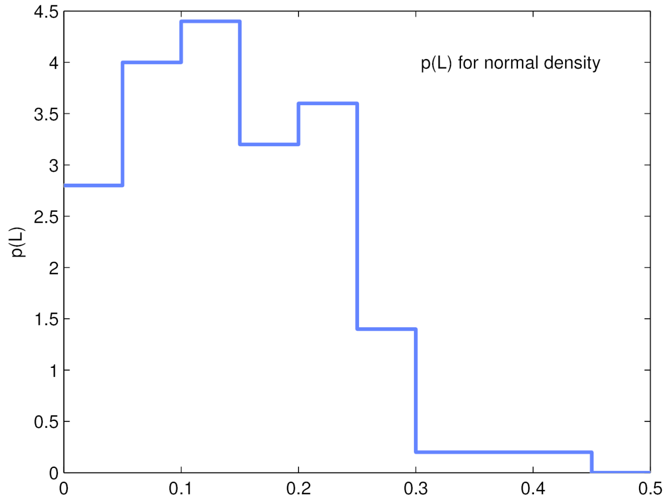


Figure 3: Sample estimation of $p(L)$ for $\mathcal{N}(x|\bar{x}, s^2)$.

The estimates of the quantiles are equal to $L_{0.1} = 7.5e-2, L_{0.2} = 1.3e-1$.

4 Summary

The main conclusions of the present paper can be summarized as follows:
(1) The L-distance can be used in the problems of the probability densities estimation over the random samples.
(2) The estimates of the L-distance can be used for creating the effective computer algorithms for the simulation of the random variates.

Acknowledgements

This work was supported in part by the RFBR project SS-5111.2014.1.

References

[1] Devroy L., Gyorfi L.,(1985). *Nonparametric Density Estimation. The L_1-View*, John Wiley and Sons.

[2] Devroy L.,(1986). *Non-Uniform Random Variate Generation*, Springer Verlag.

[3] Hammersley J.M., Handscomb D.C. (1964). *Monte Carlo methods*, Methuen, London.

[4] Makhotkin O.,(2013) Simulation of Random Variates by Approximation, Proceedings of AMSA'13, pp. 163-172, Novosibirsk, Russia.

Classification of Observation Sequences described by Hidden Markov Models

Tatyana A. Gultyaeva, Alexander A. Popov,
Valeriya V. Kokoreva and Vadim E. Uvarov
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: gultyaeva.ta@gmail.com

Abstract

This article covers several approaches to a problem of classification of multidimensional observation sequences described by Hidden Markov Models (HMM). It was shown that the method based on derivatives of likelihood function logarithm with respect to HMM parameters is effective when competing classes are similar in some way. This paper continues work of our previous articles, where we were doing research over one-dimensional observation sequences described by HMM [1], [2], [3], [4] and multidimensional observation sequences described by HMM [5], [6].

Keywords: Hidden Markov Model, classification of sequences, derivatives.

Introduction

HMM conception was developed in 60s – 70s of 20th century independently by several researchers (T.K. Vintsiuk [7], V.A. Kovalevsky [8] and L.E. Baum [9]). Sequences classification usually presents no difficulties when competing models are distinguishable enough (by probability). In that case, traditionally one would use a method based on likelihood criteria. However, classification results become spurious when competing models are similar, i.e. belonging of some sequence to any of competing HMMs becomes equally probable. Such situation usually occurs when observations are distorted which also makes them hard to distinct or when real world objects or processes are actually similar to each other in parameters and inner structure, which leads to similar observations, produced by them. Thus, it is necessary to use different approaches, which would improve HMM capabilities of distinguishing the similar alternatives. There are two approaches to that problem. First makes use of methods that work with actual model: either change structure of used models (see, e.g. works of V. Alexandrov [10], R. M. Neal [11]) or use other methods of model parameters estimation (C.J. Walder [12], S. Ikbal [13], G.D. Zhou [14]) or combines those two methods (C. Liu [15], S.P. Chatzis [16]). In other words, methods of first group are oriented on more accurate description of object or process of interest. Second approach allows changing the decision rule by getting some information from models and using it for classification. For example by transition into a space of secondary attributes, (see works of R. Solera-Urena [17], Ch. Ling [18], O. Aran [19]) followed by classification with the use of support vector machine. Second approach seems more perspective to us, because there is a particular freedom both in choosing of attribute space of classification and in choosing of actual classifier. This paper covers classification of multidimensional sequences described by HMM in space of derivatives with respect to HMM parameters. We shall note that earlier works by other authors who used such spaces were dedicated to classification of one-dimensional sequences only.

1 Hidden Markov Models

Hidden Markov process represents a mathematical model that is a two-component random process (X, Y) with hidden component Y and observable component X where random process X is Markov random process [20]. HMM is a particular case of hidden Markov process when X is Markov chain with finite set of states, which is described by a transition probabilities matrix. Current state of Markov chain q_t is interpreted as a hidden state of data source (object of interest) and moments of state changing – as discrete events that occur on development of object of interest. The sequence of hidden states modelled by such chain (i.e. realization of random process) is denoted as $Q = \{q_1, q_2, ..., q_T\}$, where T is the length of observation sequence. Random process Y is real-valued ($y_t \in \mathbb{R}$) random process of finite order (i.e. the density of transition into a new state probability depends not only on one but on several previous states) with discontinuous probability properties. Sequence $O = \{o_1, o_2, ..., o_T\}$ is a sequence of observation states (i.e. realization of random process). Random process Y by conditional probability $P\left(o_t \mid q_t = s_i\right)$. Thus, HMM can be fully described by the following parameters:

1) initial state distribution $\Pi = \{\pi_j\}$, $j = \overline{1, N}$, where $\pi_j = P\left(q_1 = o_t\right)$; set of hidden states $S = \{s_1, s_2, ..., s_N\}$, N – number of hidden states in model;

2) state transition probabilities matrix $A = \{a_{ij}\}$, $i, j = \overline{1, N}$, where $a_{ij} = P\left(q_t = s_j \mid q_{t-1} = s_i\right)$;

3) probability density function of observation symbols $B = \{b_i(t)\}$, where $b_i(t)$ are density functions of conditional probabilities $P\left(o_t \mid q_t = s_i\right)$, o_t – element from observation sequence that was observed at time $t = \overline{1, T}$.

This article covers case when conditional probability densities of observable symbols are mixtures of probability distributions:

$$b_i(t) = \sum_{m=1}^{M_i} \tau_{im} g(o_t; \Theta_{im}), \tag{1}$$

where τ_{im} is a weight of m-th component of mixture in i-th hidden state, M_i – number of components in mixture for hidden state of s_i . Let us suppose that in (1) the number of components for all hidden states are equal to M. This paper also deals with the case when a mixture of normal distributions describes probability density functions and observations are Z-dimensional. Thus, probability density function is of following form:

$$g(o_t; \mu_{im}, \Sigma_{im}) = (2\pi)^{-0.5Z} \left|\Sigma_{im}\right|^{-0.5} \exp^{-0.5(o_t - \mu_{im})\Sigma_{im}^{-1}(o_t - \mu_{im})^T},$$

where μ_{im} and Σ_{im} parameters are mean and covariance matrix respectively for m-th component of mixture for i-th hidden state $i = \overline{1, N}$, $m = \overline{1, M}$.

Thus, HMM is defined by non-observable (hidden) Markov chain, probability distribution of observation symbols and initial state distribution: $\lambda = (A, B, \Pi)$.

1.1 Classification of sequences produced by HMM

Let us define a problem of two-class classification. We have two groups of training sequences: first group was generated by HMM λ_1 and second – by HMM λ_2. It is assumed that generation HMMs differ from each other by parameters. Having unknown model parameters λ_1 and λ_2, we first estimate them (for example, with the use of Baum-Welch algorithm), and then classify some sequence O with the following decision rule:

$$\lambda = \arg \max_{i=\overline{1,2}} \left(lnP(O \mid \lambda_i) \right),$$

where $P(O \mid \lambda_i)$ – function of likelihood that sequence O was generated by model λ_i .

Usually, there occurs no problem if competing HMMs considerably differ by λ_1 and λ_2 parameters. For similar models the percent of correctly classified sequences may decrease to 50% level, i.e. models become indistinguishable. We shall name that method as “traditional”.

As an addition to the traditional HMM classification method we see the method mentioned above as the most convenient and prospective in means of improving HMMs discriminative abilities. It provides both some freedom in choosing the attributes space for classification and classifier itself. We will use space of derivatives of likelihood function logarithm with the respect to HMM parameters. Thus, characteristic vector for some sequence O will be of following block form:

$$V = \left(\frac{\partial lnP(O \mid \lambda_1)}{\partial \mu} \Big|_{\hat{\lambda}_1} \frac{\partial lnP(O \mid \lambda_2)}{\partial \mu} \Big|_{\hat{\lambda}_2} \right)^T,$$

where derivative from likelihood function logarithm is taken with the respect to some HMM parameter μ. It is calculated for first block with estimated parameters $\hat{\lambda}_1$ of first model and for second block with estimated parameters $\hat{\lambda}_2$ of second (competing) model. As classifier we will use support vector machine.

1.2 Selection of attributes information subspace

Usually researcher does not know what attributes causes the difference between models (transition matrixes or means), so it is important to choose attributes that will provide the best classification. Apart from that, sometimes models differ in large quantity of attributes, but not every such difference is informative. Therefore, it is important to choose information subspace of attributes, which would provide the best classification at the lowest cost (i.e. which includes the lowest number of attributes).

One of the key features of method that solves the problem is the criterion of attributes subspace informativity. The criterion can be direct and indirect.

Direct methods are based on cross-validation technique. Classified sample is randomly divided into training and testing parts. Training part is used to build the decision rule and testing part is classified with that rule. Keeping in mind the number of incorrect classification, the procedure is repeated with other training and testing parts. After several repetitions, the number of incorrections are summed up. The lower the sum is the better is the informativity of subsystem. Indirect methods are based on the evaluation of patterns distribution attributes. Thus, for normal distribution the Fisher information is a good criterion. The higher the distance between means of patterns is and the lower the variances of patterns are, the more information we can extract. Direct method is more resource consuming but it is believed that it gives more accurate estimation of recognition quality. That is why we have chosen the direct method of information subspace extraction, which is based on direct criteria AdDel [21]. AdDel algorithm is the combination of two simpler algorithms Add and Del.

Add algorithm (Addition). Suppose we have a N-dimensional attributes space and we need to choose attributes that are the most sufficient for classification. All N attributes are checked for informativity and the attribute, which produced the lowest number of incorrect recognitions, is added to information subsystem. Later it is added by all $N-1$ attributes. Resulted two-dimensional spaces are used to count the wrong classifications (i.e. now we build a classification based on two attributes). The most informative pair of attributes is chosen. This process continues until we get a system of n attributes.

Del algorithm (Deletion). Suppose we have a N-dimensional attributes space and we need to choose n attributes that are the most sufficient for classification. We shall evaluate a classification error when using all attributes. Then we shall step-by-step delete each attribute from that system. The attribute which deletion gives the lowest classification error shall be chosen to be excluded from system.This process continues until we get a system of n attributes.

Relaxation AdDel algorithm. Suppose we have a N-dimensional attributes space and we need to choose n attributes that are the most sufficient for classification. With the help of Add algorithm, attributes space (initially empty) is added by most informative n_1 attributes. Then with the help of Del algorithm, $n_2 < n_1$ attributes are excluded from it. Thus, one step increases attributes space by $n_2 - n_1$ attributes. This process continues until we get a system of n attributes.

2 Simulation results

We will now present the effectiveness of classifier that is based on method that uses space of derivatives of likelihood function logarithm with respect to HMM parameters. The most informative space in our research was constructed by AdDel algorithm. In addition, we compared effectiveness of this method and traditional one (based on likelihood function logarithm). Also, effectiveness of classifier that is based on method that uses space of derivatives of likelihood function logarithm with respect to HMM parameters without constructing the most informative attributes space. In that case, classification is made in space of derivatives of likelihood function logarithm with respect to HMM parameters, where the difference between models is included.

All simulations described below were conducted with following parameters: number of hidden states $N = 3$, dimensionality of observations $Z = 3$, number of components in mixture of Gaussian distributions $M = 3$, number of training sequences $K = 100$ and sequences length $T = 100$. We defined parameters of competing models and differences between them with the use of additional parameter of models proximity dA as follows:

$$A = \begin{pmatrix} 0.1 + dA & 0.7 - dA & 0.2 \\ 0.2 & 0.2 + dA & 0.6 - dA \\ 0.8 - dA & 0.1 & 0.1 + dA \end{pmatrix}.$$

For model λ_1 proximity parameter was defined to 0, and for model λ_2 difference from model λ_1 was expressed through additional proximity parameter dA.

AlDel algorithm modification of derivatives method require several attributes that can be included into information subspace. Thus, we will map corresponding attributes to numbers below:

$$A^{\lambda_1} : 1-9, A^{\lambda_2} : 130-138; \mu^{\lambda_1} : 10-36, \mu^{\lambda_2} : 139-165; \Sigma^{\lambda_1} : 37-117, \Sigma^{\lambda_2} : 166-246;$$

$$\pi^{\lambda_1} : 118 - 120, \pi^{\lambda_2} : 247 - 249; \tau^{\lambda_1} : 121 - 129, \tau^{\lambda_2} : 250 - 258.$$

For example, the first element of characteristic vector V will be a derivative of likelihood function logarithm with respect to a_{11} element of transition probabilities matrix that was calculated with estimated parameters of λ_1 model. One hundred and thirtieth element of characteristic vector V will be a derivative of likelihood function logarithm with respect to a_{11} element of transition probabilities matrix that was calculated with estimated parameters of λ_2 model. For AdDel algorithm, the values of $n_1 = 2, n_2 = 1$ were chosen on bending of quality curve (i.e. we chose value, after which the informativity of algorithm work stopped changing). Therefore, result tables were filled with the optimal number of attributes. The simulation was carried out, where competing models differed only in elements of transition probabilities matrix A. Results are shown in Table 1.

Results show us that method based on derivatives is still more effective than traditional ant its AdDel modification is even more effective: the advantage is up to 5% more of correctly classified sequences with $dA = 0.05$. In addition, we must note that resulted information subspace consists not only of attributes that are different in reality: derivatives with respect to covariance matrix elements and to initial state distribution elements (for example, when $dA = 0.2$). The results presented above show that AdDel algorithm, used for choosing the attributes information space, in addition to derivatives method work quite well for classification of sequences, which were generated by models that differ in various parameters. At the same time, its usage improves the effectiveness of this classification method. In addition, combination of this method and AdDel algorithm proved to be resistant to disturbances in observation sequences. In some cases, the combination classified correctly up to 30% more sequences compared to traditional method.

Table 1: Dependence of the percents of correctly classified sequences and numbers of attributes on additional parameter of models proximity dA

dA	Percent of correctly classified sequences with the use of traditional method	Percent of correctly classified sequences with the use of method based on derivatives	Percent of correctly classified sequences with the use of method based on derivatives (with the use of AdDel algorithm)	Numbers chosen for the highest informativity of attributes
0.01	53.5	56.5	66.5	91-99 251 121 247 125 2
0.03	53.0	66.0	71.5	7 131 148-150 109-117 10-12 184-192
0.05	65.5	73.5	79.5	131 7 9 100-108 121 123 16-18
0.1	84.0	87.5	91.0	134 135 138 148-150 1 130
0.2	96.5	97.0	99.5	1 5 7 135 55-63 82-90 118

3 Conclusions

In this paper, we described the new approach for classification of multidimensional sequences, which is based on HMM. Simulations showed that its effectiveness is preserved even if dimensionality of observations is increased. This approach proved to be significantly better for similar models (no matter in what parameters they differed), especially when AdDel algorithm was used for choosing of attributes space. Thus, this work open up new perspectives for classification of multidimensional sequences generated by Hidden Markov Models, which have many practical implementations. Classification of multidimensional sequences proved especially effective compared to traditional classification method in several cases. For example, when observation sequence was disturbed by probabilistic noise (with probability $P = 0.1$, $\epsilon \succ Cauchy(0, 0.5)$) and competing models were different in transition probabilities matrix with $dA = 0.1$ this method provided 88% of correctly classified sequences compared to 62% for traditional classification. As to its AdDel modification – it provided 91.5% correctly classified sequences (and only 7 attributes were needed for this). To sum up, we can say that method, based on derivatives of likelihood function logarithm with respect to HMM parameters, modified by AdDel algorithm, showed exceptional results and is a very perspective method for classification of multidimensional sequences, described by hidden Markov models, in case of differences in any of parameters and in case of disturbed multidimensional sequences.

Acknowledgements

This research has been supported by the Russian Ministry of Education and Science (project No 2.541.2014K).

References

[1] Gultyaeva T.A. Classification of sequences with the use of Hidden Markov Models of inexact structure / T.A. Gultyaeva, A.A. Popov // *Herald of TSU. Management, Computer Engineering and Informatics* , Tomsk : TSU publishing house, 2013. Vol. **3 (24)**, pp. 57-63. (in Russian)

[2] Gultyaeva T.A. Classification of sequences generated by Hidden Markov Models in multiclass case. / T.A. Gultyaeva, A.A. Popov // *Scientific herald of NSTU*, Novosibirsk : NSTU publishing house, 2013. Vol. **3 (52)**, pp. 40-45. (in Russian)

[3] Gultyaeva T.A. The Classification of Noisy Sequences Generated by Similar HMMs / T.A. Gultyaeva, A.A. Popov // *PReMI-2011, LNCS*, Springer-Verlag Berlin Heidelberg, 2011. Vol. **6744/2011**, pp. 30-35.

[4] Gultyaeva T.A. Application of the hidden Markov models, kNN and SVM for a classification problem modes of power supply system/ T.A. Gultyaeva, D.Y. Korotenko, A.A. Popov //*Actual problems of electronic instrument engineering (APEIE-2012) : 11th International Conference*, Novosibirsk : NSTU publishing house, 2012. Vol. **1**, pp. 36-40.

[5] Gultyaeva T.A. Classification of multidimensional observation sequences described by Hidden Markov Models / T.A. Gultyaeva, V.V. Kokoreva //*Actual problems of electronic instrument engineering (APEIE-2014) : 12th International Conference*, Novosibirsk : NSTU publishing house, 2014. Vol. **1**, pp. 556-561.

[6] Gultyaeva T.A. Graphics processing unit implementation of Hidden Markov models / T.A. Gultyaeva, A.S. Sautin, V.E. Uvarov // *Actual problems of electronic instrument engineering (APEIE-2014) : 12th International Conference*, Novosibirsk : NSTU publishing house, 2014. Vol. **1**, pp. 571-573.

[7] Vintsiuk T.K. Word recognition from speech with use of dynamic programming / T.K. Vintsiuk //*Cybernetics*, Moscow, 1968. Vol. **1**, pp. 15-22. (in Russian)

[8] Kovalevsky V.A. Optimal algorithm of recognition of some image sequences / V.A. Kovalevsky // *Cybernetics*, 1967. Vol. **4**, pp. 75-80. (in Russian)

[9] Baum L.E. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains / L.E. Baum [et. al.] // *The Annals of Mathematical Statistics*, 1970. Vol. **1 (41)**, pp. 164-171.

[10] Alexandrov V. Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures [Web source] / V. Alexandrov, M. Gerstein // BMC Bioinformatics, 2004. Vol. **5**, Iss. **2**. http://www.biomedcentral.com/1471-2105/5/2.

[11] Neal R.M. Inferring State Sequences for Non-linear Systems with Embedded Hidden Markov Models / R.M. Neal, M.J. Beal, S.T. Roweis // Published in book: *Advances in Neural Information Processing Systems 16* / S. Thrun, L.K. Saul, B. Scholkopf, 2003. pp. 401-408.

[12] Walder C.J. Towards a Maximum Entropy Method for Estimating HMM Parameters / C.J. Walder, P.J. Kootsookos, B.C. Lovell // *Proceedings of the 2003 APRS Workshop on Digital Image Computing*, 2003. Vol. **1**, pp. 45-19.

[13] Ikbal S. HMM/ANN Based Spectral Peak Location Estimation for Noise Robust Speech Recognition / S. Ikbal, H. Bourlard, M. Magimai-Doss //*Acoustics, Speech, and Signal Processing. Proceedings IEEE International Conference ICASSP'05* , 2005. Vol. **1**, pp. 453-456.

[14] Zhou G. Discriminative hidden markov modeling with long state dependence using a knn ensemble / G. Zhou // *International Conference on Computational Linguistics*, 2004, pp. 22-28.

[15] Liu C. Estimation of the t distribution using EM and its extensions, ECM and ECME / C. Liu, D.B. Rubin // *Statistica Sinica*, 1995. Vol. **5**, pp. 19-39.

[16] Chatzis S.P. A Variational Bayesian Methodology for Hidden Markov Models utilizing Student's-t Mixtures / S.P. Chatzis, D.I. Kosmopoulos //*Pattern Recognition*, 2011. Vol. **44**, Iss. **2.**, pp. 295-306.

[17] Solera-Urena R. Robust ASR using Support Vector Machines / R. Solera Urena, D. Martin-Iglesias, A. Gallardo-Antolin // *Speech Communication*, 2007. Vol. **49**, pp. 253-267.

[18] Ling Ch. Combination of Fisher Scores and Appearance Based by Features For Face / Ch. Ling, M. Hong // *Proc. of the 2003 ACM SIGMM Workshop on Biometrics Methods and Applications*, 2003, pp. 74-81.

[19] Aran O. Recognizing two handed gestures with generative. discriminative and ensemble methods via Fisher kernels / O. Aran, L. Akarun //*In LNCS: Multimedia Content Representation. Classification and Security International Workshop*, 2006. Vol. **4015**, pp. 159-166.

[20] Mottl V.V. *Hidden Markov Models in Structural Signal Analysis* / V.V. Mottl, I.B. Muchnik. Moscow: FIZMATLIT, 1999. 352 p. (in Russian)

[21] Zagoryiko N.G. *Cognitive data analysis* / N. G. Zagoryiko; Russia, Siberian department of science academy, Sobolev institute of mathematics. Novosibirsk: issued by academic publisher “Geo”, 2013. 186 p. (in Russian)

Theorem of Learning for a Competition Algorithm

Antyufeev V.S.
Institute of Computational Mathematics and Mathematical Geophysics SD RAS,
Novosibirsk State University
Novosibirsk, Russia
e-mail: ant@osmf.sscc.ru

Abstract

This paper is an extension of [1], where a new decisive algorithm was proposed. In its operation, the unit resembles artificial neural networks. However, the functioning of the algorithm proposed is based on different concepts. It does not use the concept of a net or a neuron. The theorem of learning for the new competition algorithm is proved.

Keywords: theorem of learning, probabilistic convergence, artificial neural network.

Introduction

Creation of artificial systems of pattern recognition remains a difficult theoretical and engineering problem. The necessity of pattern recognition arises in various fields of human activities: from military science and safety systems to digitization of analog signals.

ANNs are logical algorithms whose operation is associated with biological concepts of brain functioning [10, 11]. The complexity of these algorithms hampers theoretical investigations of ANNs. It is difficult for researchers to understand what actually happens "inside" the network.

We proposed [1] a new decisive algorithm, a competition algorithm This algorithm allows solving approximately the same problems as ANNs solve. The learning and operation of this algorithm, however, are based on different principles. These new principles made it possible to simplify the algorithm and to make its operation understandable.

1 Notations, definitions, auxiliary statements

The new competition algorithm is fuzzy. It can be identified with the function CA : $R^n \to [0,\ 1]$. Like the ANN, the CA is subjected to a learning procedure.

The signal z is a numerical tuple, a point in R^n. The signal coordinates $z_k,\ k = 1,\ \dots\ , n$ are the parameters of the examined object; n is a fixed dimension of the space of signals.

Let us use Z ($Z \subset R^n$) to designate the set of all admissible signals. Let Z be a set measurable with respect to the Lebesgue measure. The set Z is the main space for signals. Therefore, in what follows, we consider only the points-signals $z \in Z$ and the measurable sets $U, V, \ldots \subset Z$. Their measures are indicated by $\mu(U)$, $\mu(V)$, $\ldots$, respectively. Let us assume that $0 < \mu(Z) < \infty$ by definition.

We divide the set Z into subsets X, Y: $X \cup Y = Z$, $X \cap Y = \emptyset$, $\mu(X) > 0$: for points from the sets X and Y, the answers to the main questions are positive and negative, respectively. The set X is a decisive set. The idea of solving the recognition problem is to give an answer to the main question for an arbitrary signal $z \in R^n$: is it true that z belongs to X?

An auxiliary function $\varepsilon(z)$ is an indicator or a sign of the signal z:

$$\varepsilon(z) = \begin{cases} +1, & z \in X \\ -1, & z \in Y \end{cases}$$

Each learning signal z^k generates a scalar influence field around it [1]. The sign of this field coincides with the sign $\varepsilon(z^k)$ of the signal z_k.

Let us define a function

$$h(z) = \frac{1}{|z|^m}. \tag{1}$$

Here $m \geq n$ is an arbitrary number and n is the dimension of the space of signals. We find the meaning of the number m when proving the theorem of learning.

The intensity of the field induced by the learning signal z^k at an arbitrary point $z \in Z$ is determined by the signal influence function [1]:

$$h_k(z) = \varepsilon(z^k)\ h(z - z^k). \tag{2}$$

Let $z^1, z^2, \ldots, z^N$ be a finite sequence of learning signals. Each of the signals z^k induces its influence field with the intensity $h_k(z)$ at the point z. The superposition of these fields is determined by summation. The intensity of the superposition of the fields at the point z is indicated by $H_N(z)$:

$$H_N(z) = \sum_{k=1}^{N} h_k(z) \tag{3}$$

$H_N(z)$ is the influence function of the total set of the learning signals $\{z^1, z^2, \ldots z^N\}$.

$F(t)$ is an auxiliary function of one variable with a specific s-shaped graph:

$$F(t) = \frac{1}{2}\left[\frac{t}{\sqrt{t^2+1}} + 1\right]$$

$f_N(z)$ is an approximate decisive function corresponding to the learning signals $z^1, z^2, \ldots, z^N$:

$$f_N(z) = \begin{cases} F(H_N(z)), & z \neq z^k, & k = 1, \ldots, N \\ 1, & z = z^k \in X, & k = 1, \ldots, N \\ 0, & z = z^k \in Y, & k = 1, \ldots, N \end{cases} \tag{4}$$

The ideal decisive function

$$\chi_X(z) = \begin{cases} 1, & z \in X \\ 0, & z \notin X \end{cases}$$

is the characteristic function [16] of the decisive set X.

2 Learning theorem

In this section, we prove that the competition algorithm can be trained “to the best possible extent” if the learning signals in Z are chosen in a special manner. In other words, correct learning allows the fraction of correct answers to be brought close to unity.

In proving the theorem, we use the following statement from the mathematical analysis [17]. Let $\lim\limits_{k\to\infty} a_k = 0$ and $a_k \geq b_k \geq 0$ for all k. Then we have $\lim\limits_{k\to\infty} b_k = 0$. We refer to the assertion of a sequence.

Theorem. Let $\{z^1, z^2, \dots\}$ be an infinite sequence of mutually independent random points uniformly distributed on the set $Z \subset R^n$, $\quad 0 < \mu(Z) < \infty$, Then the sequence of approximate decisive functions $f_k(z)$ converges in terms of probability to an ideal decisive function $\chi_X(z)$ at all points $z \in Z$ except for, maybe, points of the boundary $\Gamma(X)$ of the sought set X.

In other words, if the learning points-signals are chosen in the above-described manner, then the following limiting relations are satisfied:

$$f_k(z) \xrightarrow{p} 1 \quad \forall z \in \mathrm{Int}(X) \qquad f_k(z) \xrightarrow{p} 0 \quad \forall z \in \mathrm{Int}(Y) \tag{5}$$

Here $\mathrm{Int}(C)$ is the set of internal points of the set C, i.e., points that belong to W together with a certain neighborhood. *Proof.* Let $x^* \in X$ be an arbitrary function that does not lie on the boundary of the set X. We are going to prove that $f_k(x^*) \xrightarrow{p} 1$. If $y^* \in Y$ is a point that does not lie on the boundary of X, then the limiting relation $f_k(y^*) \xrightarrow{p} 0$ can be proved in a similar way. Obviously, the theorem is valid if the point x^* coincides with one of the learning signals $z^k \in X$, because in this case $f_k(z^k) \equiv 1$ for all $k = 1, 2, \dots$ in accordance with definition (4). Let the point $x^* \in X$ coincide with none of the learning signals z^k. According to definition (4) of the decisive function f_k (see also [1]), statement (5) is equivalent to the following limiting relation:

$$H_N(x^*) \xrightarrow{p} +\infty, \quad N \to \infty \tag{6}$$

The convergence in terms of probability to an infinite limit is determined in the course of proving the theorem by analogy with the convergence to a finite limit [19].

Thus, the proof of the theorem reduces to proving the limiting relation (6). Without loss of generality, we assume for convenience that

$$x^* = \mathbf{0}. \tag{7}$$

If $x^* \neq \mathbf{0}$, we perform a parallel transposition of the coordinate system in R^n.

We divide the set of learning signals $z^1, z^2, \dots, z^N$ into subsets:

$$\{z^1, z^2, \dots, z^N\} = \{x^1, x^2, \dots, x^{N_X}\} \cup \{y^1, y^2, \dots, y^{N_Y}\}, \qquad N_X + N_Y = N \tag{8}$$

The first group contains signals z^k which fall into X; the second group consists of signals z^k which fall into Y. Write and transform the expression for $H_N(x^*)$. According to definitions (2) and (3), we have

$$H_N(x^*) = \sum_{k=1}^{N} h_k(x^*) = \sum_{k=1}^{N} \frac{\varepsilon(z^k)}{|z^k - x^*|^m}$$

As $x^* = \mathbf{0}$ (see Eq. (7)), this expression can be simplified:

$$\sum_{k=1}^{N} \frac{\varepsilon(z^k)}{|z^k - x^*|^m} = \sum_{k=1}^{N} \frac{\varepsilon(z^k)}{|z^k|^m}$$

Taking into account the relations $\varepsilon(x^i) = +1, \quad \varepsilon(y^j) = -1$, and the division of the set $\{z^1, z^2, \ \dots \ , z^N\}$, we obtain

$$\sum_{k=1}^{N} \frac{\varepsilon(z^k)}{|z^k|^m} = \sum_{i=1}^{N_X} \frac{1}{|x^i|^m} - \sum_{j=1}^{N_Y} \frac{1}{|y^j|^m}$$

For brevity, we designate

$$H_N^X = \sum_{i=1}^{N_X} \frac{1}{|x^i|^m}, \quad H_N^Y = \sum_{j=1}^{N_Y} \frac{1}{|y^j|^m}$$
$$(N_X + N_Y = N)$$

Thus, we have

$$H_N(x^*) = H_N^X - H_N^Y \tag{9}$$

Recall that, in accordance with the theorem conditions, the learning signals x^i, y^j are random points; therefore, H_N^X, H_N^Y are random quantities. Find the probability limits for the expressions $\frac{1}{N}H_N^X$ and $\frac{1}{N}H_N^Y$.

First consider the expression

$$\frac{1}{N_X} H_N^X = \frac{1}{N_X}(h(x^1) + \ \dots \ + h(x^{N_X})) \tag{10}$$

(see definition (1) of the function h) and find its limit. Let $\boldsymbol{x}$ be a random point uniformly distributed on the set X. Then $\boldsymbol{h} = h(\boldsymbol{x})$ is a random quantity. Points x^i can be considered as sampled values of $\boldsymbol{x}$. Expressions of the form (10) are used as statistical estimates for calculating multidimensional integrals by the Monte Carlo method [18] on the basis of the laws of large numbers.

According to Kolmogorov's theorem [19], a random quantity of the form (10) converges in terms of probability to the corresponding integral of h if and only if there exists a finite mathematical expectation $\mathbf{E}\boldsymbol{h}$.

This integral of h, however, diverges in the neighborhood of the singular point $\mathbf{0}$ at $m \geq n$ [17]; therefore, the finite mathematical expectation $\mathbf{E}\boldsymbol{h}$ does not exist.

Calculate the limit of Eq. (10), using only random quantities with a finite mathematical expectation.

Let us use $S(\varepsilon)$ to indicate a sphere of radius $\varepsilon > 0$ in R^n with the center at the point $\mathbf{0}$.

Eliminate the neighborhood $S(\varepsilon)$ of the singular point $\mathbf{0}$ from X and indicate the remaining set by $X_\varepsilon = X - S(\varepsilon)$ (Fig. 2). Consider the regularized function

$$h_\varepsilon = \begin{cases} h(x), & x \in X_\varepsilon \\ 0, & x \in S_\varepsilon \end{cases} \tag{11}$$

The function h_ε is bounded; therefore, the integral of h_ε over the domain X converges. Denote its value by I_ε:

$$I_\varepsilon = \int\limits_X h_\varepsilon(x)dx < \infty$$

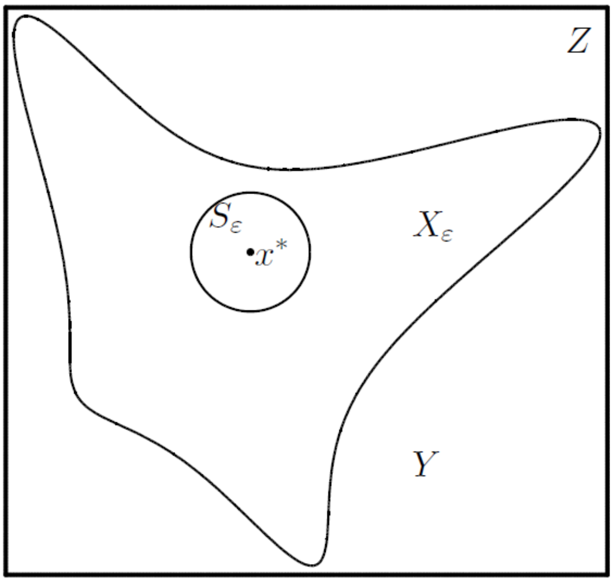


Figure 1: Division of the set X

As the integral of the function h diverges in the neighborhood of the point $\mathbf{0}$, the following limiting relation is valid:

$$\lim_{\varepsilon\to 0} I_\varepsilon = \lim_{\varepsilon\to 0} \int\limits_{X\setminus S_\varepsilon} h(x)dx = +\infty \tag{12}$$

Choose an arbitrary number $A > 0$. By virtue of Eq. (12), there exists such a value $\varepsilon_A > 0$, that the following inequality is satisfied at $\varepsilon < \varepsilon_A$:

$$I_\varepsilon > A \tag{13}$$

The function h_ε corresponds to a regularized random quantity $\boldsymbol{h}_\varepsilon = h_\varepsilon(\boldsymbol{x})$ with a finite mathematical expectation $\mathbf{E}\boldsymbol{h}_\varepsilon$. For $\boldsymbol{h}_\varepsilon$, the conditions of the above-mentioned Kolmogorov's theorem are satisfied. Therefore, we have

$$\frac{1}{N_X}(h_\varepsilon(x^1) + \ \ldots \ + h_\varepsilon(x^{N_X})) \xrightarrow{p} \mathbf{E}\boldsymbol{h}_\varepsilon = I_\varepsilon < \infty$$

as $N_X \to \infty$. In other words, for any fixed $\delta > 0$, the limiting relation

$$P\left(\left|\frac{1}{N_X}(h_\varepsilon(x^1) + \ \ldots \ + h_\varepsilon(x^{N_X})) - I_\varepsilon\right| > \delta\right) \to 0 \tag{14}$$

holds as $N_X \to \infty$. Choose a positive value of $\varepsilon < \varepsilon_A$ and $\delta = I_\varepsilon - A$. Then, according to Eq. (13), we have $\delta > 0$.

Use a simple remark: for any random quantity $\boldsymbol{a}$, the following inequality is satisfied:

$$P(|\boldsymbol{a}| > \delta) = P(\boldsymbol{a} > \delta) + P(\boldsymbol{a} < -\delta),$$

whence it follows that

$$P(|\boldsymbol{a}| > \delta) \geq P(\boldsymbol{a} < -\delta)$$

It follows from Eq. (14) and from the last remark that the following inequality is satisfied:

$$P\left(\left|\frac{1}{N_X}(h_\varepsilon(x^1) + \ \ldots \ + h_\varepsilon(x^{N_X})) - I_\varepsilon\right| > \delta\right) \geq$$

$$\geq P\left(\frac{1}{N_X}(h_\varepsilon(x^1) + \ \ldots \ + h_\varepsilon(x^{N_X})) - I_\varepsilon < -\delta\right) \geq 0$$

Taking into account the limiting relation (14) and $\delta = I_\varepsilon - A > 0$ we find from the statement on the sequence that

$$P\left(\frac{1}{N_X}(h_\varepsilon(x^1) + \ \ldots \ + h_\varepsilon(x^{N_X})) - I_\varepsilon < -\delta\right) =$$

$$= P\left(\frac{1}{N_X}(h_\varepsilon(x^1) + \ \ldots \ + h_\varepsilon(x^{N_X})) < I_\varepsilon - \delta\right) =$$

$$= P\left(\frac{1}{N_X}(h_\varepsilon(x^1) + \ \ldots \ + h_\varepsilon(x^{N_X})) < A\right) \to 0$$

as $N_X \to \infty$.

One more remark should be made: if the inequality $\boldsymbol{a} \geq \boldsymbol{b}$ is satisfied for random quantities $\boldsymbol{a}$, $\boldsymbol{b}$ and if C is an arbitrary number, then we obtain

$$P(\boldsymbol{b} < C) \geq P(\boldsymbol{a} < C).$$

As $h(z) \geq h_\varepsilon(z)$ (see Eq. (11)), it follows from this remark that

$$P\left(\frac{1}{N_X}(h_\varepsilon(x^1) + \ \ldots \ + h_\varepsilon(x^{N_X})) < A\right) \geq P\left(\frac{1}{N_X}(h(x^1) + \ \ldots \ + h(x^{N_X})) < A\right)$$

Therefore, in accordance with the statement on the sequence, we have

$$P\left(\frac{1}{N_X}(h(x^1) + \ \ldots \ + h(x^{N_X})) < A\right) \equiv P\left(\frac{1}{N_X}H_N^X < A\right) \to 0$$

Recall that the number A here is arbitrary; therefore, its value can be chosen to be arbitrarily large. By analogy with determining the convergence in terms of probability to a finite limit, we can assume that the last limiting relation means the convergence of the sequence of random quantities $\frac{1}{N_X} H_N^X$ to infinity:

$$\frac{1}{N_X} H_N^X \xrightarrow{p} +\infty \tag{15}$$

Now we have to find the probability limit of the random quantity $\frac{1}{N} H_N^X$ ($N = N_X + N_Y > N_X$). We rewrite the expression for $\frac{1}{N} H_N^X$:

$$\frac{1}{N} H_N^X = \left(\frac{N_X}{N}\right)\left(\frac{1}{N_X} H_N^X\right)$$

The limit of the multiplier $\left(\frac{1}{N_X} H_N^X\right)$ has been already found (see Eq. (15)). We rewrite the multiplier $\left(\frac{N_X}{N}\right)$:

$$\frac{N_X}{N} = \frac{\chi_X(z^1) + \ \dots \ + \chi_X(z^N)}{N}$$

In the numerator of the last fraction, the terms $\chi_X(z^k)$ are random Bernoulli quantities. Those N_X of them for which $z^k = x^i \in X$ are equal to unity, and the remaining terms are equal to zero. According to the law of large numbers, we have

$$\frac{\chi_X(z^1) + \ \dots \ + \chi_X(z^N)}{N} = \frac{\chi_X(x^1) + \ \dots \ + \chi_X(x_X^N)}{N} \xrightarrow{p}$$

$$\xrightarrow{p} E(\chi_X(\boldsymbol{x})) = \int\limits_Z \chi_X(x)dx = \frac{\mu(X)}{\mu(Z)} > 0, \tag{16}$$

because $\mu(X) > 0$ and $\mu(Z) < \infty$ by definition. Equations (15) and (16) yield the sought limiting relation (in the sense defined above):

$$\frac{1}{N} H_N^X \xrightarrow{p} +\infty \tag{17}$$

2. Now return to Eq. (9) and find the estimate from above for the expression jjHJj. This estimate can be obtained without using the probability theory. Let us indicate the distance from the point x^* to the set Y by d:

$$d = \inf_{y \in Y} |y - x^*| = \inf_{y \in Y} |y - \mathbf{0}| = \inf_{y \in Y} |y| \tag{18}$$

According to the conditions of the theorem, we have $x^* \in \mathrm{Int}(X)$; therefore, $d > 0$. Obviously, the following inequality is satisfied:

$$H_N^Y = \sum_{j=1}^{N_Y} \frac{1}{|y^j|^m} \le N_Y \cdot \max_{y^j \in Y} \frac{1}{|y^j|^m} \tag{19}$$

Let us estimate both multipliers in the right-hand side of the last inequality. As $N_X + N_Y = N$ (see Eq. (8)), then

$$N_Y \leq N. \tag{20}$$

As $\inf\limits_{y \in Y} |y| = d$ (see Eq. (18)), the inequality $|y^j| \geq d$ is valid for all points $y^j \in Y$. Therefore, we have

$$\max_{y^j \in Y} \frac{1}{|y^j|^m} = \frac{1}{\min\limits_{y^j \in Y} |y^j|^m} \leq \frac{1}{d^m}. \tag{21}$$

From Eqs. (19)–(21), we obtain the estimate from above for $\frac{1}{N} H_N^Y$: Here the constant C is independent of the number N of the learning points. Thus, we proved (17) that

$$\frac{H_N^X}{N} \xrightarrow{p} +\infty,$$

and inequality

$$\frac{H_N^Y}{N} \leq C < +\infty$$

is satisfied for all N. It follows from here that

$$\frac{H_N^X - H_N^Y}{N} \xrightarrow{p} +\infty,$$

and, moreover,

$$[H_N^X - H_N^Y] \xrightarrow{p} +\infty.$$

Acknowledgements

The reported study was partially supported by RFBR, research project No 13-01-00441 a.

References

[1] Antyufeev V.S. (2012). Solution of recognition problems by the Monte Carlo method. *Russian Journal of Numerical Analysis and Mathematical Modelling*. Vol. **27**, No. **2**, pp. 113-130.

[2] Feller W. (1970). *Introduction to probability theory and its application*. John Wiley, New York.

[3] Fichtenholz G.M.(1969). *Course of Differential and Integral Calculus.*. Nauka, Moscow.

Research of the Lambda Wilks Statistic Distribution under Conditions of Violation of the Main Assumptions in the Discriminant Function Analysis

V. M. Volkova and A. A. Sanina
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: volkova@ami.nstu.ru, anastas.sanina@gmail.com

Abstract

This paper studies how the lambda Wilks statistic distribution in the discriminant function analysis changes under conditions of violation of data normality and how the variables with the low tolerance value influence on the statistic distribution. Monte Carlo method has been used for statistical regularities simulation. The article is also addressed the test power based on the lambda Wilks statistic.

Keywords: lambda Wilks statistic, discriminant function analysis, violation of normality assumptions.

Introduction

The discriminant function analysis is a subdiscipline of the multiple statistical analysis, which allows us to study the differences between two or more object groups considering several variables simultaneously [2].

As usual, classification methods are associated with building one or more discriminant functions, providing the possibility of assigning a new value to one of the object groups.

The creation of discriminant functions is necessary for decision-making in the discriminant function analysis. The significance of the discriminant functions indicates whether the differences between the average values of variables in the groups are really statistically significant, or these differences are due to the random fluctuations around an overall average value [2, 4]. The statistical significance is tested using the lambda Wilks statistic.

Basic assumptions of the discriminant function analysis are related to the data normality, variance homogeneity and the absence of redudant variables (i.e. the absence of variables with a low tolerance value). The previous experiments were made when the first assumption was false and the assumption of independent variable and variance homogeneity was true. The obtained results were published in [5].

This work is devoted to the results obtained under condition of violation of the third assumption too. It is tried to answer the question how the input variable with a low tolerance value influences on the final result and conclusion.

1 Problem Definition

1.1 The cases of null and alternative hypotheses

The null hypothesis is that the average values of the discriminant function are equal for both groups. In order to enforce the terms of null hypothesis the simulation was performed with equal average values of the variables for all groups. Those let the null-hypothesis be presented in a general form as:

$$\mu_{1k} = \mu_{2k} = \mu,$$

where $k = \overline{1,\, p}$, μ_{ik} is the average value in the i-th group for the k-th variable; p is the number of variables. The concrete alternative hypotheses are presented by the expression:

$$H_1 = \mu_{1k} = \mu_{2k} + c\sigma_{1k}^2,$$

where c is a numerical coefficient, σ_{ik}^2 is the variance in the i-th group for the k -th variable. In doing so, the equality of average values is observed for one variable in the case of the H_{11} - hypothesis, for two variable in the case of the H_{12} - hypothesis and for all variables in the case of the H_{13}-hypothesis.

1.2 Two-group Lambda Wilks Statistics

The assumption of variable normality and variance equality is postulated. The lambda Wilks statistics is calculated in several steps to test the null hypothesis [2].

Primarily, it is necessary to build the scatter matrix T, its elements are calculated using the formula:

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} \left(X_{ikm} - \overline{X}_i \right) \left(X_{jkm} - \overline{X}_j \right),$$

where $i,\, j = \overline{1,\, p}$, X_{ikm} is the value of the discriminant variable X_i for a m-th object in the k-th class; $\overline{X}_i$ is an average value of an X_i variable for all classes; n is total observations; g is the number of classes.

After that the intra group variation matrix W is calculated. This matrix determines the variance of values within the classes. Its elements are calculated by the formula:

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} \left(X_{ikm} - \overline{X}_{ik} \right) \left(X_{jkm} - \overline{X}_{jk} \right),$$

where $\overline{X}_{ik}$ is an average value of an X_i variable in the k-th class. The other notations are similar to the one used previously.

The next step is the calculation of the B-matrix:

$$B = T - W,$$

which determines the intergroup variance. Its elements are calculated according to the formula:

$$b_{ij} = t_{ij} - w_{ij}.$$

The next step is solving a set of equations which presented in the following form:

$$\left(B - \lambda_i W\right) v_j = 0,$$

where λ_j is eigenvalue of the BW^{-1}-matrix. Thus, the result of solving the set of equations allows us to estimate the values of the eigenvectors, corresponding to the discriminant functions. The formula

$$\Lambda = \prod_{i=1}^{2} \frac{1}{1+\lambda_i},$$

is used to calculate the lambda Wilks statistic, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ [2]. If the null-hypothesis of equality of average values is true, then the lambda Wilks statistic

$$\chi^2 = -\left\{n - \frac{p+2}{2} - 1\right\} \ln\left(\Lambda\right) \tag{1}$$

is distributed according to the χ^2-law with the p degrees of freedom, where p is the number of discriminant variables X_i.

2 Experimental Conditions

The discriminant function analysis with two groups and three variables ($p = 3$) has been considered in this article. The investigations were performed under conditions of the intergroup variance homogeneity. Experiments where data is distributed according to the Normal law and to the law with "light" or "heavy" tails have been carried out using the Double Exponential Law ($De(\lambda)$) with different shape parameter value. Its probability density function is

$$f\left(x,\, \lambda,\, \theta_1,\, \theta_2\right) = \frac{\lambda}{2\sqrt{2}\theta_2 \Gamma\left(1/\lambda\right)} \exp\left\{-\left(\frac{|x-\theta_1|}{\sqrt{2}\theta_2}\right)^2\right\}.$$

The cases of independence and multicollinearity random vector variables has been considered. For case of random dependent vector variables the formula for the covariance matrix has been obtaned. The experiments were performed with a program system developed in the framework of a common methodology for a statistical simulation. This technique is developed by a scientific school under the guidance of Professor B. Yu. Lemeshko [3] in Novosibirsk State Technical University at the Faculty of Applied Mathematics and Computer Science. The program toolkit allows us to simulate pseudo-random values which can be distributed according to the Gauss law or any other laws.

3 Main Results

3.1 The Results of Investigations in Case of Variables Independence

The investigations have been conducted under condition of sample size equality for all groups while data is distributed according to the law with “heavy” or “light” tails (De(0.5) or De(10)), to the Normal law (De(2)) and to significantly asymmetric laws (in this case data is distributed according to the Exponential law). When conducting the investigations, the assumption of variable independence is postulated.

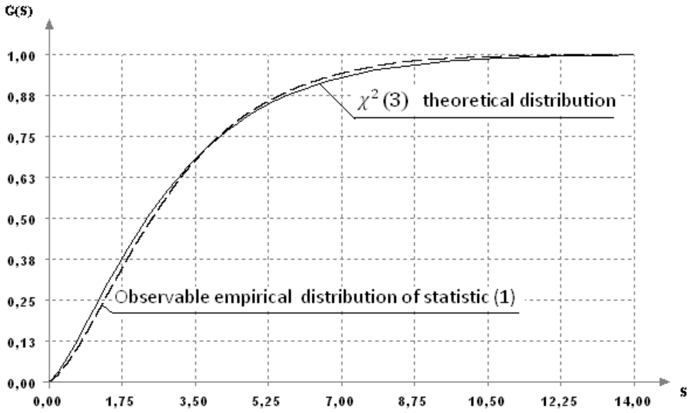


Figure 1: The Lambda Wilks statistic distribution while data is distributed according to the $De(0.5)$-law and $n_1 = n_2 = 20$.

There are some graphic deviations of the experimental lambda Wilks statistic distribution from the parent distribution if data is distributed according to the law with “heavy” tails with the sample size ($n_1 = n_2 < 100$) being small. Pic. 1 does show it. And you can also see despite the fact that the emperical distribution does not converge with a theoretical distribution the maximum error is not more than 0.015 having calculated the first kind error using a theoretical distribution instead of the empirical distribution. In addition, it should be noted that in the case of a big sample size ($n_1 = n_2 = 100$) the experimental lambda Wilks statistic distribution is visually the same as the theoretical χ^2-distribution. The same character of the lambda Wilks statistic distribution is also observed if data is distributed according to the significantly asymmetric law. The goodness-of-fit hypothesis between the experimental lambda Wilks statistic distribution and the parent distribution is not rejected if data is distributed according to the law with “light” tails or to the Normal law, with being equal to 0.01. Pic. 2 shows the degree of agreement of the theoretical and empirical statistic distribution if data is distributed according to the law with “light” tails.

In doing so, the lambda Wilks test power study with different c-parameter values has revealed that it does not depend on the law. Figure 3 illustrates the trend depending on the c-parameter value if data is distributed according to the De(0.5) law. As it was expected the more variables are a part of the alternative hypothesis

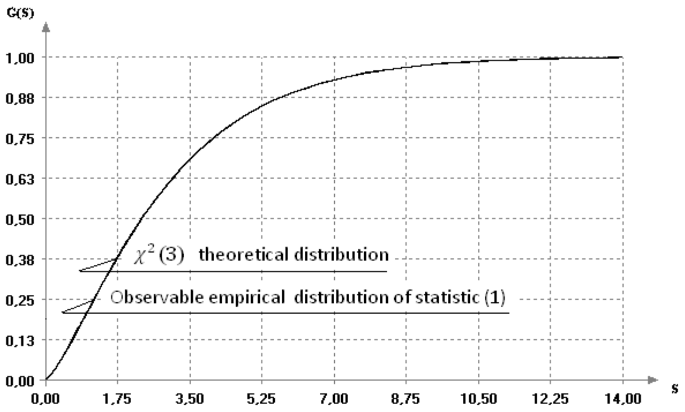


Figure 2: The Lambda Wilks statistic distribution while data is distributed according to the $De(10)$-law and $n_1 = n_2 = 20$.

(i.e. the assumption of the means equality in groups is not true for more variables) the more statistic power. But if the mean values in groups differ from each other by the value which equals to double variance value the statistic power is almost one for all alternative hypotheses.

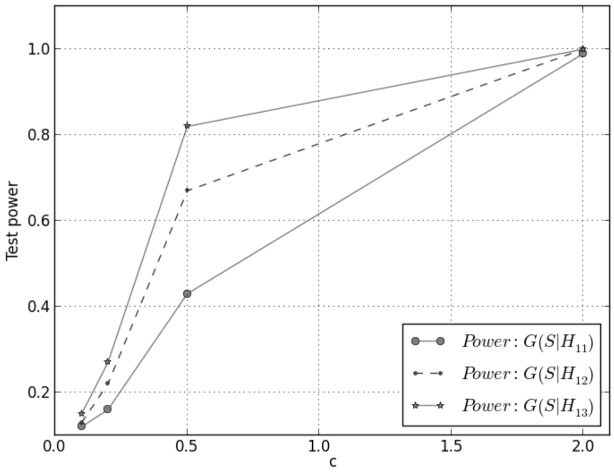


Figure 3: The comparison of the lambda Wilks test power when data is distributed according to the $De(0.5)$-law and $n_1 = n_2 = 20$ for different alternative hypotheses depending on the c- parameter value.

3.2 Simulation of Random Dependent Vector Variables

The simulation of pseudo-random normal vectors is based on the well-known generating algorithm [1]. Multivariate normally distributed random vector $\overline{X} = [X_1, X_2, \ldots, X_p]^T$ with p-dimension is completely determined by the vector

$\overline{M} = [\mu_1, \mu_2, \ldots, \mu_p]^T$ of mathematical expectations and covariance matrix $\Sigma = [\sigma_{ij}]$ where $i, j = \overline{1, p}$.

Suppose, there are a set of random values $\{Z_i\}$ and Z_i-value is distibuted according to the standard Normal Law where $i = \overline{1, p}$. Then the vector $\overline{X}$ distributed according to the multivariate Normal Law with the M and Σ parametes is determined by the linear transformation

$$\overline{X} = A\overline{Z} + M. \tag{2}$$

In (2) it is supposed that A is a lower triangular matrix.

Then coefficients a_{ij} and covariance matrix elements are easy determined by the recurrent procedure and (2)-d formula respectively:

$$\begin{aligned} a_{ij} &= \frac{\sigma_{ij} - \sum_{k=1}^{g-1} a_{ik}a_{jk}}{\sqrt{\sigma_{jj} - \sum_{k=1}^{g} a_{jk}^2}}, & (3) \\ \sigma_{ij} &= E\left[(X_i - \mu_i)(X_j - \mu_j)\right]. & (4) \end{aligned}$$

One-dimensional samples cosisted of normal random variables $\{Z_i\}$ are simulated by the method of inverse functions.

It is proposed to realise the multidimensional variable simulation procedure similarly to the algorithm described above (2)-(3), (4) [1]. The variables are distributed according to a law which is different from the Normal Law with a mathematical expectation vector and covariance matrix. It has been proved the correctness of this approach. It is also obtained conditions for the mathematical expectation vector and covariance matrix to get the situation when the first vector is linear depended on the second vector in the case of two dimensions. The similar conditions were obtained by authors of the article when the third variable is depended on others in the case of three dimensions. These results are shown below.

Vector $\overline{X}$ is distributed according to the multivariate Normal Law with the M and Σ parameters and determined by the (2)-d linear transformation. Coefficients for matrix A are calculated using the (3)-d formula. We can get three equations by substituting matrix A into the (2)-d formula:

$$\begin{aligned} X_1 &= \sqrt{\sigma_{11}} Z_1 + \mu_1, \\ X_2 &= \frac{\sigma_{12}}{\sqrt{\sigma_{11}}} Z_1 + \sqrt{\sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}} Z_2 + \mu_2, \\ X_3 &= \frac{\sigma_{31}}{\sqrt{\sigma_{11}}} Z_1 + \frac{\sigma_{32} - (\sigma_{31}\sigma_{21})/\sigma_{11}}{\sqrt{\sigma_{22} - \sigma_{21}^2/\sigma_{11}}} Z_2 + \\ &+ \sqrt{\sigma_{33} - \frac{\sigma_{31}^2}{\sigma_{11}} - \frac{(\sigma_{32} - (\sigma_{31}\sigma_{21})/\sigma_{11})^2}{\sigma_{22} - \sigma_{21}^2/\sigma_{11}}} Z_3 + \mu_3, \end{aligned}$$

where variables Z_i are distributed according to the standard Normal Law.

Just after that we introduce the following equations:

$$\begin{aligned}\sigma_{31} &= c_1\sigma_{11},\ \sigma_{32}=\sigma_{22},\ \sigma_{err}=\sqrt{\sigma_{33}-\sigma_{22}-c_1^2\sigma_{11}},\\ \mu_3 &= c_1\mu_1+\mu_2,\end{aligned}$$

and assume that X_1 and X_2 are the independent variables. The next formula

$$\sigma_{12}=\sigma_{21},$$

follows from the previous equations.

The next step is to group like terms in order to get an expression for calculating the linear dependent variable.

$$X_3=c_1\left(\sqrt{\sigma_{11}}z_1+\mu_1\right)+\mu_2+\sqrt{\sigma_{22}}Z_2+\sqrt{\sigma_{33}-c_1^2\sigma_{11}-\sigma_{22}}Z_3. \tag{5}$$

The (5)-th formula can be represented as

$$X_3=c_1X_1+X_2+X_{err}, \tag{6}$$

where the variable X_3 is dependent on the variables X_1 and X_2. X_1, X_2 and X_{err} are distributed according to the Normal Law with (M_1, σ_{11}), (M_2, σ_{22}) and (0, σ_{err}) parametes respectively, c_1 is a constant.

Thus, if it is requerid to simulate a three-dimensional sample which is like (5) where the X_3 is dependent on $X_1$and X_2 you should only define a mathematical expectation vector and a covariance matrix as

$$\mu=\begin{bmatrix}\mu_1\\ \mu_2\\ c_1\mu_1+\mu_2\end{bmatrix},\ \Sigma=\begin{bmatrix}\sigma_{11} & 0 & c_1\sigma_{11}\\ 0 & \sigma_{22} & \sigma_{22}\\ c_1\sigma_{11} & \sigma_{22} & \sigma_{33}\end{bmatrix}.$$

Since our studies is conducted under true null hypothesis so

$$\mu_1=\mu_2=0,\ \mu=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

Suppose, that

$$\sigma_{22}=k\sigma_{11},$$

in order to simulate different values of data heteroscedasticity rate.

It should be noted, that to link the elements of the covariance matrix with a tolerance value of a variable is another problem. It allows us to do investigation with the given tolerance value. The tolerance value is calculated using the formula: $1-R^2$.

We can obtain the R^2-value using the covariance matrix elements:

$$R^2=\frac{\sigma_{11}}{\sigma_{33}}\left(k+c_1^2\right)\to\sigma_{33}=\frac{\sigma_{11}}{R^2}\left(k+c_1^2\right).$$

It was decided to confine the problem and supposed that only variable X_3 was depend on variables $X_1$and X_2. It led to the next formula:

$$\operatorname{corr}(X_1,\, X_3) = \operatorname{corr}(X_2,\, X_3)\,. \tag{7}$$

It is clearly seen that the restriction

$$c_1 = \sqrt{k} \ and \ \sigma_{33} = \frac{2k\sigma_{11}}{R^2}$$

follows from the (7)-th formula.

3.3 The Results of Investigations in Case of Data Multicollinearity

For example, if $R^2 = 0.999$ the covariance matrix for the first and second groups looke like

$$\Sigma = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2.002 \end{pmatrix}.$$

Studies show that even if the tolerance value for one of a variable is 0.001 (i.e., $R^2 = 0,999$) the redundant variable presence in the model has no effect on the statistics distribution.

It was also studied the influence of different scale values on statistical results (i.e., different variance values correspond to different variables). These variance values differ from each other up to 250 000 times, that is

$$\sigma_{22} = 250\,000\sigma_{11},$$

for both groups. It means that 95.4% of observations for the first and the second variables under condition of data normality should belong to the $[-2,\, 2]$ and $[-1000,\, 1000]$ interval respectively. The empirical and theoretical statistics distributions do not differ from each other in both cases (i.e., all variable are independent or the third variable is dependent on others with $R^2 = 0.999$).

Conclusions

In conclusion, we can give some recommendations concerning the application of the lambda Wilks criterion. When determining the significance of the discriminant functions, you can use the statistical criterion correctly if observations in groups are distributed according to the Normal law or to the similar law with “light” tails. In contrast, if the distribution is very asymmetric or has “heavy” tails the application of the lambda Wilks criterion is not recommended as it can yield to erroneous results. This is most vividly seen with small sample sizes. The variables with the low tolerance value or variables measured on a different scale do not influence on the statistics distribution.

References

[1] Ermakov S.M., Mihailov G.A. (1982). *Statisticheskoe modelirovanie*. Nauka, Moscow.

[2] Karimov R.N. (2002). *Osnovy diskriminantnogo analiza [Fundamentals of discriminant analysis]*. SGTU, Saratov.

[3] Lemeshko B.Yu., Lemeshko S.B. (2011). *Statistical data analysis, simulation and study of probability regularities. Computer approach: monograph*. NSTU, Novosibirsk.

[4] Naresh K. Malhotra (2002). *Marketing Research: An Applied Orientation*. Williams, Moscow.

[5] Volkova V.M., Sanina A.A. (2014). Research of the lambda Wilks statistic distribution under conditions of violation of basic assumptions in the discriminant function analysis. *Actual problems of electronic instrument engineering (APEIE-2014)*. Vol. **1**, pp. 548-551.

A Comparison of the “Fixed-Effect” and “Random-Effect” Gamma Degradation Models[1]

Ekaterina V. Chimitova and Evgeniya S. Chetvertakova
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: chimitova@corp.nstu.ru, chetvertakova@corp.nstu.ru

Abstract

In this paper, the “fixed-effect” and “random effect” gamma degradation models are compared in terms of the accuracy of estimation of regression parameters and trend parameters. We then propose an algorithm for testing goodness-of-fit of the “random effect” gamma degradation model, which is based on the parametric bootstrap procedure and application of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling type tests by the sample of residuals.

Keywords: degradation process, gamma degradation model, “random-effect” model, “fixed-effect” model, testing goodness-of-fit, reliability.

Introduction

The problems of quality control and research of reliability of technical devices are very important nowadays, especially if a human life and health depend on their performance. If we have highly reliable products, the failure data may not be sufficient to assess the reliability function, because the failures occur extremely rare during the experiment. There are two possible ways to get additional information about the reliability of items: the first one is to carry out accelerated tests when items are under the high stress and as a result failure occurs earlier; the second method is to consider the degradation paths of items. The moment of time, when the degradation path reaches a critical level, is called failure time. Both approaches can be combined, observing degradation processes of the items and the failures under the high stress. Temperature, pressure, voltage, mechanical stresses and others may act as covariates. One of the most popular models describing a degradation process is the gamma degradation model, where the gamma distribution is used as the distribution of degradation increments. This model is described in [1], [5], [6], [8], in [9] the authors consider the problem of mis-specification of Wiener and gamma degradation processes.

In degradation modeling and analysis, “random-effect” formulation facilitates the handling of unit-to-unit variability in a convenient way. In [3], [10], the authors consider the gamma degradation model with random effects, where the scale parameter is a random variable from gamma distribution. In this paper, we compare the statistical properties of the maximum likelihood estimates for the parameters of the “fixed-effect” and “random-effect” gamma degradation models. Then, we formulate the parametric bootstrap procedure for testing goodness-of-fit of the “random-effect” gamma degradation model.

[1] This research has been supported by the Russian Ministry of Education and Science (project 2.541.2014K).

1 Gamma degradation model

Stochastic process $Z(t)$ characterizing degradation process is referred to as the gamma degradation process, if

- $Z(0) = 0$;
- $Z(t)$ is a stochastic process with independent increments;
- increments $\Delta Z(t) = Z(t + \Delta t) - Z(t)$ have the gamma distribution with probability density function
$$f_{Gamma}(t; \sigma; \Delta\nu(t)) = \left(\frac{t}{\sigma}\right)^{\Delta\nu(t)-1} \frac{e^{-t/\sigma}}{\sigma \cdot \Gamma(\Delta\nu(t))},$$
where $\Delta\nu(t) = \nu(t + \Delta t) - \nu(t)$ is the shape parameter and σ is the scale parameter.

If random variates ξ_1 and ξ_2 follow the gamma distribution with scale parameter σ and shape parameters ν_1 and ν_2, correspondingly, then $\xi_1 + \xi_2$ follows the gamma distribution with scale parameter σ and shape parameter $\nu_1 + \nu_2$. This property explains the fact of using the gamma distribution as a distribution of increments.

Let degradation process is observed under a constant in time stress (covariate) x, the range of values of which is defined by the conditions of the experiment. We assume, that the covariate influences the degradation as in the accelerated failure time model [7]: $Z_x(t) = Z\left(\frac{t}{r(x;\beta)}\right)$, where $r(x;\beta)$ is a positive covariate function. The most popular models of the covariate function are:

- log linear model – $r(x, \beta) = e^{\beta_0 + \beta_1 x}$;
- power rule model – $r(x, \beta) = e^{\beta_0 + \beta_1 \ln x}$;
- Arrhenius model – $r(x, \beta) = e^{\beta_0 + \beta_1 / x}$.

Let the mathematical expectation of degradation process $Z_x(t)$ is

$$M\left(Z_x\left(t\right)\right) = m_x\left(t\right),$$

where $m_x\left(t\right) = \sigma\nu\left(\frac{t}{r(x;\beta)}\right)$ is a positive increasing trend function and $r(x;\beta)$ is a positive covariate function.

The time to failure, which depends on covariate x, is equal to

$$T_x = \sup\{t : Z_x(t) < \tilde{z}\},$$

where $\tilde{z}$ is the critical value of the degradation path. Then, the reliability function for gamma degradation model is given by

$$S_x(t) = P\{T_x > t\} = P\{Z_x(t) < \tilde{z}\} = F_{Gamma}\left(\tilde{z}; \sigma, \frac{m_x(t)}{\sigma}\right).$$

In [3], [10], the unit-to-unit variability is included into the gamma degradation model by considering the scale parameter σ as a random variable from the gamma distribution with parameters δ and θ. In this case, the reliability function can be written as

$$S_x(t) = P\{T_x > t\} = P\{Z_x(t) < \tilde{z}\} = F_{Gamma}\left(\tilde{z}; M(\sigma), \frac{m_x(t)}{M(\sigma)}\right),$$

where $M(\sigma)$ is the mathematical expectation of σ.

Suppose, that we have the degradation path $Z^i(t)$ and covariate value x^i for n items, $i = \overline{1,n}$. The degradation path for i-th item is given by

$$(0, Z_0^i),\ (t_1^i, Z_1^i), ..., (t_{k_i}^i, Z_{k_i}^i), j = \overline{1, k_i},$$

where k_i is the number of moments of measuring degradation. Suppose, that the initial value of the degradation index $Z_0^i = 0$, $i = \overline{1,n}$.

Let us denote the sample of increments of the degradation path as

$$\mathbf{X}_n = \left\{X_j^i = Z_j^i - Z_{j-1}^i,\ i = \overline{1,n},\ j = \overline{1,k_i}\right\}.$$

Following the assumption, that the observed random processes $Z_{x^i}^i(t)$, $i = \overline{1,n}$ are the gamma degradation processes with the trend function $m_x(t;\beta,\gamma)$ and covariate function $r(x;\beta)$, we can estimate the model parameters, maximizing the logarithm of the likelihood function

$$L(\mathbf{X}_n) = \prod_{i=1}^{n}\prod_{j=1}^{k_i}\int_0^\infty \Gamma(X_{ij};\omega,\Delta\nu(t))\Gamma(\omega;\delta,\theta)d\omega =$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{k_i}\int_0^\infty X_{ij}^{\Delta\nu(t)-1}\frac{1}{\Gamma(\Delta\nu(t))\Gamma(\theta)}\frac{1}{\delta^{-\theta}}\omega^{\theta-1}\omega^{\Delta\nu(t)}e^{-X_{ij}\omega}e^{-\omega\delta}d\omega =$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{k_i} X_{ij}^{\Delta\nu(t)-1}\frac{1}{\Gamma(\Delta\nu(t))\Gamma(\theta)}\frac{1}{\delta^{-\theta}}\int_0^\infty \omega^{\Delta\nu(t)+\theta-1}e^{-(X_{ij}+\delta)\omega}d\omega =$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{k_i} X_{ij}^{\Delta\nu(t)-1}\frac{1}{\Gamma(\Delta\nu(t))\Gamma(\theta)}\frac{1}{\delta^{-\theta}}\frac{1}{(X_{ij}+\omega)^{\Delta\nu(t)+\theta}}\Gamma(\Delta\nu(t)+\theta) =$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{k_i}\frac{X_{ij}^{\Delta\nu(t)-1}}{(X_{ij}+\omega)^{\Delta\nu(t)+\theta}}B^{-1}(\Delta\nu(t);\theta).$$

It is natural, that the degradation processes are different for various units. Thus, the construction of the degradation model with random effects seems to be reasonable. However, the "random-effect" gamma degradation model is more complicated, and the dimension of parameters vector is larger. So, we need to understand whether taking into account the unit-to-unit variability allows to obtain more accurate estimates of the trend and regression parameters.

By means of Monte-Carlo simulations, we have compared the accuracy of estimates of the trend and regression parameters for "fixed-effect" gamma degradation model and the gamma degradation model with random effects, when data are generated from the "random-effect" model. The next plan of experiment has been used for simulation of the degradation path:

- scalar covariate x is equal to $x^1 = 1$ and $x^2 = 2$;
- items are randomly divided to 2 groups corresponding to 2 values of covariate;
- moments of measuring degradation for all items are equal to 10, 15, 25, 30.

The true values of parameters for both models were taken equal to $\sigma = 1$, $\beta_0 = 1$,$\beta_1 = 1$, $\delta = 0.1$, $\theta = 10$.

Let $\mu = (\gamma^T, \beta^T)$ denotes the vector of trend and regression parameters. We have compared the accuracy of estimation of parameter μ for considered models, calculating the Euclidean norm of relative deviation of estimates from the true value:

$$\psi = \|\frac{\mu - \hat{\mu}}{\mu}\|,$$

which was averaged by $M = 10000$ samples:

$$\bar{\psi} = \sum_{i=1}^{M} \psi_i, i = \overline{1, M}.$$

The obtained values of relative accuracy $\bar{\psi}$ of estimates of model parameters $\hat{\mu}$ in the case of the linear trend function

$$m_x(t) = \frac{t}{e^{\beta_0 + \beta_1 x}}$$

for the "fixed-effect" and "random-effect" gamma degradation models are presented in Table 1.

The obtained values of relative accuracy $\bar{\psi}$ of estimates of model parameters $\hat{\mu}$ in the case of the power trend function

$$m_x(t) = \left(\frac{t}{e^{\beta_1 + \beta_2 x}}\right)^{\gamma_0}$$

for the "fixed-effect" and "random-effect" gamma degradation models are presented in Table 2.

As can be seen from Tables 1 and 2, in the case of smaller sample sizes the better estimates have been obtained for "fixed-effect" gamma degradation model and in the case of the sample sizes $n = 200$ and $n = 500$, the better estimates have been obtained for "random-effect" gamma degradation model. Such results can be explained by the fact, that in the case of "fixed-effect" model we estimate only the scale parameter σ of the degradation increments distribution, trend and regression parameters, but in the case of "random-effect" model we estimate the scale and form parameters of the scale distribution δ and θ, and also trend and regression parameters.

Table 1: The relative accuracy of parameters estimates for gamma degradation models in the case of the linear trend function

Gamma degradation model	Sample size			
	20	60	200	500
"Fixed-effect" model	0.0336	0.0161	0.0039	0.0021
"Random-effect" model	0.0514	0.0129	0.0021	0.0016

Table 2: The relative accuracy of parameters estimates for gamma degradation models in case of the power trend function

Gamma degradation model	Sample size			
	20	60	200	500
"Fixed-effect" model	0.2999	0.1606	0.0581	0.0351
"Random-effect" model	0.3310	0.1814	0.0472	0.0288

Testing goodness-of-fit of the gamma degradation model

An important stage in the construction of the gamma degradation model is testing the goodness-of-fit hypothesis:

$$H_0 : X^i_j \sim F_{Gamma}(t; \hat{\sigma}, \hat{p}^i_j), i = \overline{1,n},\ j = \overline{1,k},$$

where $\hat{p}^i_j = \frac{m_{x^i}(t^i_j;\hat{\gamma},\hat{\beta}) - m_{x^i}(t^i_{j-1};\hat{\gamma},\hat{\beta})}{\hat{\sigma}}$.

The main problem of using the gamma degradation model is the absence of mathematical methods for testing the statistical hypothesis of goodness-of-fit for the model. In this paper, we propose an approach to testing goodness-of-fit of gamma degradation model with covariates, which is based on the investigation of test statistic distributions with computer simulation methods in interactive mode of testing hypothesis. The goodness-of-fit tests of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling type are recommended for testing this hypothesis [2].

Let us take the residuals of increments of the degradation path in the form:

$$R^i_j = F_{Gamma}(X^i_j; \hat{\sigma}, \hat{p}^i_j),\ i = \overline{1,n},\ j = \overline{1,k_i}.$$

If the hypothesis H_0 is correct, then

$$R^i_j \sim Rav(0,1), i = \overline{1,n},\ j = \overline{1,k_i}.$$

Thus, we need to test the uniform distribution for the random variates $R^i_j,\ i = \overline{1,n},\ j = \overline{1,k_i}$. To test this hypothesis we use the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling type tests [4]. Let $R^*_{(1)} \leq R^*_{(2)} \leq ... \leq R^*_{(N)}, N = \sum\limits_{i=1}^{n} k_i$

are the elements of variational series constructed by the sample

$$\mathbf{R}_N = \left\{R_j^i, i = \overline{1,n},\ j = \overline{1,k_i}\right\}.$$

The Bolshev statistic is used as a distance between empirical and theoretical distributions in Kolmogorov test:

$$S_k = \frac{6ND_N + 1}{6\sqrt{N}},$$

where $D_N = \max(D_N^+, D_N^-)$, $D_N^+ = \max\limits_{1\le i\le N}\left\{\frac{i}{N} - R_{(i)}^*\right\}$, $D_N^- = \max\limits_{1\le i\le N}\left\{R_{(i)}^* - \frac{i-1}{N}\right\}$.

The Cramer-von Mises-Smirnov statistic can be written as follows:

$$S_\omega = N\omega_N^2 = \frac{1}{12N} + \sum_{i=1}^{N}\left\{R_{(i)}^* - \frac{2i-1}{2N}\right\}^2,$$

and the Anderson-Darling statistic has the form:

$$S_\Omega = -N - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2N}\log R_{(i)}^* + \left(1 - \frac{2i-1}{2N}\right)\log\left(1 - R_{(i)}^*\right)\right\}.$$

When testing composite hypotheses, the distributions of test statistics depend on several factors: the method for parameter estimation, the type and the number of parameters being estimated and the specific value of the shape parameter (as in the case of the gamma distribution). Therefore, for calculation of the p-value, the distribution of test statistics under true null hypothesis can be simulated according to the following algorithm:

1. Generate the sample of increments $\mathbf{X}_n$ basing on the gamma degradation model under test and the appropriate plan of the experiment using the scale parameter, generated from the gamma distribution for each item.
2. Estimate model parameters with the sample $\mathbf{X}_n$ using maximum likelihood method.
3. Calculate the sample of residuals $\mathbf{R}_N$.
4. Calculate goodness-of-fit test statistics values (S_K, S_ω or S_Ω) for the sample $\mathbf{R}_N$.
5. Repeat points 1-4 M times, and obtain the empirical distribution $G_M(S|H_0)$.

Thus, we can calculate the p-value $\alpha_N = 1 - G_M(S_N|H_0)$, where S_N is a value of test statistics, calculated for the sample, which is used to test the hypothesis H_0. If α_N is less than the significance level α, then hypothesis H_0 is rejected.

Let us study the dependence of test statistic distributions on the choice of covariate function for degradation model with random effects. We have considered three different covariate functions: loglinear, Arrhenius and power rule models.

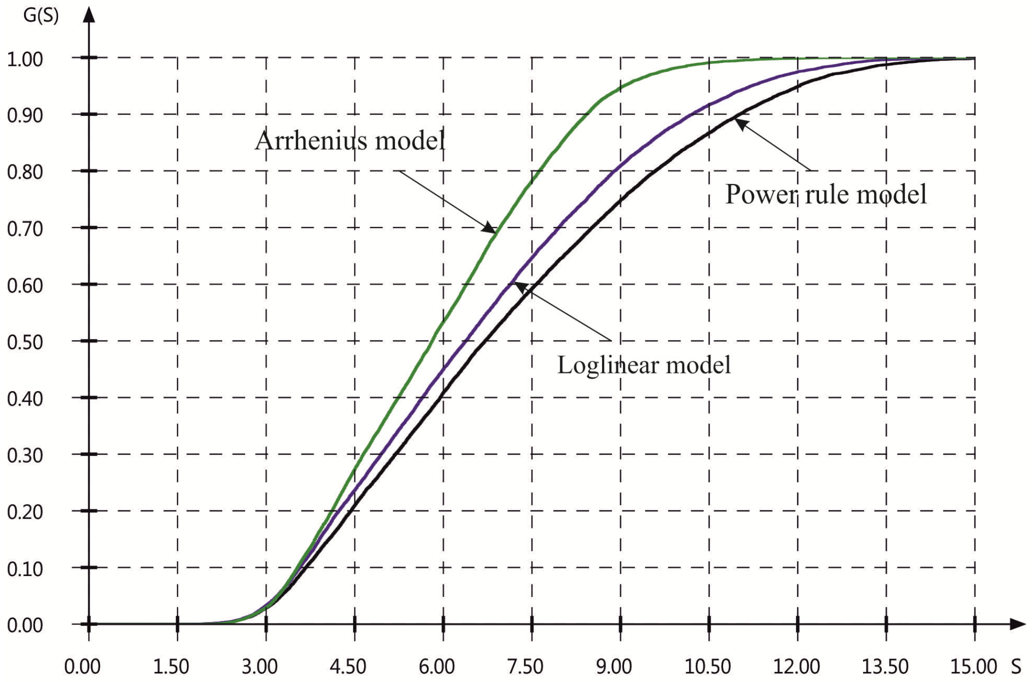


Figure 1: The empirical Kolmogorov distributions for "random-effect" gamma degradation model in case of loglinear, Arrhenius and power rule covariate functions.

In this study, we have used the same plan of experiment as in the previous section. Models parameters are equal to $\sigma = 1$, $\delta = 0.1$, $\theta = 10$. Regression parameters are equal $\beta_0 = 1$, $\beta_1 = 0.5$ for loglinear and Arrhenius models; $\beta_0 = 1$, $\beta_1 = -0.5$ for power rule model.

As can be seem from Figure 1, the form of the distribution of Kolmogorov statistic changes for different covariate functions. The same results were obtained for Cramer-von Mises-Smirnov and Anderson-Darling tests.

Conclusions

In this paper, we have considered the problems of constructing the gamma degradation model with random effects using the degradation data under the constant in time stress. The comparison of the statistical properties of maximum likelihood estimates of model parameters for the "fixed-effect" gamma degradation model and gamma degradation model with random effects has been carried out. It has been shown, that in the case of smaller sample sizes the more accurate estimates have been obtained for the "fixed-effect" gamma degradation model and in the case of the sample sizes $n = 200$ and $n = 500$, the better results have been obtained for the "random-effect" gamma degradation model. So, the "fixed-effect" gamma degradation model can be recommended to use for samples with a small size. Probably, the accuracy of the estimates for "random-effect" model will be improved for other values of the scale and form parameters of the scale distribution. The method of testing the goodness-of-fit for the gamma degradation model with random effects using Kolmogorov, Cramer von Mises-Smirnov and Anderson-Darling type tests has been proposed. While investigating, it has been shown that the distributions of considered test statistics depend on the trend and covariate functions, moments of measuring degradation and the sample size.

References

[1] Bordes, L., Paroissin, C. and Salami, A. (2010). Parametric inference in a perturbed gamma degradation process. *Preprint/Statistics and Probability Letters*, Vol. **13**.

[2] Chimitova, E., Chetvertakova, E. (2014). The construction of the gamma degradation model with covariates. *Tomsk State University Journal of Control and Computer Science*, Vol. **85**, pp. 51-60.

[3] Lawless, J., Crowder, M. (2004). Covariates and Random Effects in a Gamma Process Model with Application to Degradation and Failure. *Life Data Analysis*, Vol. **10**, pp. 213-227.

[4] Lemeshko, B.Yu., Lemeshko, S.B., Postovalov, S.N. and Chimitova E.V. (2011). *Statistical data analysis, simulation and study of probability regularities. Computer approach: monograph*, NSTU Publisher, Novosibirsk.

[5] Liao, C.M. and Tseng, S.-T. (2006). Optimal design for step-stress accelerated degradation test. *IEEE Trans. Reliab.*, Vol. **55**, pp. 59-66.

[6] Meeker, W.Q. and Escobar, L.A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.

[7] Nikulin, M. and Bagdonavicius, V. (2001). *Accelerated Life Models: Modeling and Statistical Analisys*, Chapman & Hall/CRC, Boca Raton.

[8] Tang, L.C., Yang, G.Y. and Xie M. (2004). Planning of step-stress accelerated degradation test. *Reliability and Maintainability Annual Symposium*, Los Angeles.

[9] Tsai, C.-C., Tseng, S.-T. and Balakrishnan, N. (2011). Mis-specification analyses of gamma and Wiener degradation processes. *Journal of Statistical Planning and Inference*, Vol. **141**, pp. 25-35.

[10] Tsai, C.-C., Tseng, S.-T. and Balakrishnan, N. (2012). Optimal Design for Degradation Tests Based on Gamma Processes With Random Effects. *IEEE Trans. Reliab.*, Vol. **61**, pp. 604-613.

On Survival Categorical Methods Based on an Extended Actuarial Estimator

Sergey V. Malov[1,2] and Stephen J. O'Brien[1]
[1] *Theodosius Dobzhansky Center for Genome Bioinformatics, St.-Petersburg State University, St.-Petersburg, Russia*
[2] *St.-Petersburg Electrotechnical University, St.-Petersbrg, Russia*
e-mail: `malovs@sm14820.spb.edu, lgdchief@gmail.com`

Abstract

We consider methods of categorical data analysis applicable for the survival experimental design. The actuarial (or life table) estimator for groupped right censored survival data that is consistent under very special cases only is not perfect in the estination problem, but it is applicable anyway, because there is no consistent estimator of survival function of failure time under general nonparametric model of groupped survival data with independent right censoring. We revisited the actuarial estimator for groupped right censored data and create survival categorical tests based on the extended actuarial estimator.

Keywords: survival data, right censoring, independent censoring, hypothesis testing, categorical tests, contrasts, Wald's test, actuarial estimator, Kaplan–Meier estimator.

Introduction

Common experimental design in epidemiology is to screen a cohort of individuals for disease endpoints during some time interval. Study participants are disease free at the baseline (time point zero) and they are followed up until a failure time or missing at follow up. In most cases failure times are not observed precisely and the investigator observes time interval containing a failure time for each of not missed at follow up individulals having sympthoms of desease at the endpoint. The case of fixed inspection times measured from the baseline is widely applicable. The goal is to quantify difference in rate of disease progression among a population of study participants.

Let T be a failure time or time of appearance symptoms of disease. Distribution of T depends on covariate z and can be given by a distribution function $F_z(x) = P(T \le x|z)$ or by a survival function $S_z(x) = 1 - F_z(x)$. Assume that the covariate z is a categorical variable having d levels. We are interesting to compare distributions of failure time under different values of covariate. Let $\gamma_T = \min_{i \in 1,\dots,d} \sup\{x : F_i(x) < 1\}$. The null hypothesis is

$$H_0^* : S_1(x) = \ldots = S_d(x) \quad \text{for all} \quad x \in [0, \gamma_T].$$

To formulate the problem in terms of categorical data analysis set $0 < t_1 < \ldots < t_s < \gamma_T$. Consider $p_{1|z} = P(T \in [0, t_j]|z)$ and $p_{j|z} = P(T \in]t_{j-1}, t_j]|z)$, $j = 2, \ldots, s+1$, where $t_{s+1} = \infty$. We formulate weaker null hypothesis

$$H_0 : p_{j|1} = \ldots = p_{j|d} \quad \text{for all} \quad j = 1, \ldots, s$$

or, in terms of the survival function,

$$H_0 : S_1(t_j) = \ldots = S_d(t_j) \quad \text{for all} \quad j = 1, \ldots, s. \tag{1}$$

It is clear that H_0 is closing to H_0^* if $p_{j|z} \to 0$ as $s \to \infty$.

A contingency table experimental design is universal for wide number of applications. There are examples of application of classical categorical tests in right censored data case [8, 13, 14]. Some limitations on application of classical categorical tests for right censored survival data are discussed in [11].

The right censored survival data model is commonly used for such kind of experimental design. Categorical tests using grouping for right censored survival data are presented widely in literature. Likelihood ratio tests with grouped right censored survival data were investigated in [15]. A chi-square type test for survival data due to Habib & Thomas [7]. Advanced properties of chi-square type tests are obtained in [1, 2]. Hollander & Pena [10] consider chi-square test statistic for simple null hypotheses in censored data case and its limit behaviour. Contrasts based categorical tests on independence for survival data obtained from limit theorems for Nelson–Aalen and Kaplan–Meier estimators are given in [11]. Exact event (failure or censoring) times are required for all these approaches.

In most cases event times are not observed exactly and an investigator is only observes a time interval between successive observations containing an event time for each of individuals. We consider the case of fixed observation times. There is no consistent nonparametric estimator for survival function of failure time even in the observation times by such kind of data. The Actuarial life table estimator was rather famous in early classic of survival analysis [3, 5, 6]. Breslow & Crowley [4] investigate conditions on distributions of failure and survival times to the actuarial estimator be consistent under independent censoring for any choice of fixed observation times and obtain asymptotic properties of the estimator.

We consider an extention of the actuarial estimator. In section 1 we introduce an extended actuarial estimator and investigate conditions to the extended actuarial estimator be consistent. Asymptotic properties of the extended actuarial estimator and categorical survival tests for groupped right censored survival data are discussed in section 2.

1 An Extended Actuarial Estimator

In this section we consider categorical methods applicable for interval censored data, when intervals of censoring coincides with the categorical bounds. First we consider right censored data under independent censoring. Let (T_i, U_i) be the independent pairs of independent failure and censoring times respectively; (X_i, δ_i), where $X_i = T_i \wedge U_i$ and $\delta_i = \mathbb{1}_{\{T_i \le U_i\}}$, $i = 1, \ldots, n$, be the observed data.

An alternative representation of the Kaplan-Meier estimator. Introduce the sequential times of the observed events $X_{(1)} < \ldots < X_{(m)}$. Note

$$D_i^* = \sum_{j=1}^{n} \mathbb{1}_{\{X_j = X_{(i)}\}}, \quad D_i^f = \sum_{j=1}^{n} \mathbb{1}_{\{X_j = X_{(i)}, T_j \le U_j\}}, \quad D_i^c = D_i^* - D_i^f,$$

is number of observations (failures and censoring, respectively) at time $X_{(i)}$, $i = 1,\dots,m$. The Kaplan–Meier estimator is given by

$$\hat{S}(X_{(k)}) = \prod_{i=1}^{k}\left(1 - \frac{D_i^f}{Y_i^*}\right).$$

The corresponding discrete distribution has atoms

$$\delta(X_{(k)}) = \hat{S}(X_{(k-1)}) - \hat{S}(X_{(k)}) = \prod_{i=1}^{k-1}\left(1 - \frac{D_i^f}{Y_i^*}\right) D_k^f/Y_k^* = \hat{S}(X_{(k-1)})\, D_k^f/Y_k^*,$$

at points $X_{(k)}$, $k = 1,\dots,m$. Using the equations

$$1 - \frac{D_i^f}{Y_i^*} = \frac{Y_i^* - D_i^f}{Y_i^*} = \frac{Y_i^* - D_i^*}{Y_i^*}\frac{Y_i - D_i^f}{Y_i^* - D_i^*} = \frac{Y_{i+1}^*}{Y_i^*}\left(1 + \frac{D_i^* - D_i^f}{Y_{i+1}^*}\right),$$

one can write that

$$\hat{S}(X_{(k)}) = \frac{Y_{k+1}^*}{n}\prod_{i=1}^{k}\left(1 + \frac{D_i^c}{Y_{i+1}^*}\right)$$

and

$$\delta(X_{(k)}) = \frac{D_k^f}{n}\prod_{i=1}^{k-1}\left(1 + \frac{D_i^c}{Y_{i+1}^*}\right).$$

It is the baseline formula for alternative sequential method to construct the Kaplan-Meier estimator. We start from the empirical distribution by $\{X_1,\dots,X_n\}$ with $\delta(X_{(i)}) = D_i^*/n$, $i = 1,\dots,m$. Then, for any i in order to increase the values $X_{(i)}$ one use the following procedure. If $T_k \le U_k$ and $X_k = X_{(i)}$ then one doesn't change $\delta(X_{(i)})$. Otherwise, if $T_k > U_k$ and $X_k = X_{(i)}$ then one distribute the corresponding mass between all $X_{(j)}$, $j > i$, in proportion of number of failures having values $X_{(j)}$.

The actuarial estimator. Let $0 = x_0 < x_1 < \dots < x_r < \gamma_T$ be some fixed timepoints (observation times); $I_1 = [0, x_1]$ and $I_j = (x_{j-1}, x_j]$. The survival function in the fixed timepoints takes the following values:

$$S(x_k) = \prod_{i=1}^{k}(1 - \lambda_i),$$

where $\lambda_i = (S(x_{i-1}) - S(x_i))/S(x_{i-1})$. The observed data are given by δ_i and $\kappa_{ij} = \mathbb{1}_{\{X_i \in I_j\}}$, $j = 1,\dots,r$, $i = 1,\dots,n$. The Kaplan–Meier estimator is not applicable in this case. Introduce the following notations:

$$D_{1k} = \sum_{i=1}^{n}\mathbb{1}_{\{X_i \in I_k, U_i > a_k, \delta_i = 1\}}, \quad D_{2k} = \sum_{i=1}^{n}\mathbb{1}_{\{X_i \in I_k, U_i \in I_k, \delta_i = 1\}}, \quad W_k = \sum_{i=1}^{n}\mathbb{1}_{\{X_i \in I_k, \delta_i = 0\}},$$

$$Y_{1k} = \sum_{i=1}^{n}\mathbb{1}_{\{X_i > x_{k-1}, U_i > x_k\}}, \quad Y_{2k} = \sum_{i=1}^{n}\mathbb{1}_{\{X_i > x_{k-1}, U_i \in I_k\}},$$

$$Y_k = Y_{1k} + Y_{2k}, \quad D_k = D_{1k} + D_{2k}.$$

The actuarial estimator of λ_k discussed in [4] is

$$\hat{\lambda}_k = D_k/(Y_k - W_k/2).$$

It was proved that the actuarial estimator is consistent for any $x_1, \ldots, x_r$: $x_r \leq M < \gamma_T$ iff

$$S(x) = 1/(1 + c\,G(x))^{1/2}, \text{ for all } x \in [0, M],$$

where c is a constant. We consider the following extension of the actuarial estimator:

$$\hat{\lambda}_k(a) = D_k/(Y_k - aW_k), \text{ for some } a \in [0,1], \ k = 1, \ldots, r.$$

By using the alternative representation of Kaplan–Meier estimator it is easy to see that the case $a = 0$ ($a = 1$) is corresponding to $T_i < U_j$ ($T_i > U_j$) for all $i : \delta_i = 1$, $j : \delta_j = 0$ in terms of the Kaplan–Meier estimator, $X_i \in I_k$ and $X_j \in I_k$, and the case $a = 1/2$ is corresponding to the classical actuarial estimator. On the analogy of [4] we note that

$$\hat{\lambda}_k(a) = \frac{Y_{1k}}{Y_k - aW_k} \cdot \frac{D_{1k}}{N_{ik}} + \frac{D_{2k}}{Y_k - aW_k}$$

Because D_{lk}, Y_{lk} and W_k are sums of independent and identically distributed random variables $\hat{\lambda}_k(a) \to \lambda_k(a)$ a.s. and

$$\lambda_k(a) = \frac{\mathbb{P}(T > x_{k-1}, U > x_k)}{\mathbb{P}(X > x_{k-1}) - a\mathbb{P}(U \in I_k, T > U)}\lambda_k + \frac{\mathbb{P}(T \in I_k, U \in I_k, T \leq U)}{\mathbb{P}(X > x_{k-1}) - a\mathbb{P}(U \in I_k, T > U)}$$

$$= \frac{(1 - G(x_k))}{(1 - G(x_{k-1})) - a\int_{x_{k-1}}^{x_k} S_k^*\, dG}\lambda_k + \frac{\int_{x_{k-1}}^{x_k} F_k^*\, dG}{(1 - G(x_{k-1})) - a\int_{x_{k-1}}^{x_k} S_k^*\, dG}$$

$$= \frac{1 - G_k^*(x_k)}{1 - a\int_{x_{k-1}}^{x_k} S_k^*\, dG_k^*}\lambda_k + \frac{\int_{x_{k-1}}^{x_k} F_k^*\, dG_k^*}{1 - a\int_{x_{k-1}}^{x_k} S_k^*\, dG_k^*},$$

where $S_k^*(x) = S(x)/S(x_{k-1})$, $x \geq x_{k-1}$, and $F^* \equiv 1 - S^*$ are survival and distribution functions of the truncated distribution of T respectively and G^* is the truncated distribution function of U. Then the corresponding bias of the estimator is given by

$$b_k = \lambda_k(a) - \lambda_k = \frac{a\int_{x_{k-1}}^{x_k} S_k^*\, dG_k^* - G_k^*(x_k)}{1 - a\int_{x_{k-1}}^{x_k} S_k^*\, dG_k^*}F_k^*(x_k) + \frac{\int_{x_{k-1}}^{x_k} F_k^*\, dG_k^*}{1 - a\int_{x_{k-1}}^{x_k} S_k^*\, dG_k^*} \tag{2}$$

Under fixed F and G the estimator $\hat{\lambda}_k(a)$ is consistent for λ_k if

$$a = \frac{G_k^*(x_k) - F_k^*(x_k)^{-1}\int_{x_{k-1}}^{x_k} F_k^*\, dG_k^*}{\int_{x_{k-1}}^{x_k} S_k^*\, dG_k^*}. \tag{3}$$

2 Categorical survival tests

Let $0 < t_1 < \ldots < t_s < \gamma_T$ are such that $S_i(t_{k_1}) - S_i(t_{k_2}) > 0$ for all $1 \le k_2 < k_1 < \gamma_T$. We assume for simplicity that the set of breakpoints $\{t_1, \ldots, t_s\}$ coincides with the set of observation times $\{x_1, \ldots, x_s\}$. Generalization of tests below to the case $\{t_1, \ldots, t_s\} \subseteq \{x_1, \ldots, x_{s'}\}$, $s \le s'$, is trivial. Note

$$p_k^D = \mathbb{P}(X \in I_k, \delta = 1) = \int_{x_{i-1}}^{x_i} (1-G)\, dF;$$

$$p_k^Y = \mathbb{P}(X > x_{k-1}) = S(x_k)(1 - G(x_k));$$

$$p_k^W = \mathbb{P}(X \in I_k, \delta = 0) = \int_{x_{k-1}}^{x_k} S\, dG.$$

The asymptotic normality of the actuarial estimator $(\hat\lambda_1(a_1), \ldots, \hat\lambda_r(a_r))$ under $a_1 = \ldots = a_r = 1/2$ was given in [4] and the asymptotic normality for an arbitrary $a_1, \ldots, a_r \in [0,1]$ is following in the same way

$$\sqrt{n}(\hat\lambda_k(a_k) - \lambda_k(a_k)) \Rightarrow N(0, \sigma_k^2) \tag{4}$$

where $\sigma_k^2 = \lambda_k(a_k)/(p_k^Y - a_k p_k^W) - (p_k^Y - a_k^2 p_k^W)\lambda_k(a_k)^2/(p_k^Y - a_k p_k^W)^2$, and $\hat\lambda_k(a)$ are asymptotically independent, $k = 1, \ldots, r$. A consistent estimator $\hat\sigma_k^2$ of σ_k^2 can be obtained by using the actuarial estimator $\hat\lambda_k(a_k)$ and the empirical estimators $\hat p_k^Y$ and $\hat p_k^W$ of p_k^Y and p_k^W respectively.

Introduce the parameters $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{is})'$, where $\theta_{iz} = S_z(t_i; \boldsymbol{a}) = \prod_{l: x_l \le t_i}(1 - \lambda_{l|z}(a_l))$, $\lambda_{l|z}(a)$ is the limit mean of the actuarial estimator discussed in section 1 under fixed z, $\boldsymbol{a} = (a_1, \ldots, a_r)$ and $a_i = a_i(z) \in [0,1]$ can be different under different levels of z. Tests based on the actuarial estimators require the null hypothesis

$$\widetilde{H}_0 : \theta_{j1} = \ldots = \theta_{jd} \quad \text{for all} \quad j = 1, \ldots, s,$$

that is different in general of the categorical null hypothesis H_0 in (1), but its rejection implies rejection of the nonparametric homogeneity hypothesis H_0^* if the right censoring distribution defined by G is not dependent on z. Note that (4) implies convergence

$$\sqrt{n_i}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \Rightarrow N(0, \boldsymbol{\Sigma}_i)$$

with $\Sigma_i = \|\sigma_{i:jj'}\|$ and $\sigma_{i:jj'} = \theta_{ij}\theta_{ij'} \sum_{k=1}^{j} \sigma_k^2/(1 - \lambda_{k|i}(a_k))$ if $j \le j'$. The consistent estimator of Σ_i under $n_i \to \infty$ can be obtained by using the actuarial estimator $\hat\lambda_{k|i}(a_k)$ and $\hat\sigma_{k|i}^2$, $k = 1, \ldots, s$, $i = 1, \ldots, d$. Introduce $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1s}, \ldots, \theta_{d1}, \ldots, \theta_{ds})'$ and the corresponding estimator $\hat{\boldsymbol{\theta}} = (\hat\theta_{11}, \ldots, \hat\theta_{1s}, \ldots, \hat\theta_{d1}, \ldots, \hat\theta_{ds})'$. Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \Rightarrow N(0, \boldsymbol{\Sigma}), \tag{5}$$

and $\boldsymbol{\Sigma} = \mathrm{diag}(l_1\boldsymbol{\Sigma}_1, \ldots, l_n\boldsymbol{\Sigma}_d)$ is the block-diagonal matrix, $l_i = n/n_i$ for $i = 1, \ldots, d$.

Let $\boldsymbol{\psi}_i = \mathbf{A}\boldsymbol{\theta}_i$, where $\mathbf{A} = \|a_{ij}\|$ is $(d-1)\times d$ -matrix of linearly independent contrasts, i. e. $\sum_{j=1}^{d} a_{ij} = 0$ for all i and $\mathbf{rk}(\mathbf{A}) = d-1$. Then, $\widetilde{H}_0$ can be written in terms of contrasts

$$\widetilde{H}_0 : \boldsymbol{\psi}_1 = \ldots = \boldsymbol{\psi}_{d-1} = 0.$$

Associate with any a_{ij} the diagonal matrix $\mathbf{A}_{ij} = a_{ij}\mathbf{I}_s$, where $\mathbf{I}_s$ is the identity matrix of size s and construct the matrix $\mathbf{B}$ of size $(d-1)s\times ds$ from blocks $\mathbf{A}_{ij}$ in appropriate order. It is obviously that $\mathbf{B}$ is a matrix of linearly independent contrasts and the null hypothesis can be rewritten in vector form

$$\widetilde{H}_0 : \mathbf{B}\boldsymbol{\theta} = 0.$$

Taking into account (5) we obtain that under null hypothesis

$$n\,\hat{\boldsymbol{\theta}}'\hat{\mathbf{Q}}^{-1}\hat{\boldsymbol{\theta}} \Rightarrow \chi^2_{(d-1)s},$$

where $\hat{\mathbf{Q}} = \mathbf{B}'(\mathbf{B}\widehat{\boldsymbol{\Sigma}}\mathbf{B}')^{-1}\mathbf{B}$.

Analogous tests for

$$\widetilde{H}_0^* : \lambda_{j|1}(a_j) = \ldots = \lambda_{j|d}(a_j) \quad \text{for all} \quad j = 1,\ldots,s,$$

can be obtained directly from (4).

Acknowledgements

This research was supported in part by St.-Petersburg State University (grant No. 1.37.165.2014) and Russian Ministry of Education and Science Mega-grant No. 11.G34.31.0068 (Dr. Stephen J. O'Brien, Principal Investigator).

References

[1] Bagdonavičius, V., Levuliene, R, Nikulin, M. S. & Tran, Q. X. (2012). On Chi-square Type Tests and Their Applications in Survival Analysis and Reliability. *Zapiski nauchnih seminarov POMI* **408**, 43–61.

[2] Bagdonavičius, V. & Nikulin, M. S. (2011). Chi-squared Goodness-of-fit Test for Right Censored Data. *International Journal of Applied Mathematics and Statistics* **24**, 30–50.

[3] Berkson and Gage (1950). Calculation of survival rates for cancer. *Proc.Mayo Clinic*, Vol. **25**, 270–286.

[4] Breslow, N. and Crowley, J. (1974) A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *The Annals of Statistics*, Vol. **2**(3), 437–453.

[5] Cutler and Ederer (1958). Maximum utilization of the life table method in analyzing survival. *J. Chron.Dis.* **8**, 699–712.

[6] Gehan (1969). Estimating survival function from the life table. *J.Chron.Dis.*, Vol. **21**, 629–644.

[7] Habib & Thomas (1986). Chi-Square Goodness-if-Fit Tests for Randomly Censored Data. *The Annals of Statistics*, Vol. **14**(2), 759–765.

[8] Hendrickson, S.L., Lautenberger, J.A., Chinn, L.W., Malasky, M., Sezgin, E., Kingsley, L.A., Goedert, J.J., Kirk, G.D., Gomperts, E.D., Buchbinder, S.P., Troyer, J.L. and O'Brien, S.J. (2010). Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression *PLoS ONE*, Vol. **5**(9), art. no. e12862, pp. 1-8

[9] Hjort, N.L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates. *The Annals of Statistics*, Vol. **18**, 1221–1258.

[10] Hollander & Pena (1992). A Chi-Squared Goodness-of-Fit Test for Randomly Censored Data. *Journal of the American Statistical Association*, Vol. **87**(418), 458-463.

[11] Malov S.V. & O'Brien S.J. (2013). On Survival Categorical Methods with Applications in Epidemiology and AIDS Research. In *Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control.* Proceedings of the International Workshop AMSA'13 (Novosibirsk, September 25-27, 2013), 173–180.

[12] O'Brien, S.J., Hendrickson, S.L. (2013). Host genomic influences on HIV/AIDS. *Genome Biology*, Vol. **14**: 201.

[13] O'Brien, S.J., Nelson, G.W., Winkler, C.A., Smith, M.W. (2000). Polygenic and multifactorial disease gene association in man: Lessons from AIDS. *Annual Review of Genetics*, Vol. **34**, 563–591.

[14] Shin, H.D., Winkler, C., Stephens, J.C., Bream, J., Young, H., Goedert, J.J., O'Brien, T.R., Buchbinder, S.f, Giorgi, J.h, Rinaldo, C.i, Donfield, S.g, Willoughby, A.j, O'Brien, S.J. , and Smith, M.W. (2000) Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. **97**(26), 14467-14472.

[15] Turnbull & Weiss (1978). A Likelihood Ratio Statistic for Testing Goodness of Fit with Randomly Censored Data. *Biometrics*, Vol. **34**(3), 367–375.

Estimation of Survival Function in Partially Informative Random Censoring Model in the Presence of a Covariate

Abdushukurov A.A.
National University of Uzbekistan, Tashkent, Uzbekistan
e-mail: a_abdushukurov@rambler.ru

Abstract

In paper we consider a specific dependent informative competing risks model in the presence of a covariate and propose three types of estimators for survival function. We show uniform strong consistency and weak convergence of estimators to the same Gaussian process.

Keywords: Competing risks, informative censoring, exponential-hazard, product-limit, power-type estimators.

Introduction

Let $\{(X_i, Y_{1i}, Y_{2i}),\ i \geq 1\}$ be a sequence of independent and identically distributed triples of positive random variables (r.v.-s), where the components X_i, Y_{1i} and Y_{2i} are supposed to be conditionally independent given a covariate Z_i. The X_i 's are lifetimes with a common continuous distribution function (d.f.) $F(t),\ t \in \mathbb{R}^+$.The Y_{ki} 's, $k = 1, 2$, are censoring times with common continuous d.f.-s $G_k(t),\ k = 1, 2,\ t \in \mathbb{R}^+$, respectively. At the n-th stage of experiments the observed data consists a sample of triples $\{(\xi_i, \delta_i, Z_i),\ i = 1, ..., n\} = C^{(n)}$ with $\xi_i = \min(X_i, Y_{1i}, Y_{2i})$ and

$$\delta_i = \begin{cases} 1, & \text{if} \quad X_i \leq \min(Y_{1i}, Y_{2i}), \\ 0, & \text{if} \quad Y_{1i} \leq \min(X_i, Y_{2i}), \\ -1, & \text{if} \quad Y_{2i} \leq \min(X_i, Y_{1i}). \end{cases}$$

In sample $C^{(n)}$ the r.v.-s of interest X_i 's are censored from the right by r.v.-s $\min(Y_{1i}, Y_{2i})$ and observable partially only in case of $\delta_i = 1$. The estimation of d.f. F and its functionals from sample $C^{(n)}$ is one of the main goals in survival analysis. Let's define the conditional d.f.-s of r.v.-s X_i and $Y_{ki},\ k = 1, 2$, given a covariate $Z_i = z$ as

$$F(t/z) = P(X_i \leq t/Z_i = z), \qquad (t, z) \in \mathbb{R}^+ \times \mathbb{R},$$

$$G_k(t/z) = P(Y_{ki} \leq t/Z_i = z),\ k = 1, 2, \quad (t, z) \in \mathbb{R}^+ \times \mathbb{R}.$$

We also suppose that the censoring by r.v.-s Y_{1i} for a given covariate is informative, i.e. the pairs (X_i, Y_{1i}) follows the conditionally proportional hazards model (PHM) in which the d.f.$G_1(t/z)$ is related to $F(t/z)$ as

$$1 - G_1(t/z) = (1 - F(t/z))^\beta, \quad (t, z) \in \mathbb{R}^+ \times \mathbb{R}. \tag{1.1}$$

Here β is a some fixed but unknown censoring parameter. This kind of partially informative random censoring model with nuisance parameter (β, G_2) in lack of covariate Z_i was considered by authors [1,4,9,11]. Adapting some of ideas from [4] here we propose three asymptotical equivalent estimators of $F(t)$ through estimation of conditional d.f. $F(t/z)$ by exponential-hazard, product-limit and relative-risk power estimators using data from sample $C^{(n)}$.

1 Estimators of Survival Function

Let $H(t/z)$ is a conditional d.f. of ξ_i. Then by supposed independence of r.v.-s for a given covariate and from (1.1) we have $1 - H(t/z) = (1 - K(t/z))(1 - G_2(t/z))$, where $K(t/z) = P(\min(X_i, Y_{1i}) \leq t/Z_i = z) = 1 - (1 - F(t/z))(1 - G_1(t/z)) = 1 - (1 - F(t/z))^{\beta+1}$, $t \in \mathbb{R}^+$, $z \in \mathbb{R}$. For any d.f. $L(t)$, let

$$\tau_L = \sup\left\{t \in \mathbb{R}^+ : L(t) = 0\right\}, \quad T_L = \inf\left\{t \in \mathbb{R}^+ : L(t) = 1\right\},$$
$$L(t-) = \lim_{s \uparrow t} L(s), \quad \Delta L(t) = L(t) - L(t-).$$

Then by (1.1), $\tau_F = \tau_{G_1} = \tau_K$, $T_F = T_{G_1} = T_K$and $\tau_H = \max(\tau_F, \tau_{G_2}) \geq 0$,$T_H = \min(T_F, T_{G_2}) \leq \infty$.Let $Q(z)$ is d.f. of r.v.-s Z_i. Then

$$F(t) = \int F(t/z) dQ(z), \quad G_k(t) = \int G_k(t/z)\, dQ(z), \quad k = 1, 2,$$
$$K(t) = \int K(t/z)\, dQ(z) = P(\min(X_i, Y_{1i}) \leq t),$$
$$H(t) = \int H(t/z) dQ(z) = P(\xi_i \leq t).$$

In order to constructing the estimators of $F(t)$, we need the following conditional subdistribution functions:

$$\tilde{H}(t/z) = P(\xi_i \leq t, \delta_i \neq -1/Z_i = z) = P(\min(X_i, Y_{1i}) \leq$$
$$\leq \min(t, Y_{2i})/Z_i = z) = \int\limits_0^t (1 - G_2(u/z))\, dK(u/z),$$
$$\tilde{\tilde{H}}(t/z) = P(\xi_i \leq t, \delta_i = -1/Z_i = z) = P(Y_{2i} \leq \min(t, X_i, Y_{1i}) \,/\, Z_i = z) =$$
$$= \int\limits_0^t (1 - K(u/z))\, dG_2(u/z),$$

with $\tilde{H}(t/z) + \tilde{\tilde{H}}(t/z) = H(t/z)$ for all $t \in \mathbb{R}^+$, $z \in \mathbb{R}$. Let $\gamma = \frac{1}{\beta+1}$ and $p_m = \mathbb{P}(\delta_i = m)$, $m = -1, 0, 1$. Then

$$\mathbb{P}(\delta_i \neq -1) = \int \lim_{t \to \infty} \tilde{H}(t/z)\, dQ(z) = \int \int\limits_0^\infty (1 - G_2(u/z)) dK(u/z) dQ(z) =$$

$$= \int \left(\int_0^\infty (1 - G_2(u/z))\, d\left[1 - (1 - F(u/z))^{\beta+1} \right] \right) dQ(z) =$$

$$= \frac{1}{\gamma} \int \left(\int_0^\infty (1 - G_2(u/z))(1 - F(u/z))^\beta dF(u/z) \right) dQ(z), \tag{2.1}$$

and

$$\mathbb{P}(\delta_i = 1) = \int \mathbb{P}(\delta_i = 1/Z_i = z)\, dQ(z) = \int \mathbb{P}(X_i \le \min(Y_{1i}, Y_{2i})/Z_i = z)\, dQ(z) =$$

$$= \int \left(\int_0^\infty (1 - F(u/z))^\beta (1 - G_2(u/z))\, dF(u/z) \right) dQ(z). \tag{2.2}$$

From (2.1) and (2.2), we get $\gamma = \mathbb{P}(\delta_i = 1)/\mathbb{P}(\delta_i \neq -1) = \mathbb{P}(\delta_i = 1/\delta_i \neq -1)$. Hence the parameter $\gamma = \frac{1}{\beta+1}$ can be consistently estimated by statistics

$$\gamma_n = \frac{\sum_{i=1}^n I(\delta_i = 1)}{\sum_{i=1}^n I(\delta_i \neq -1)} = \frac{p_{1n}}{p_{0n} + p_{1n}}, \tag{2.3}$$

where $p_{mn} = \frac{1}{n}\sum_{i=1}^n I(\delta_i = m)$ are estimators of probabilities $p_m, m = -1, 0, 1$. Let's define cumulative hazard functions (c.h.f.-s)

$$\tilde{\Lambda}(t/z) = -\log(1 - K(t/z)) = -\frac{1}{\gamma} \cdot \log(1 - F(t/z)),$$

$$\tilde{\tilde{\Lambda}}(t/z) = -\log(1 - G_2(t/z)), \tag{2.4}$$

and

$$\Lambda(t/z) = -\log(1 - H(t/z)) = \tilde{\Lambda}(t/z) + \tilde{\tilde{\Lambda}}(t/z).$$

We suppose that d.f.$Q(z)$ have a density $q(z)$. Then c.h.f.-s (2.4) can be represented as

$$\tilde{\Lambda}(t/z) = \int_0^t \frac{q(z)\, d\tilde{H}(u/z)}{q(z)(1 - H(u/z))} = \int_0^t \frac{d\tilde{A}(u;z)}{B(u;z)},$$

$$\tilde{\tilde{\Lambda}}(t/z) = \int_0^t \frac{q(z)\, d\tilde{\tilde{H}}(u/z)}{q(z)(1 - H(u/z))} = \int_0^t \frac{d\tilde{\tilde{A}}(u;z)}{B(u;z)}, \tag{2.5}$$

and

$$\Lambda(t/z) = \int_0^t \frac{q(z)\, dH(u/z)}{q(z)(1 - H(u/z))} = \int_0^t \frac{dA(u;z)}{B(u;z)},$$

where

$$A(u;z) = \tilde{A}(u;z) + \tilde{\tilde{A}}(u;z), \quad \tilde{A}(u;z) = q(z)(1 - G_2(u/z)),$$

$$\tilde{\tilde{A}}(u;z) = q(z)(1 - K(u/z)) \quad \text{and} \quad B(u/z) = q(z)(1 - H(u/z)).$$

These functions can be estimated by statistics

$$A_n(u;z) = \tilde{A}_n(u;z) + \tilde{\tilde{A}}_n(u;z),$$

$$\tilde{A}_n(u;z) = \frac{1}{n}\sum_{i=1}^{n} I(\xi_i \le u,\, \delta_i \ne -1)\ \pi_a(z;Z_i),$$

$$\tilde{\tilde{A}}_n(u;z) = \frac{1}{n}\sum_{i=1}^{n} I(\xi_i \le u,\ \delta_i = -1)\ \pi_a(z;Z_i), \tag{2.6}$$

and

$$B_n(u;z) = \frac{1}{n}\sum_{i=1}^{n} I(\xi_i \ge u)\ \pi_a(z,Z_i),$$

where $\pi_a(z;t) = \frac{1}{a_n}\pi\left(\frac{z-t}{a_n}\right)$ with kernel function $\pi(z)$ and bandwidth sequence $a = a_n \downarrow 0$ as $n \to \infty$. By substitution of estimators (2.6) into formulas (2.5) we obtain the estimators for c.h.f.-s

$$\tilde{\Lambda}_n(t/z) = \int_0^t \frac{d\tilde{A}_n(u;z)}{B_n(u;z)}, \qquad \tilde{\tilde{\Lambda}}_n(t/z) = \int_0^t \frac{d\tilde{\tilde{A}}_n(u;z)}{B_n(u;z)},$$

and

$$\Lambda_n(t/z) = \tilde{\Lambda}_n(t/z) + \tilde{\tilde{\Lambda}}_n(t/z) = \int_0^t \frac{d\,A_n(u;z)}{B_n(u;z)}. \tag{2.7}$$

In order to estimate the conditional d.f. $F(t/z)$ we use representation

$$1 - F(t/z) = (1 - K(t/z))^{\gamma}, \quad (t,z) \in \mathbb{R}^{+} \times \mathbb{R}, \tag{2.8}$$

following from (1.1). For $1 - K(t/z)$ we use the following exponential hazard type estimator of Altschuler-Breslow, product-limit type estimator of Kaplan-Meier and relative-risk power type estimator of Abdushukurov (see, [1-6]):

$$1 - K_{1n}(t/z) = \exp\left(-\tilde{\Lambda}_n(t/z)\right),$$

$$1 - K_{2n}(t/z) = \prod_{u \le t}\left(1 - \Delta\tilde{\Lambda}_n(u/z)\right), \tag{2.9}$$

$$1 - K_{3n}(t/z) = \left[\prod_{u \le t}(1 - \Delta\Lambda_n(u/z))\right]^{R_n(t;z)},$$

where $R_n(t;z) = \tilde{\Lambda}_n(t/z)(\Lambda_n(t/z))^{-1}$ is estimator of $R(t;z) = \tilde{\Lambda}(t/z)(\Lambda(t/z))^{-1}$, $\Delta\tilde{\Lambda}_n(u/z) = \tilde{\Lambda}_n(u/z) - \tilde{\Lambda}_n(u-/z)$,$\Delta\Lambda_n(u/z) = \Lambda_n(u/z) - \Lambda_n(u-/z)$. According to (2.8) using estimators (2.3) and (2.9) we get corresponding estimators of $1-F(t/z)$ as

$$1-F_{ln}(t/z)=(1-K_{ln}(t/z))^{\gamma_n},\ \ l=1,2,3,\ \ (t,z)\in\mathbb{R}^+\times\mathbb{R}. \tag{2.10}$$

Finally, using statistics (2.10) we construct estimators of $1-F(t)$ by averaging as follows:

$$1-F_{ln}(t)=\int(1-F_{ln}(t/z))dQ_n(z)\,,\ \ l=1,2,3,\ t\in\mathbb{R}^+, \tag{2.11}$$

where

$$Q_n(z)=\frac{1}{n}\sum_{i=1}^{n}I(Z_i\leq z)\,,\ \ z\in\mathbb{R},$$

is the empirical estimator of d.f. $Q(z)$. Note that in lack of censoring by r.v.-s Y_{1i}-s (i.e. $G_1(t)\equiv 0$) the estimator $K_{1n}(t/z)$ in (2.9) coincides with one considered in [11].

2 Asymptotic properties of estimators of survival function

In order to investigate the asymptotic properties of estimators (2.11) we need the following conditions.

Conditions I:

(I.1) The kernel function π is bounded and Lipschitz continuous of order 1 with respect to the Euclidean distance on $\mathbb{R}$.

(I.2) $\int\pi(z)\,dz=1,\ \int z\pi(z)\,dz=0,\ \int z^2\,|\pi(z)|dz<\infty$.

(I.3) The bandwith sequence $\{a_n,n\geq 1\}$ satisfies: $a_n\to 0$ and $\frac{\log n}{na_n}\to 0$ as $n\to\infty$.

(I.4) The partial derivatives $\frac{\partial F(t/z)}{\partial t}$ and $\frac{\partial G_2(t/z)}{\partial t}$ exist and are continuous in t for each z.

(I.5) The functions $q(z)\,,\ F(t/z)$ and $G_2(t/z)$ have bounded continuous first and second partial derivatives with respect to z.

(I.6) For any closed interval $[a,b]\subset\mathbb{R}^+$, there exists constants $\rho,\ \delta(\varepsilon)>0$ such that

$$\mathbb{P}(\xi_i>\rho/Z_i=z)\geq\delta(\varepsilon)\,,\quad\forall z\in[a,b]\,,$$

with $q(z)\geq\varepsilon$ and $\varepsilon>0$ arbitrary small.

Note that in view of (1.1) the conditions (I.4) and (I.5) for d.f.$G_1(t/z)$ are hold too. Moreover, from (I.6) we also have a chain of inequalities

$$1-K(\rho)=\mathbb{P}(\min(X_i,Y_{1i})>\rho)\geq\mathbb{P}(\xi_i>\rho)=$$

$$=\int q(z)\,\mathbb{P}(\xi_i>\rho/Z_i=z)\,dz\geq\varepsilon\cdot\int\limits_{\{z:q(z)\geq\varepsilon\}}\mathbb{P}(\xi_i>\rho/Z_i=z)\,dz>0.$$

The properties of estimators (2.11) are established from the corresponding properties of estimators (2.3), (2.6), (2.7) and (2.9). In accordance with results of Cheng [7] under Conditions I we obtain asymptotic unbiasedness of $\tilde{A}_n, \tilde{\tilde{A}}_n, A_n, B_n$ and hence the uniform strong consistency of $\tilde{\Lambda}_n(t/z)$ and $\tilde{\tilde{\Lambda}}_n(t/z)$ over a rectangle $[0,\tau]\times[a,b]$ with $\tau\in(0,T_H)$ and rate of convergence as follows:

$$\sup_{(t,z)\in[0,\tau]\times[a,b]}\left|\tilde{\Lambda}_n(t/z)-\tilde{\Lambda}(t/z)\right|\overset{\text{a.s.}}{=}O\left(\left(\frac{\log n}{na_n}\right)^{1/2}\right)+O\left(a_n^2\right), \tag{2.12}$$

$$\sup_{(t,z)\in[0,\tau]\times[a,b]}\left|\tilde{\tilde{\Lambda}}_n(t/z)-\tilde{\tilde{\Lambda}}(t/z)\right|\overset{\text{a.s.}}{=}O\left(\left(\frac{\log n}{na_n}\right)^{1/2}\right)+O\left(a_n^2\right). \tag{2.13}$$

Then

$$\sup_{(t,z)\in[0,\tau]\times[a,b]}\left|\Lambda_n(t/z)-\Lambda(t/z)\right|\overset{\text{a.s.}}{=}O\left(\left(\frac{\log n}{na_n}\right)^{1/2}\right)+O\left(a_n^2\right), \tag{2.14}$$

$$\sup_{(t,z)\in[0,\tau]\times[a,b]}\left|K_{1n}(t/z)-K(t/z)\right|\overset{\text{a.s.}}{=}O\left(\left(\frac{\log n}{na_n}\right)^{1/2}\right)+O\left(a_n^2\right), \tag{2.15}$$

where (2.14) is consequence of (2.12), (2.13) and triangular inequality, (2.15) follows from (2.12) and inequality $|a-b|\leq|\log a-\log b|$, for $0<a,b\leq 1$. It is easy to see that statistics (2.3) is strong consistent and asymptotically unbiased estimator of γ. From Consequence 3 in [3] for each $m=-1,0,1$ and any $\varepsilon>0$ we have

$$\mathbb{P}\left(|p_{mn}-p_m|>\left(\frac{\varepsilon\log n}{n}\right)^{1/2}\right)\leq 2n^{-\varepsilon}, \tag{2.16}$$

also if $\min(p_m,1-p_m)\geq 2\left(\frac{\varepsilon\log n}{n}\right)^{1/2}$, then

$$\mathbb{P}\left(p_{mn}^{-1}>2p_m^{-1}\right)\leq 2n^{-\varepsilon}, \tag{2.17}$$

and

$$\mathbb{P}\left((1-p_{mn})^{-1}>2(1-p_m)^{-1}\right)\leq 2n^{-\varepsilon}. \tag{2.18}$$

Hence from (2.16)-(2.18) by Borel-Cantelly lemma for $m=-1,0,1$ we have

$$|p_{mn}-p_m|\overset{\text{a.s}}{=}O\left(\left(\frac{\log n}{n}\right)^{1/2}\right),$$

and with probability one

$$\frac{1}{p_{mn}}<\frac{2}{p_m},\quad\frac{1}{1-p_{mn}}<\frac{2}{1-p_m}.$$

Adapting characterization of simple proportional hazard model under independent random censoring from the right (see, [1,3-6,8,9]) we get following property of considered conditionally partially informative competing risks model. For a given covariate $Z_i=z$ partially observable (only if $\delta_i\neq -1$) r.v.-s $\min(X_i,Y_{1i})$ and indicators $I(X_i\leq Y_{1i})$ are independent if and only if the representation (1.1) is satisfied. Hence under occurrence of events $A_z^{(i)}=\{Z_i=z\}\cap\{\delta_i\neq -1\}$ the r.v.-s

$\xi_i = \min(X_i, Y_{1i}, Y_{2i})$ and δ_i are conditionally independent if and only if the representation (1.1) is satisfied. This characterization property of considered model is very useful in investigating of asymptotical properties of estimators (2.11).

In the next theorem we show that all three statistics $F_{nl}(t)$, $l = 1, 2, 3$, are aconsistent estimators for the unconditional d.f.$F(t)$.

Theorem. Under Conditions I statistics $\{F_{ln}(t),\ l = 1, 2, 3\}$ are a uniformly strongly consistent estimators for d.f. $F(t)$ on $[0, \tau]$:

$$\sup_{0 \le t \le \tau} |F_{ln}(t) - F(t)| \underset{n \to \infty}{\to} 0, \text{ a.s.}$$

Moreover we also prove weak convergence of normed processes

$$\left\{V_{ln}(t) = \sqrt{n}\left(F_{ln}(t) - F(t)\right),\ t \in [0, \tau], l = 1, 2, 3\right\}$$

to the same Gaussian process.

References

[1] Abdushukurov A.A.(1998). Nonparametric estimation of the distribution function based on relative risk function *Commun Statist:.Theory & Meth.* Vol. **27**, N. **8**, pp. 1991-2012.

[2] Abdushukurov A.A.(1999). On nonparametric estimation of reliability indices by censored sample. *Theory Probab. Appl.*. Vol. **43**, N. **1**, pp. 3-11.

[3] Abdushukurov A.A., Nedzvedsky D.T. (2005). Asymptotic properties of empirical processes on censored samples of ramdomsices. *J. Math. Sciences*. Vol. **127**, N. **1**, pp. 931-939.

[4] Abdushukurov A.A., Makhmudova D. (2008). Semiparametric estimation of the distribution function in the informative competing risks model. *In: Statistical Methods of Estimation and Hypoteses Testing. Perm. Perm State University.* pp. 98-106. (In Russian).

[5] Abdushukurov A.A.(2011). Nonparametric estimation based on incomplete observations. *In: International Enciclopedia of Statistical Sciences. (Prof. Miodrag Lovric, Editor). Springer*. Pt.**14.** pp. 962-964.

[6] Abdushukurov A.A. (2011). *Estimates of unknown distributions from incomplete observations and its properties.*. LAMBERT Academic Publishing. 301p. (In Russian).

[7] Cheng P.E. (1989). Nonparametric estimation of survival curve under dependent censorship. *J. Statist. Plann. Infer.* Vol. **23**, pp. 181-191.

[8] Csörgő S.(1988). Estimation in the Proportional Hazards Model of random censorship. *Statistics*. Vol. **19**, N. **3**, pp. 437-467.

[9] Chather U., Pawlischko J. (1998). Estimating the survival function under a generalized Koziol-Green model with partially informative censoring. *Metrika*. Vol. **48**, pp. 189-207.

[10] Jensen U., Wiedmann J. (2000). Estimation of a Survival Curve under Dependent Cenoring *Second Internat.Conf. on Math.Meth-s in Reliability. Bordeaux. France. July 4-7.* Vol. **2**, pp. 571-574.

[11] Zhang H., Rao M.B. (2004). On generalized maximum likelihood estimation in the proportional hazards model with partially informative censoring *Metrika*. Vol. **59**, pp. 125-136.

Asymptotic Results for Copula Estimator of Survival Function under Random Right Censored Observations at Fixed Covariate Values

Rustamjon S. Muradov
Institute of Mathematics and National University of Uzbekistan, Tashkent, Uzbekistan
e-mail: r_muradov@myrambler.ru

Abstract

In this work we consider estimator of survival function under random censored observations in the presence of covariate, where the dependence between lifetime and censoring variable is expressed by a given Archimedian copula. We present some asymptotic results of estimator.

Keywords: Censored observations, covariate, asymptotic representation, weak convergence, Archimedian copula, Gaussian process.

Introduction

In survival analysis our interest focuses on a nonnegative random variables (r.v.-s) denoting death times of biological organisms or failure times of mechanical systems. A difficulty in the analysis of survival data is the possibility that the survival times can be subjected to random censoring by other nonnegative r.v.-s and therefore we observe incomplete data. There are various types of censoring mechanisms. In this article we consider only right censoring model and problem of estimation of conditional survival function when the survival times and censoring times are dependent and new estimates of conditional survival function assuming that the dependence structure is described by a known copula function. We also consider integral-type estimator of survival function under random right censored observations at fixed covariate values, where the dependence between a life time and a censoring variable may expressed by a given Archimedean copula. We demonstrate almost sure asymptotic representation which provides a key tool for obtaining weak convergence result for estimator.

1 A short introduction to the concept of copulas

Without a doubt the dependence relations between random variables plays a very important role in many fields of mathematics and is one of the most widely studied subjects in probability and statistics. M. Fr´echet and G. Dall'Aglio (see [8]) did some interesting works about this matter in the fifties, studying the bivariate and trivariate distribution functions (d.f.-s) with given univariate margins. The answer to this problem for the univariate margins case was given by A. Sklar creating a new class of functions which he called copulas. The concept of copulas was introduced in 1959 (see [10]) to study the linkage between multivariate distribution functions and their univariate marginals. Since then, copulas have gained growing importance as a tool for modeling statistical dependence of random variables in many fields. We begin with a short introduction to the concept of copulas. For more details with an emphasis on the statistical and mathematical foundations of copulas see Nelsen [8].

Definition 1. *A copula* $C\left(u,v\right):\left[0,1\right]^2\rightarrow\left[0,1\right]$ *is a bivariate distribution function with uniform marginals.*

A first example of copulas is the product copula $C\left(u,v\right)=uv,$ which characterizes independent r.v.-s when the d.f.-s are continuous. The importance of copulas in statistics is described in Sklar's Theorem.

Theorem 1 (8)**.** *Let* H *be a joint d.f. with margins* F *and* G*. Then there exists a copula* C *such that for all* x,y *in* R,

$$H\left(x,y\right)=C\left(F\left(x\right),G\left(y\right)\right). \tag{1}$$

If F *and* G *are continuous, then* C *is unique; otherwise,* C *is uniquely determined on* $Ran\left(F\right)\times Ran\left(G\right)$*. Conversely, if* C *is a copula and* F *and* G *are d.f.-s, then the function* H *defined by (1) is a joint d.f. with margins* F *and* G*. Thus copulas link joint d.f.-s to their one-dimensional margins.*

The representation (1) suggests that if the copula C were known, then substituting continuous marginal estimators for F and G would yield a plug-in estimate of their associated joint d.f. H. Moreover, in light of Sklar's result with arrive at the following functional definition of a copula.

Definition 2. *Given a bivariate d.f.* H *with marginals* F *and* G, *the function defined as*

$$C\left(u,v\right)=H\left(F^{-1}\left(u\right),G^{-1}\left(v\right)\right),$$

for $\left(u,v\right)\in\left[0,1\right]^2$*, where* $F^{-1}\left(u\right)$ *and* $G^{-1}\left(v\right)$ *are the inverse functions of* F *and* G *respectively, is the copula corresponding to* H*.*

In many applications, the r.v.-s of interest represent the lifetimes of individuals or objects in some population. The probability of an individual living or surviving beyond time x is given by the survival function $S(x)=P\left(X>x\right)=1-F\left(x\right)$, where, as before, F denotes the d.f. of X. Let C be the copula function of the bivariate distribution of $\left(X,Y\right)$. We have

$$\bar{H}\left(x,y\right)=P\left(X>x,Y>y\right)=1-F\left(x\right)-G\left(y\right)+H\left(x,y\right)=$$

$$=S\left(x\right)+S\left(y\right)-1+C\left(1-S\left(x\right),1-S\left(y\right)\right)=C^*\left(S\left(x\right),S\left(y\right)\right),$$

where $C^*\left(u,v\right)=u+v-1+C\left(1-u,1-v\right)$-survival copula function.

Let φ be a continuous, strictly decreasing function from $[0,1]$ to $[0,\infty]$such that $\varphi\left(1\right)=0$.

Definition 3. *The pseudo-inverse of φ is the function $\varphi^{[-1]}$ with $Dom\varphi^{[-1]} = [0, \infty]$ and given by*

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 < t < \varphi(0), \\ 0, & \varphi(0) \leq t \leq \infty. \end{cases}$$

Note that $\varphi^{[-1]}$ is continuous and no increasing on $[0, \infty]$, and strictly decreasing on $[0, \varphi(0)]$. Furthermore,

$$\varphi\left(\varphi^{[-1]}(t)\right) = \begin{cases} t, & 0 < t < \varphi(0), \\ 0, & \varphi(0) \leq t \leq \infty, \end{cases} = \min(t, \varphi(0)).$$

If $\varphi(0) = \infty$, then $\varphi^{[-1]} = \varphi^{-1}$.

Definition 4. *Copulas of the form $C(u, v) = \varphi^{[-1]}(\varphi(u) + \varphi(v))$ are called Archimedean copulas, where the function φ is called a generator of the copula $\varphi(1) = 0$.*

2 The right censoring model

In such research areas as bio-medicine, engineering, insurance, social sciences and many others researchers are interested in positive variables, which are expressed as a time until a certain event. For example, in medicine the survival time of individual, while in industrial trials, time until breakdown of a machine are non-negative r.v.-s of interest. But in such practical situations, the observed data may be incomplete, that is censored. This is the case, for example, in medicine when the event of interest-death due to a given cause and the censoring event is death due to other cause. In industrial study, it may occur that some piece of equipment is taken away (that is censored) because it shows some sign of future failure. Moreover, the r.v.-s of interest (lifetimes, failure times) and censoring r.v.-s usually can be influenced by other variable, often called prognostic factor or covariate. In medicine, dose of a drug and in engineering some environmental conditions (temperature, pressure) are influenced to the observed variables. The basic problem consist in estimation of distribution of lifetime by such censored dependent data. The aim of paper is considering this problem in the case of right random censoring model in the presence of covariable.

Let's consider the case when the support of covariate C is the interval $[0, 1]$ and we describe our results on fixed design points $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1$ at which we consider responses (survival or failure times) $X_1, ..., X_n$ and censoring times $Y_1, ..., Y_n$ of identical objects, which are under study. These responses are independent and nonnegative r.v.-s with conditional distribution function (d.f.) at x_i, $F_{x_i}(t) = P(X_i \leq t/C_i = x_i)$. They are subjected to random right censoring, that is for X_i there is a censoring variable Y_i with conditional d.f. $G_{x_i}(t) = P(Y_i \leq t/C_i = x_i)$ and at n-th stage of experiment the observed data is

$$S^{(n)} = \{(Z_i, \delta_i, C_i), 1 \leq i \leq n\},$$

where $Z_i = min(X_i, Y_i), \delta_i = I(X_i \leq Y_i)$ with $I(A)$ denoting the indicator of event A. Note that in sample $S^{(n)}$ r.v. X_i is observed only when $\delta_i = 1$. Commonly, in survival analysis to assume independence between the r.v.-s X_i and Y_i conditional on the covariate C_i. But, in some practical situations, this assumption does not hold. Therefore, in this article we consider a dependence model in which dependence structure is described through copula function. So let

$$S_x(t_1,t_2)=P(X_x>t_1,Y_x>t_2),t_1,t_2\geq 0,$$

the joint survival function of the response X_x and the censoring variable Y_x at x. Then the marginal survival functions are $S_x^X(t)=1-F_x(t)=S_x(t,0)$ and $S_x^Y(t)=1-G_x(t)=S_x(0,t),t\geq 0$. We suppose that the marginal d.f.-s F_x and G_x are continuous. Then according to the Theorem of Sklar (see, section 1), the joint survival function $S_x(t_1,t_2)$ can be expressed as

$$S_x(t_1,t_2)=C_x(S_x^X(t_1),S_x^Y(t_2)),t_1,t_2\geq 0, \tag{2}$$

where $C_x(u,v)$ is a known copula function depending on x, S_x^X and S_x^Y in a general way.

3 Asymptotic results for estimator

Assume that at the fixed design value $x\in(0,1)$, C_x in (2) is Archimedean copula, i.e.

$$S_x(t_1,t_2)=\varphi_x^{[-1]}(\varphi_x(S_x^X(t_1))+\varphi_x(S_x^Y(t_2))),t_1,t_2\geq 0, \tag{3}$$

where copula generator function φ_x is strict, i.e. $\varphi_x(0)=\infty$ and hence $\varphi_x^{[-1]}=\varphi_x^{-1}$. From (3), it follows that

$$P(Z_x>t)=1-H_x(t)=\overline{H_x(t)}=S_x^Z(t)=S_x(t,t)=$$

$$=\varphi_x^{-1}(\varphi_x(S_x^X(t))+\varphi_x(S_x^Y(t))),\ \ t\geq 0, \tag{4}$$

Let $H_x^{(1)}(t)=P(Z_x\leq t,\delta_x=1)$ be a subdistribution function and $\Lambda_x(t)$ is crude hazard function of r.v. X_x subjecting to censoring by Y_x, that is

$$\Lambda_x(dt)=\frac{P(X_x\in dt,X_x\leq Y_x)}{P(X_x\geq t,Y_x\geq t)}=\frac{H_x^{(1)}(dt)}{S_x^Z(t-)}. \tag{5}$$

From (4) and (5) one can obtain following expression of survival function S_x^X :

$$S_x^X(t)=\varphi_x^{-1}[-\int_0^t S_x^Z(u-)\varphi_x'(S_x^Z(u))d\Lambda_x(u)]=$$

$$=\varphi_x^{-1}[-\int_0^t \varphi_x'(S_x^Z(u))dH_x^{(1)}(u)],\ \ t\geq 0, \tag{6}$$

(see, for example, [1-4]). In order to constructing the estimator of S_x^X according to representation (6), we introduce some smoothed estimators of $S_x^Z, H_x^{(1)}$ and regularity conditions for them. Similarly to Breakers and Veraverbeke [2], we will also use the Gasser-Müller weights

$$\omega_{ni}(x,h_n)=\frac{1}{q_n(x,h_n)}\int_{x_{i-1}}^{x_i}\frac{1}{h_n}\pi(\frac{x-z}{h_n})dz,\ i=1,...,n,$$

with

$$q_n(x,h_n)=\int_0^{x_n}\frac{1}{h_n}\pi(\frac{x-z}{h_n})dz,$$

where $x_0=0$, π is a known probability density function(kernel) and $\{h_n,n\geq 1\}$ is a sequence of positive constants, tending to zero as $n\to\infty$, called bandwidth sequence. Let's introduce the weighted estimators of H_x, S_x^Z and $H_x^{(1)}$ respectively as

$$H_{xh}(t)=\sum_{i=1}^n\omega_{ni}(x,h_n)I(Z_i\leq t),$$

$$S_{xh}^Z(t)=1-H_{xh}(t), \tag{7}$$

$$H_{xh}^{(1)}(t)=\sum_{i=1}^n\omega_{ni}(x,h_n)I(Z_i\leq t,\delta_i=1).$$

Then pluggin in (6) estimators (7) we get corresponding estimator of $S_x^X(t)$ as

$$S_{xh}^X(t)=1-F_{xh}(t)=\varphi_x^{-1}[-\int_0^t\varphi_x'(S_{xh}^Z(u))dH_{xh}^{(1)}(u)],\ t\geq 0, \tag{8}$$

Remark that in the case of no covariate, estimator (8) reduces to estimator first obtained by Zeng and Klein [11]. In the case of the independent copula $\varphi(y)=-logy$, Zeng and Klein estimate reduces to a exponential-hazard estimate (see, [1,2,9,11]). Also it is well-known that under independent censoring case Kaplan-Meier's product-limit estimator and exponential-hazard estimators are asymptotical equivalent. Therefore, we will show that estimator and copula-graphic estimator of Breakers and Veraverbeke[2] have the same asymptotic behaviours.

For the design points $x_1,...x_n$, denote

$$\underline{\Delta_n}=\min_{1\leq i\leq n}(x_i-x_{i-1}),\ \ \overline{\Delta_n}=\max_{1\leq i\leq n}(x_i-x_{i-1}).$$

For the kernel π, let

$$\|\pi\|_2^2=\int_{-\infty}^{\infty}\pi^2(u)du,\ \ m_\nu(\pi)=\int_{-\infty}^{\infty}u^\nu\pi(u)du,\ \nu=1,2.$$

Moreover, we use next assumptions on the design and on the kernel function:
$(A1)$ As $n\to\infty,\ x_n\to 1,\ \overline{\Delta_n}=O(\frac{1}{n}),\ \overline{\Delta_n}-\underline{\Delta_n}=o(\frac{1}{n})$.
$(A2)$ π is a probability density function with compact support $[-M,M]$ for some $M>0$, with $m_1(\pi)=0$ and $|\pi(u)-\pi(u')|\leq C(\pi)|u-u'|$, where $C(\pi)$ is some constant.

Let $T_{H_x} = \inf\{t \geq 0 : H_x(t) = 1\}$. Then $T_{H_x} = \min(T_{F_x}, T_{G_x})$. We need some smoothness conditions on functions $H_x(t)$ and $H_x^{(1)}(t)$. We formulate them for a general (sub)distribution function $N_x(t), 0 \leq x \leq 1, t \in R$ and for a fixed $T > 0$.
$(A3)$ $\frac{\partial}{\partial x}N_x(t) = \dot{N}_x(t)$ exists and is continuous in $(x,t) \in [0,1] \times [0,T]$.
$(A4)$ $\frac{\partial}{\partial t}N_x(t) = N'_x(t)$ exists and is continuous in $(x,t) \in [0,1] \times [0,T]$.
$(A5)$ $\frac{\partial^2}{\partial x^2}N_x(t) = \ddot{N}_x(t)$ exists and is continuous in $(x,t) \in [0,1] \times [0,T]$.
$(A6)$ $\frac{\partial^2}{\partial t^2}N_x(t) = N''_x(t)$ exists and is continuous in $(x,t) \in [0,1] \times [0,T]$.
$(A7)$ $\frac{\partial^2}{\partial x \partial t}N_x(t) = \dot{N}'_x(t)$ exists and is continuous in $(x,t) \in [0,1] \times [0,T]$.
$(A8)$ $\frac{\partial \varphi_x(u)}{\partial u} = \varphi'_x(u)$ and $\frac{\partial^2 \varphi_x(u)}{\partial u^2} = \varphi''_x(u)$ are Lipschitz in the x-direction with a bounded Lipschitz constant and $\frac{\partial^3 \varphi_x(u)}{\partial u^3} = \varphi'''_x(u) \leq 0$ exists and is continuous in $(x,u) \in [0,1] \times (0,1]$.

It is clear that for existence of right hand side of representation (6) we must require the conditions (A 4) for functions $H_x(t)$ and $H_x^{(1)}(t)$ in $[0,1] \times [0,T]$ with $T < T_{H_x}$ and existence of $\varphi'_x(u)$ on $[0,1] \times (0,1]$.

We present almost sure representation result with rate.

Theorem 2 (1,2). *Assume (A 1), (A 2), $H_x(t)$ and $H_x^{(1)}(t)$ satisfy (A 5)-(A 7) in $[0,T]$ with $T < T_{H_x}$, φ_x satisfies (A 8) and $h_n \to \infty$, $\frac{logn}{nh_n} \to 0$, $\frac{nh_n^5}{logn} = O(1)$. Then, as $n \to \infty$,*

$$F_{xh}(t) - F_x(t) = \sum_{i=1}^{n} \omega_{ni}(x,h_n)\Psi_{tx}(Z_i,\delta_i) + r_n(t),$$

where

$$\Psi_{tx}(Z_i,\delta_i) = \frac{-1}{\varphi'_x(S_x^X(t))}[\int_0^t \varphi''_x(S_x^Z(u))(I(Z_i \leq u) - H_x(u))dH_x^{(1)}(u) -$$

$$-\varphi'_x(S_x^Z(t))(I(Z_i \leq t, \delta_i = 1) - H_x^{(1)}(t)) - \int_0^t \varphi''_x(S_x^Z(u))(I(Z_i \leq u, \delta_i = 1) - H_x^{(1)}(u))dH_x(u)]$$

and

$$\sup_{0 \leq t \leq T} |r_n(t)| \overset{a.s.}{=} O\left(\left(\frac{logn}{nh_n}\right)^{3/4}\right).$$

The weak convergence of the process $(nh_n)^{1/2}F_{xh}(\cdot) - F_x(\cdot)$ in the space $\ell^\infty[0,T]$ of uniformly bounded functions on $[0,T]$, endowed with the uniform topology is the contents of the next theorem.

Theorem 3 (1,2). *Assume (A 1), (A 2), $H_x(t)$ and $H_x^{(1)}(t)$ satisfy (A 5)-(A 7) in $[0,T]$ with $T < T_{H_x}$, and that φ_x satisfies (A 8).*
(I) If $nh_n^5 \to 0$ and $\frac{(logn)^3}{nh_n} \to 0$, then, as $n \to \infty$,

$$(nh_n)^{1/2}\{F_{xh}(\cdot) - F_x(\cdot)\} \Longrightarrow \boldsymbol{W}_x(\cdot) \;\; in \;\; \ell^\infty[0,T].$$

(II) If $h_n = Cn^{-1/5}$ for some $C > 0$, then, as $n \to \infty$,

$$(nh_n)^{1/2}\{F_{xh}(\cdot) - F_x(\cdot)\} \Longrightarrow \boldsymbol{W}^*_x(\cdot) \;\; in \;\; \ell^\infty[0,T],$$

where $\boldsymbol{W}_x(\cdot)$ *and* $\boldsymbol{W}^*_x(\cdot)$ *are Gaussian processes with means*

$$E\,\boldsymbol{W}_x(t) = 0, \quad E\,\boldsymbol{W}^*_x(t) = a_x(t),$$

and same covariance

$$Cov(\boldsymbol{W}_x(t), \boldsymbol{W}^*_x(s)) = Cov(\boldsymbol{W}^*_x(t), \boldsymbol{W}^*_x(s)) = \Gamma_x(t,s),$$

with

$$a_x(t) = \frac{-C^{5/2} m_2(\pi)}{2\varphi'_x(S^X_x(t))} \int_0^t [\varphi''_x(S^Z_x(u))\ddot{H}_x(u) dH^{(1)}_x(u) - \varphi'_x(S^Z_x(u)) d\ddot{H}^{(1)}_x(u)],$$

and

$$\Gamma_x(t,s) = \frac{\|\pi\|_2^2}{\varphi'_x(S^X_x(t))\varphi'_x(S^X_x(s))}\{\int_0^{\min(t,s)} (\varphi'_x(S^Z_x(z)))^2 dH^{(1)}_x(z) +$$

$$+\int_0^{\min(t,s)} [\varphi''_x(S^Z_x(w))S^Z_x(w) + \varphi'_x(S^Z_x(w))] \int_0^w \varphi''_x(S^Z_x(y)) dH^{(1)}_x(y) dH^{(1)}_x(w) +$$

$$+\int_0^{\min(t,s)} \varphi''_x(S^Z_x(w)) \int_w^{\max(t,s)} (\varphi''_x(S^Z_x(y))S^Z_x(y) + \varphi'_x(S^Z_x(y))) dH^{(1)}_x(y) dH^{(1)}_x(w) -$$

$$-\int_0^t [\varphi''_x(S^Z_x(y))S^Z_x(y) + \varphi'_x(S^Z_x(y))] dH^{(1)}_x(y) \int_0^s [\varphi''_x(S^Z_x(w))S^Z_x(w) + \varphi'_x(S^Z_x(w))] dH^{(1)}_x(w)\}.$$

References

[1] Abdushukurov A.A.,Muradov R.S. (2014). Estimation of survival and mean residual life functions from dependent random censored data, *New Trends in Mathematical Sciences*, Vol.**2**, 35-48.

[2] Abdushukurov A.A., Muradov R.S.(2014). On estimation of conditional distribution function under dependent random right censored observations. *Journal of Siberian Federal University. Mathematics & Physics.* Vol.**7**. p.409-416.

[3] Breakers R.,Veraverbeke N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring, *The Canadian Journal of Statistics*, Vol.**33**, 429-447.

[4] Fleming T.R.,Harrington D.P. (1991). *Counting Processes and Survival Analysis*, Wiley, New York.

[5] Kaplan E.L.,Meier P. (1958). Nonparametric estimation from incompelet observations, *Journal of American Statistical Association*, Vol.**53**, 457-481.

[6] Keilegom I.Van.,Veraverbeke N. (1997). Estimation and the bootstrap with censored data in fixed design nonparametric regression, *Ann. Inst. Statist. Math.*, Vol.**49**, 467-491.

[7] Muradov R.S.,Abdushukurov A.A. (2011). *Estimation of multivariate distributions and its mixtures*. LAMBERT Academic Publishing. LAP, Germany, (In Russian).

[8] Nelsen R.B. (1999). *An Introduction to Copulas*. Springer, New York.

[9] Rivest L.P.,Wells M.T. (2001). A martingall aproach to the copula-graphic estimator for the survival function under dependent censoring, *Journal of Multivariate Analysis*, Vol.**79**, 138-155.

[10] Sklar A. (1959). Fonctions de repartition à n dimensions et leurs marges, Publications de l'Institut de Statistique de l'Université de Paris,Vol.**8**, 229-231

[11] Zeng M.,Klein J.P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika*, Vol.**82**, 127-138.

On Inference for the Generalized Pareto Distribution

Hideki Nagatsuka[1] and N. Balakrishnan[2,3]

[1] *Department of Industrial and Systems Engineering, Chuo University, Tokyo, Japan*

[2] *Department of Mathematics and Statistics, McMaster University, Ontario, Canada*

[3] *Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia*

e-mail: `hideki@indsys.chuo-u.ac.jp, bala@mcmaster.ca`

Abstract

The generalized Pareto distribution (GPD) is widely used to model exceedances over thresholds. The estimation of the parameters of the GPD is a difficult problem, and existing methods for estimating parameters have theoretical or computational defects. In this article, we introduce the method of inference for the GPD, proposed by Nagatsuka and Balakrishnan (2015, submitted paper), which can successfully estimates the parameters, constructs the confidence intervals and likelihood ratio tests, over the entire parameter space. In this method, the estimates always exist uniquely, and the estimators are also consistent over the entire parameter space. The method are compared with some prominent methods through a Monte Carlo simulation study.

Keywords: Generalized Pareto Distribution, Exceedances over thresholds, Consistency, Existence, Uniqueness.

Introduction

Statistical modeling of the largest or smallest values (extreme values) of certain natural phenomena (e.g., waves, floods, earthquakes, winds, temperatures etc) is of interest in various practical applications. For example, the distributions of high waves and large floods are important in the designs of dikes and dams, respectively. The traditional approach to the analysis of extreme values is based on the generalized extreme value distribution (GEVD), which is a limiting distribution for extreme values, including the Gumbel, Frechet and Weibull distributions ([3], [2], [1], and [4]). Although the GEVD is appropriate to be fitted to the data consisting of the set of maxima, there has been some criticism since using only maxima leads to the loss of information contained in other values in the given data set. This problem is remedied by considering some largest values in the given period instead of the largest value, that is, considering all values larger than a given threshold (exceedances over the threshold). The generalized Pareto distribution (GPD), which is a limiting distribution for exceedances over the threshold, offers a unifying approach to the modelling of such values ([3], [2], [1], and [4]). This distribution was initially introduced by [9], and has been widely used to analyze exceedances over the threshold in various areas (see, for example, [10]). The cumulative distribution function (cdf) of the GPD is

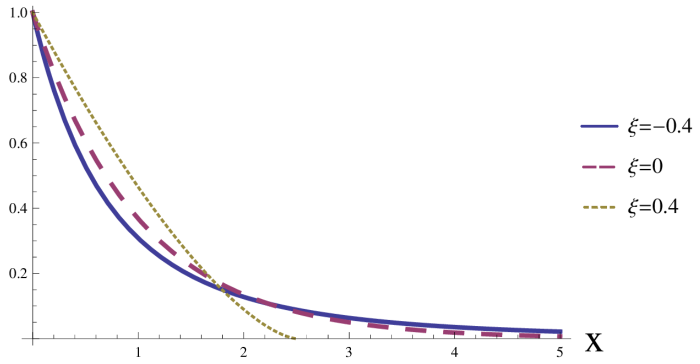


Figure 1: Pdfs of GPDs for different values of the shape parameter ξ, with $\sigma = 1$

given by

$$F(x;\,\xi,\sigma) = \begin{cases} 1 - \left(1 - \xi\dfrac{x}{\sigma}\right)^{1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-\dfrac{x}{\sigma}\right), & \xi = 0. \end{cases} \tag{1}$$

where $\xi \in \mathbb{R}$ and $\sigma > 0$ are the shape and scale parameters, respectively. For $\xi \leq 0$, the range is $0 \leq x < \infty$, while for $\xi > 0$, $0 \leq x \leq \sigma/\xi$. The corresponding probability density function (pdf) is

$$f(x;\,\xi,\sigma) = \begin{cases} \dfrac{1}{\sigma}\left(1 - \xi\dfrac{x}{\sigma}\right)^{1/\xi-1}, & \xi \neq 0, \\ \dfrac{1}{\sigma}\exp\left(-\dfrac{x}{\sigma}\right), & \xi = 0. \end{cases} \tag{2}$$

for $\xi \in \mathbb{R}$ and $\sigma > 0$.

The shape of the pdf varies with respect to the shape parameter ξ (see Figure 1). The smaller the value of ξ becomes, the heavier tailed the distribution becomes, that is, more very large values can be observed, whereas the larger the value of ξ becomes, the more light tailed the distribution becomes. In fact, $-1/\xi$ is known as the tail index, and we can know the risk in the situation from the value of ξ or the tail index. For example, the small value of ξ (or the tail index) indicates that the events associated with large values occur with high probability.

Although the GPD is useful for modeling exceedances over the threshold, it is well known that parameter estimation for the GPD is a difficult problem. We refer the reader to [5] and [6], which are excellent survey papers on methods of parameter estimation for the GPD. For $\xi > 1$, the maximum likelihood estimators do not exist. For $\xi \leq -1/r$, $r \in \mathbb{N}$, the rth moment does not exist, and therefore, all the moment-based estimators such as the method of moments (MOM) estimators, the probability weighted moments (PWM) estimators and the L-moments estimators proposed by

[7] exist only for ξ in certain ranges. Recently, two empirical Bayesian methods of parameter estimation have been proposed by [11] and [12]. They have shown good performances for moderate or small ξ ($\xi \leq 0.5$), by Monte Carlo simulations. However, for large ξ, their estimators have considerable large bias and RMSE even if the sample size is large, shown later by a Monte Carlo simulation. This result indicates that their estimators may not have consistency for large ξ. In spite of many papers dealing with the parameter estimation for the GPD, there does not appear to be any work wherein it is established formally that estimates always exist uniquely and that the estimators are also consistent over the entire parameter space.

In this paper, we introduce the new method of estimation for the parameters of the GPD, proposed by [8]. In this method, under mild conditions, the estimates always exist uniquely, and the estimators also have consistency over the entire parameter space.

The rest of this article is organized as follows. In Section 1, we describe the new estimators and discuss some of their properties. In Section 2, we show that the new method performs well in comparison with some prominent methods of estimation of parameters, in terms of bias and root mean squared error (RMSE). Finally, some concluding remarks are made.

1 Method of Estimation

In this section, we describe the new estimators of ξ and σ, and discuss some of their properties. All proofs of theorems, lemmas, and corollaries are omitted due to space constraints.

1.1 Estimation of the shape parameter

Let $X_1, \cdots, X_n$ be i.i.d. random variables from the GPD with the cdf in (1), and $X_{1:n} \leq \cdots \leq X_{n:n}$ be the order statistics obtained by arranging the above X_i's in increasing order of magnitude.

For any fixed j, $1 \leq j \leq n$, we derive the joint density of $\boldsymbol{S}_n^{(j)}$, where $\boldsymbol{S}_n^{(j)} = (S_{1:n}, \ldots, S_{j-1:n}, S_{j+1:n}, \ldots, S_{n:n})$, and $S_{i:n} = X_{i:n}/X_{j:n}$, $i \neq j$, $1 \leq i \leq n$.

Theorem 1. *For $\xi \in \mathbb{R}$ and any fixed j, $1 \leq j \leq n$, the joint density of $\boldsymbol{S}_n^{(j)}$ is given by*

$$\phi\left(\boldsymbol{s}_n^{(j)}; \xi\right) = \begin{cases} n! \displaystyle\int_\chi \frac{1}{|\xi|} \left(\frac{u}{\xi}\right)^{n-1} \prod_{i=1}^{n} (1 - us_i)^{1/\xi - 1}\, du, & \xi \neq 0, \\[2ex] \dfrac{n!(n-1)!}{\left(\sum_{i=1}^n s_i\right)^n}, & \xi = 0, \end{cases}$$
$$s_1 \leq \cdots \leq s_n,$$

where $\chi = \{u : -\infty < u < 0,\ \mathit{if}\, \xi < 0,\ or\ ,0 < u < 1/s_n,\ \mathit{if}\, \xi > 0\}$, $\boldsymbol{s}_n^{(j)} = (s_1, \ldots, s_{j-1}, s_{j+1}, \ldots, s_n)$, $s_1 \leq \cdots \leq s_n$, and $s_j = 1$.

From Theorem 1, we can obtain the likelihood function of ξ based on $\boldsymbol{S}_n^{(j)}$ as

$$l\left(\xi; \boldsymbol{s}_n^{(j)}\right) \;=\; \phi\left(\boldsymbol{s}_n^{(j)}; \xi\right), \tag{3}$$

where $\boldsymbol{s}_n^{(j)}$ are the vector consisting of the realized values of $S_{i:n}$, $i \neq j$, $1 \leq i \leq n$. Then, the MLE of ξ based on $\boldsymbol{S}_n^{(j)}$, denoted by $\hat{\xi}$, is obtained by maximizing $l(\xi; \boldsymbol{s}_n^{(j)})$ with respect to ξ, substituting $\boldsymbol{S}_n^{(j)}$ for $\boldsymbol{s}_n^{(j)}$.

It might be noted that the likelihood function $l(\xi; \boldsymbol{s}_n^{(j)})$ depends on j, and considered that which j is optimal. However, from Theorem 2, we note that we do not need to worry about the choice of j, in the ML method based on $\boldsymbol{S}_n^{(j)}$.

Theorem 2. *The MLE of ξ based on $\boldsymbol{S}_n^{(j)}$ does not depend on j.*

We further observe that the ML method based on $\boldsymbol{S}_n^{(j)}$ is equivalent to the ML methods based on $X_{i:n}/X_{i+1:n}$'s and $X_{i+1:n}/X_{i:n}$'s in the sense of the following theorem.

Theorem 3. *MLE of ξ based on $X_{i:n}/X_{i+1:n}$'s (and $X_{i+1:n}/X_{i:n}$'s) agrees with those based on $\boldsymbol{S}_n^{(j)}$.*

The following theorem gives the derivative of $l(\xi; \boldsymbol{s}_n^{(j)})$, which is continuous with respect to ξ.

Theorem 4. *For $\xi \in \mathbb{R}$ and any given $\boldsymbol{s}_n^{(j)}$, the derivative $l'(\xi; \boldsymbol{s}_n^{(j)}) = (\partial/\partial\xi) l(\xi; \boldsymbol{s}_n^{(j)})$ is given by*

$$\begin{aligned} & l'(\xi; \boldsymbol{s}_n^{(j)}) \\ &= \begin{cases} n! \displaystyle\int_{\chi} \left(-\frac{n}{\xi} - \frac{\sum_{i=1}^n \log(1 - us_i)}{\xi^2} \right) \frac{1}{|\xi|} \left(\frac{u}{\xi} \right)^{n-1} \prod_{i=1}^n (1 - us_i)^{1/\xi - 1} \, du, & \xi \neq 0, \\ n! \displaystyle\int_0^\infty u^{n-1} \left\{ \sum_{i=1}^n \left(1 - \frac{us_i}{2} \right) us_i \right\} \exp\left(-u \sum_{i=1}^n s_i \right) du, & \xi = 0, \end{cases} \\ & s_1 \leq \cdots \leq s_n, \end{aligned}$$

and is continuous with respect to ξ on $\mathbb{R}$.

The following theorem and the ensuing corollary imply that the estimate of ξ obtained by maximizing $l\left(\xi; \boldsymbol{s}_n^{(j)}\right)$ or solving the equation $l'\left(\xi; \boldsymbol{s}_n^{(j)}\right)$ always exists uniquely over the entire parameter space.

Theorem 5. *For $\xi \in \mathbb{R}$, any fixed j, $1 \leq j \leq n$, and any given $\boldsymbol{s}_n^{(j)}$, the likelihood equation $l'\left(\xi; \boldsymbol{s}_n^{(j)}\right) = 0$ always has a unique solution with respect to ξ.*

Corollary 1. *For $\xi \in \mathbb{R}$, and any fixed j, $1 \leq j \leq n$, and any given $\boldsymbol{s}_n^{(j)}$, the likelihood function $l\left(\xi; \boldsymbol{s}_n^{(j)}\right)$ is unimodal with respect to ξ.*

For proving the main result that the estimator of ξ has consistency, the following lemma is needed.

Lemma 1. *For any fixed $\xi \neq \xi_0$, where ξ_0 is the true value of the parameter ξ, and for any fixed j, $1 \leq j \leq n$,*

$$\lim_{n\to\infty} \Pr\left(l\left(\xi;\boldsymbol{S}_n^{(j)}\right) < l\left(\xi_0;\boldsymbol{S}_n^{(j)}\right)\right) = 1.$$

Theorem 6. *The estimator $\hat{\xi}$ is consistent for $\xi \in \mathbb{R}$.*

1.2 Estimation of the scale parameter

Once we obtain the estimate of ξ, by using the method outlined above, we can adopt the usual ML method to obtain the estimates of σ, in which the shape parameter ξ is replaced by $\hat{\xi}$. Then, the MLE of σ is given by, after replacing ξ by $\hat{\xi}$,

$$\hat{\sigma} \;=\; \begin{cases} \text{Solution of} \quad n - \left(\dfrac{1}{\hat{\xi}} - 1\right)\displaystyle\sum_{i=1}^{n} \frac{X_i}{\left(\sigma/\hat{\xi} - X_i\right)} = 0, & \hat{\xi} < 1 \text{ and } \neq 0, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i, & \hat{\xi} = 0, \\ \hat{\xi}\, X_{n:n}, & \hat{\xi} \geq 1. \end{cases}$$

Analogous to the estimator of ξ, it is observed that the estimate of σ always exist uniquely and that the estimator of σ are also consistent for σ over the entire parameter space.

Theorem 7. *For $\xi \in \mathbb{R}$, $\sigma > 0$ and any given the observations $x_1, \ldots, x_n$, the estimate of the σ, given by Eq.(4), where $\hat{\xi}$ is substituted for its realized value, uniquely exists.*

Theorem 8. *The estimator $\hat{\sigma}$ is consistent for $\sigma > 0$.*

2 Empirical Evaluation of the New Method of Estimation

Here, we show an extensive Monte Carlo simulation study to evaluate the performance of the new estimators of the parameters. In this Monte Carlo simulation study, the new method (New) was compared with the following prominent methods of estimation; the usual maximum likelihood method (ML), the L-moments method proposed by [7] (Lmom), the Bayesian method proposed by [11] (Bayes 1), and the Bayesian method proposed by [12] (Bayes 2).

Figures 2 and 3 depict the simulation results of the bias and the root mean squared error ($RMSE = \sqrt{variance + bias^2}$) of the estimators of ξ and σ by all methods, based on 1,000 Monte Carlo runs, for $-3 \leq \xi \leq 5$, $\sigma = 1$, and $n = 100$.

From these results, we observe that only the new method successfully obtains the estimates and shows good performances with respect to bias and RMSE in all the cases. The simulation results also indicate that the competitors do not exist or do not have consistency for certain ranges of ξ.

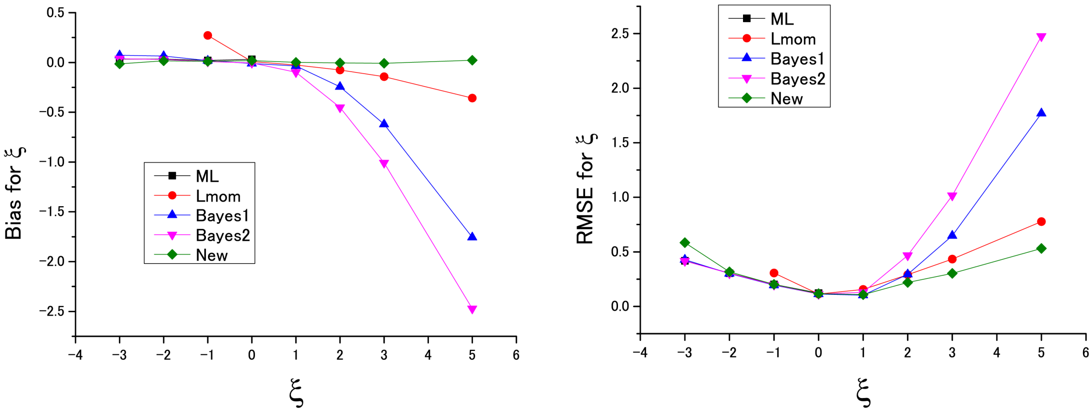


Figure 2: Bias and RMSE of the ML, L-moments, Bayes 1, Bayes 2 and the new methods for the estimation of parameter ξ when $n = 100$ (No dot indicates that the number of the estimates obtained was zero or quite small)

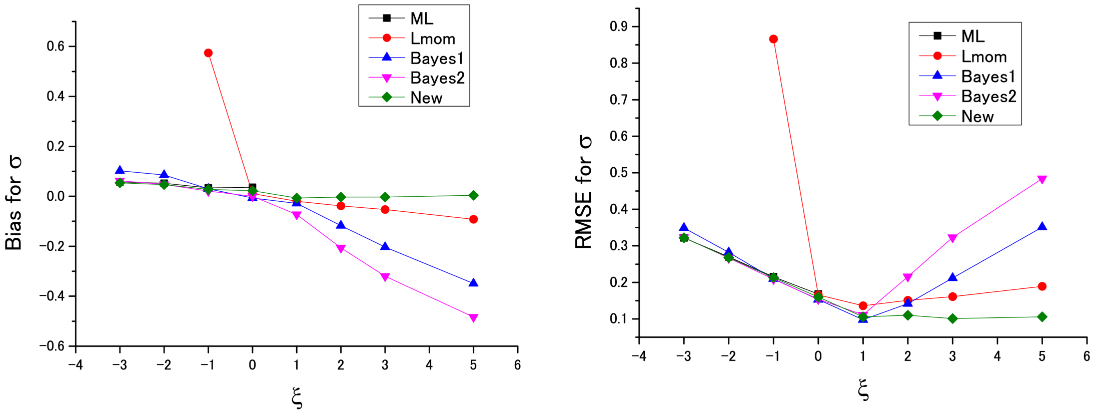


Figure 3: Bias and RMSE of the ML, L-moments, Bayes 1, Bayes 2 and the new methods for the estimation of parameter σ when $n = 100$ (No dot indicates that the number of the estimates obtained was zero or quite small)

Conclusions

We have introduced here a new method of estimation for the GPD, in which the estimates of all the parameters in the new method always exist uniquely, and that they also have consistency over the entire parameter space. We have shown a simulation study to examine the performance of the new method and to compare it with some other prominent methods, and the results have shown that only the new method successfully obtains the estimates and shows good performances with respect to bias and RMSE in all the cases.

We will present the confidence intervals for all parameters and likelihood-based tests based on the new method at our talk.

Acknowledgements

Hideki Nagatsuka was partially supported by the Grant-in-Aid for Scientific Research (C), Japan Society for Promotion of Science KAKENHI, Japan, Grant Number 15K01205, which facilitated his collaborative research visit to the second author. Research of N. Balakrishnan was supported by a Discovory Grant from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Beirlant J., Goegebeur Y., Segers J., Teugels J. (2004). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, Chichester, West Sussex.

[2] Castillo E., Hadi A. S., Balakrishnan N., Sarabia J. M. (2004). *Extreme Value and Related Models with Applications in Engineering and Science*. John Wiley & Sons, Hoboken, New Jersey.

[3] Coles S. (2001). *An Introduction to Statistical Modeling of Extreme Values.*, Springer, London.

[4] de Haan L., Ferreira A. (2006). *Extreme Value Theory: An Introduction*. Springer, New York.

[5] de Zea Bermudeza P., Kotz S. (2010). Parameter estimation of the generalized Pareto distribution – Part I. *Journal of Statistical Planning and Inference*. Vol. **140**, pp. 1353–1373.

[6] de Zea Bermudeza P., Kotz S. (2010). Parameter estimation of the generalized Pareto distribution – Part II. *Journal of Statistical Planning and Inference*. Vol. **140**, pp. 1374–1388.

[7] Hosking J.R.M. (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society. Series B*. Vol. **52**, pp. 105–124.

[8] Nagatsuka H., Balakrishnan N. (2015). A new approach to parameter estimation for the generalized Pareto distribution. *submitted*.

[9] Pickands J (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*. Vol. **3**, pp. 119–131.

[10] Salvadori G., De Michele C., Kottegoda N.T., Rosso R. (2007). *Extremes in Nature An Approach Using Copulas*. Springer, Dordrecht.

[11] Zhang J., Stephens M.A. (2009). A New and Efficient Estimation Method for the Generalized Pareto Distribution. *Technometrics*. Vol. **51**, pp. 316–325.

[12] Zhang J. (2010). Improving on Estimation for the Generalized Pareto Distribution. *Technometrics*. Vol. **52**, pp. 335–339.

Property of LM Tests for Reliability Distributions in case of Small Samples

Toshinari Kamakura and Takenori Sakumura
Chuo University, Tokyo, Japan
e-mail: kamakura@chuo-u.ac.jp, sakumura@chuo-u.ac.jp

Abstract

Weibull distributions are wildly used in engineering and medical fields. Many researchers are interested in devising methods of testing and estimating parameters for applications to reliability and survival data. In this article we will investigate small sample behaviors of testing and estimating Weibull models based on the likelihood including censored data. The Lagrange multiplier (LM) test is simple and may have good performance according to our experience recently. We will derive several LM tests for Weibull parameters; shape parameter, scale parameter and mean parameter with handling censored data. For comparing performance with other statistics we will conduct simulations studies on nominal and actual significance levels for the LM test, the likelihood ratio test and the Wald test. In case of shape parameter with one the derived LM test is very simple without any likelihood calculations, which include iterations. We find that the LM test tend to assures α-size test in that the true significance levels are smaller than the predefined size of α for any sample size and the true α approaches the nominal α from undersides in accordance with sample size n larger.

Keywords: Weibull distribution, LM test, Likelihood ratio test, Wald test, Small sample.

Introduction

The Weibull distribution is largely used in the field of reliability engineering and medical filed and have many literatures in the history. In this article we revisit several important problems on testing and estimating of Weibull parameters, which may be still left unsolved.

In the Rinne's textbook [16] the Weibull distribution is described in detail including its genesis and more than hundreds or even thousands of papers have been written on this distribution. The log-transformation of the Weibull distribution gives the log-Weibull distribution or the extreme value distribution, which is belonging to the location-scale family, and we can use the standard technique devised for the location-sale family such as the general least squares (GLS) based on Gauss-Markov theorem [12]. Later Lawless published many papers of the confidence intervals of parameters based on the extreme value distribution [7, 8, 9, 10]. Kalbfleisch and Prentice [5] showed that the only log-linear models that are also proportional hazards models are the exponential and Weibull regression models [6]. The Weibull model plays a primary role in the survival analysis in that we can use this model in place of the Cox regression model by extending the scale parameter with regression covariates. Nagatsuka et al. [15] proposed the new method of estimation for 3-parameter Weibull distributions using order statistics.

Testing

In this article we consider 2-parameter Weibull modeling which can be expressed by the following density function:

$$f(x;\ \eta,\ m) = \left(\frac{m}{\eta}\right)\left(\frac{x}{\eta}\right)^{m-1} \exp\left\{-\left(\frac{x}{\eta}\right)^{m}\right\} \quad (x > 0,\ \eta > 0,\ m > 0). \tag{1}$$

Jarque and Bera [4] derived a new test statistic for normality against Pearson family which includes normal, beta, gamma, Students's t and F distributions based on the Lagrange multiplier methods (LM test). This statistic is very simple to compute and asymptotically efficient. In this article we can show that we can construct the LM test for the Weibull distribution and also has good property especially for small samples. The general asymptotic theory supports asymptotic efficiency for Weibull case. Firstly, we have a complete sample without any censoring for the Weibull distribution $W(m, \eta)$. For a given set of n independent observations, say $x_1, x_2, \ldots, x_n$ their likelihood would be expressed by

$$\log L = \sum_{i=1}^{n} \log f(x_i; \theta).$$

Here we note that $\theta^T = (\eta, m)$.

Define $s_j = \frac{\partial \log L}{\partial \theta_j}$, $I_{jk} = E\left[-\frac{\partial^2 \log L}{\partial \theta_j \partial \theta_k}\right]$ for $j = 1, 2,\ k = 1, 2$, and $\theta_1 = \eta,\ \theta_2 = m$. Under the general conditions, the LM-test statistic is given by

$$LM = \hat{s}_2^T \left(\hat{I}_{22} - \hat{I}_{21}\hat{I}_{11}^{-1}\hat{I}_{12}\right)\hat{s}_2, \tag{2}$$

under $H_0 :\ m = m_0$, asymptotically as chi-square distribution with one degree of freedom, χ_1^2. $\hat{s}_j$ and $\hat{I}_{jk}$ are the corresponding plug-in estimate by replacing θ with its suitable constant estimate, $\hat{\theta}$. The derived LM-test statistic becomes

$$T_1 = \frac{6m_0^2}{\pi^2 n}\left\{\hat{\eta}^{-m_0}\left(\log\hat{\eta}\sum_{j=1}^{n} x_j^{m_0} - \sum_{j=1}^{n} x_j^{m_0}\log x_j\right)\right. \\ \left. + \frac{n}{m_0} - n\log\hat{\eta} + \sum_{j=1}^{n}\log x_j\right\}^2. \tag{3}$$

For comparison we describe the likelihood ratio statistic and the Wald statistic:

$$\lambda = \frac{\sup_{\eta} \prod_{i=1}^{n} \left(\frac{m_0}{\eta}\right)\left(\frac{x_i}{\eta}\right)^{m_0-1} \exp\left\{-\left(\frac{x_i}{\eta}\right)^{m_0}\right\}}{\sup_{m,\eta} \prod_{i=1}^{n} \left(\frac{m}{\eta}\right)\left(\frac{x_i}{\eta}\right)^{m-1} \exp\left\{-\left(\frac{x_i}{\eta}\right)^{m}\right\}}, \tag{4}$$

$$w = \frac{\pi\left(\hat{m} - m_0\right)}{\sqrt{\frac{6{m_0}^2}{n}}}. \tag{5}$$

Rinne([16],[18],[19]), introduced the table for testing an uncensored sample of size $n = 20$. The ML estimate of m is $\hat{m} = 2.5957$ and the test statistic for $H_0 : m = m_0 = 2$ is

$$\frac{\hat{m}}{m_0} = 1.2979.$$

Consulting their table of $\ell_1(n, \alpha)$ calculating the probabilities,

$$\Pr\{\hat{m}/m \leq \ell_1(n, \alpha)\} = \alpha,$$

we get $\ell_1(20,\ 0.95) = 1.449$ and then cannot reject null hypothesis. Our LM-test statistic is $1.694 < \chi_1^2(1 - 2 * 0.05) = 2.706$. The likelihood ratio statistic $-2\log\lambda$ is $2.088 < 2.706$ and the Wald statistic is $1.708 > 1.645$.

In the case of the special case of $m = 1$ (testing exponential distribution), the above test statistic becomes simply as follows:

$$T_1^* = \frac{6}{\pi^2 n}\left(n + \sum_{j=1}^{n} \log x_j - \frac{1}{\bar{x}}\sum_{j=1}^{n} x_j \log x_j\right)^2, \tag{6}$$

$$\lambda = \frac{\hat{\eta}^{n\hat{m}}}{\left(\bar{x}\,\hat{m}\right)^n \prod_{i=1}^{n} x_i^{\hat{m}-1}}, \tag{7}$$

and

$$w = \frac{\pi\left(\hat{m} - 1\right)}{\sqrt{\frac{6}{n}}}. \tag{8}$$

Here we note that the LM- test statistic is very simple to compute the statistic because of no need of likelihood calculation demanding numerical iterations. For previous data of White [19], $\hat{m}/m_0 = 2.596 > 1.449$, the LM-test statistic is 8.344, the likelihood ratio test statistic is 20.0179, and the Wald statistic is $9.152 > 1.645$. All tests succeed in rejecting exponentiality ($H_0 : m = 1$).

Mean parameter

The mean parameter is one of the most important parameters of distributions, but inference on the Weibull mean has not been studied so well especially in small samples. Harter and Dubey [3] created the vast table for testing Weibull mean parameter based on the Weibull-T statistics. They consider the two cases for testing $H_0 :\ \mu = \mu_0$. One case is that the standard deviation is known and another case is that the standard deviation is not known. Both cases require the information on the shape parameter m. Their example data set as regards 20 failure times in hours of electronic parts of an equipment is as follows: 154, 419, 590, 603, 770, 845, 848,891, 899, 953, 954, 982, 1044, 1059, 1126, 1127, 1294, 1678, 1831,1847. Here n=20, $\bar{x}$=995.7 hours, μ_0=886.2 hours and the sample standard deviation s=429.0 hours. Then we can get the Weibull-T statistic,

$$\frac{\sqrt{20}}{429.0}(995.7 - 886.2) = 1.142,$$

and the null hypothesis cannot be rejected at the 10 percent level of significance against alternative $H_1 : \mu > H_0 = 886.2$ consulting their table assuming $m = 2$. For deriving the LM-test statistic of the mean parameter μ we will express the distribution by two parameter m and μ:

$$f(x;\ m,\ \mu) = \left(\frac{m\Gamma_1}{\mu}\right)\left(\frac{\Gamma_1}{\mu}x\right)^{m-1}\exp\left[-\left(\frac{\Gamma_1}{\mu}x\right)^{m}\right]. \tag{9}$$

Here we use the following notations:

$$\Gamma_1 = \Gamma\left(1+\frac{1}{m}\right),\ \hat{\Gamma}_1 = \Gamma\left(1+\frac{1}{\hat{m}}\right).$$

After some calculations we have the following LM test statistic:

$$T_2 = \frac{\mu_0^{-2\hat{m}}}{\pi^2 n}\left(6H_{\frac{1}{\hat{m}}}^2 - 12H_{\frac{1}{\hat{m}}} + \pi^2 + 6\right)\left(n\,\mu_0^{\hat{m}} - \hat{\Gamma}_1^{\hat{m}}\sum_{j=1}^{n} x_j^{\hat{m}}\right)^2. \tag{10}$$

Here H_α is the harmonic function:

$$H_\alpha = \int_0^1 \frac{1-x^\alpha}{1-x}dx = \psi^{(0)}(1+\alpha) + \gamma.$$

The $\psi^{(0)}(\cdot)$ is the digamma function and γ is Euler's γ $(= 0.5771\ldots)$. The above test statistic requires the restricted ML estimate of parameter m under the restriction of $\mu = \mu_0$. Basically the LM-test statistic is designed for two-sided testing, and then we can use the two sided test setting 2α for significance level. The above data set [3] gives the following statistic value:

$$\hat{m} = 2.281,$$

$$T_2 = \ 1.356 < 1.642 = \chi_1^2(0.80).$$

In case of censored data no table is available, but we can use the LM-test extended to censored data which can expressed by followings:

$$T_3 = \frac{\hat{m}\mu_0^{-2\hat{m}}\left(d\mu_0^{\hat{m}} - \hat{\Gamma}_1^{\hat{m}}\sum\limits_{j=1}^{n} x_j^{\hat{m}}\right)^2}{(m+1)n - d - \frac{6\hat{m}^2(d+n(\gamma+\hat{\psi}_1^{(0)}-2))^2}{6d(\hat{m}-\hat{\psi}_1^{(1)})+n\left(\hat{m}\left(6(\gamma-2)\gamma+6\left(\hat{\psi}_1^{(0)}\right)^2+12(\gamma-1)\hat{\psi}_1^{(0)}+\pi^2\right)+6\hat{\psi}_1^{(1)}\right)}}. \tag{11}$$

$$\hat{\Gamma}_1 = \Gamma\left(1+\frac{1}{\hat{m}}\right),\ \hat{\psi}_1^{(l)} = \psi^{(l)}\left(1+\frac{1}{\hat{m}}\right),\ l = 0, 1,$$

The $\psi^{(1)}(\cdot)$ is the trigamma function.

Confidence interval

As for the confidence interval we can easily construct it without any effort to make it. All we need is just to calculate the testing statistic under some specified parameters and to solve the inequality including testing statistic:

$$\Pr\{T(m) \geq t(\alpha)\} \leq \alpha.$$

In case of LM-test statistic we will illustrate this as Fig. 1. For the test statistic function of m, say, $T(m)$, we solve the following nonlinear equation:

$$T(m) = \chi_1^2(1-\alpha).$$

In Fig. 1 two crossing points are depicted, which are corresponding to two confidence bounds, 1.645 and 3.355 for the White data appeared in previous sections. We note that this confidence interval (1.645, 3.355) does not include $m = 1$ and does include $m = 2$.

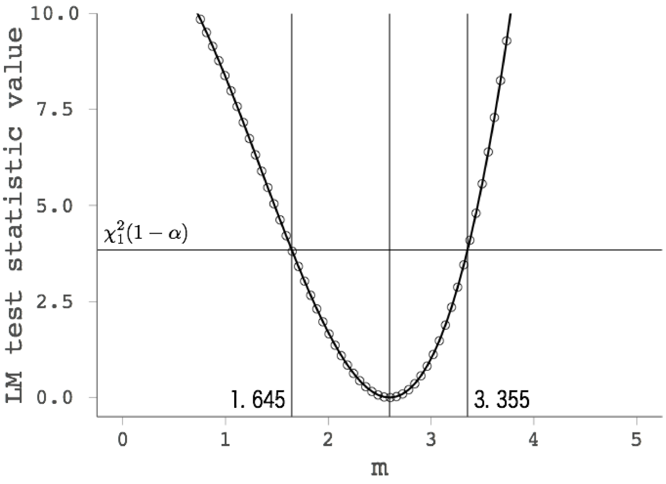


Figure 1: LM test statistic value and confidence interval

Simulation Studies

We conducted simulations studies with with $1,000,000$ number of replications for each test statistic.

The Table 1 shows that LM-test assures the significance levels in that it is always less than the specified α level. Other test statics will violate α-size test specification especially in case of small samples. For $n = 3$ we sometimes encounter faults in ML calculations, but we can calculate the LM test statistic because of needs of estimation for just one parameter. However we are required to estimate two parameters for other test statistics.

Other simulation results also make clear that the LM-test does not depend on the transformation of data or reparameterization such as log transformation of data

Table 1: True significance level estimated α by simulation in case of $H_0 : m = 1$ under true Weibull with shape parameter $m = 1$.

		nominal α size			
n	statistics	1%	2.5%	5%	10%
3	LM	0.00429	0.00764	0.01236	0.02071
	LR	-	-	-	-
	Wald	-	-	-	-
5	LM	0.00594	0.01063	0.01722	0.03549
	LR	0.03150	0.06187	0.10366	0.17340
	Wald	0.17927	0.21453	0.25003	0.29808
10	LM	0.00738	0.01372	0.02779	0.07243
	LR	0.01837	0.04071	0.07361	0.13384
	Wald	0.09265	0.12226	0.15478	0.20387
20	LM	0.00800	0.01804	0.03862	0.08687
	LR	0.01386	0.03229	0.06120	0.11592
	Wald	0.05076	0.07467	0.10383	0.15302
50	LM	0.00891	0.02190	0.04529	0.09493
	LR	0.01136	0.02764	0.05428	0.10639
	Wald	0.02630	0.04547	0.07228	0.12182
100	LM	0.00945	0.02318	0.04773	0.09737
	LR	0.01051	0.02614	0.05191	0.10332
	Wald	0.01799	0.03517	0.06100	0.11149

The notation “-” indicates all estimates are not available.

corresponding to extreme value distribution, and the orthogonalization of parameters by Cox and Reid [1].

Conclusions

We investigated testing and estimating for the Weibull models and we found that the LM-test statistic is simple and has good performance. In case of shape parameter with one the derived LM test is very simple without any likelihood calculations, which include iterations. We find that the LM test tends to assures α-size test in that the true significance levels are smaller than the predefined size of α for any sample size and the true α approaches the nominal α from undersides in accordance with sample size n larger.

Further investigations are required for the case of an LM-test version of Lawless type conditional inferences.

References

[1] D. R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference, 1987.

[2] H. W. Hager, L. J. Bain, and C. E. Antle. Reliability Estimation for the Generalized Gamma Distribution and Robustness of the Weibull Reliability Estimation for the Generalized Gamma Distribution and Robustness of the Weibull Model. *Technometrics*, Vol. 13, No. 3, pp. 547–557, 1971.

[3] H. L. Harter and S. D. Dubey. Theory and tables for tests of hypotheses concerning the mean and the variance of a Weibull population. *Wright-Patterson Air Force Base, Ohio: Aerospace Research Laboratories, Office of Aerospace Research*, Vol. 67-0059, , 1967.

[4] C. M. Jarque and A. K. Bera. A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, Vol. 55, No. 2, pp. 163–172, 1987.

[5] J. D. Kalbfleisch and R. L. Prentice. The statistical analysis of failure time data. *John Wiley & Sons, Inc.*, 1980.

[6] J. D. Kalbfleisch and R. L. Prentice. The statistical analysis of failure time data. *Canadian Journal of Statistics*, Vol. 10, No. 1, pp. 64–66, 1982.

[7] J. F. Lawless. Confidence intervals for the parameters of the Weibull distribution. *Utilitas Mathematica*, Vol. 2, pp. 71–87, 1972.

[8] J. F. Lawless. Conditional Versus Unconditional Confidence Intervals for the Parameters of the Weibull Distribution. *Journal of the American Statistical Association*, Vol. 68, No. 343, pp. 665–669, 1973.

[9] J. F. Lawless. Approximations to confidence intervals for parameters in the extreme value and Weibull distributions. *Biometrika*, Vol. 61, No. 1, pp. 123–129, 1974.

[10] J. F. Lawless. Confidence Interval Estimation for the Weibull and Extreme Value Distributions. *Technometrics*, Vol. 20, No. 4, pp. 355–364, 1978.

[11] D. N. Lawley. A General Method for Approximating To the Distribution of Likelihood Ratio Criteria. *Biometrika*, Vol. 43, pp. 295–303, 1956.

[12] E. H. Lloyd. Least-Squares Estimation of Location and Scale Parameters using Order Statistics. *Biometrika*, Vol. 39, No. 1, pp. 88–95, 1952.

[13] H. Moulton, L. A. Weissfeld, and R. T. St. Laurent. Bartlett correction factors in logistic regression models. *Computational Statistics & Data Analysis*, Vol. 15, pp. 1–11, 1993.

[14] P. B. Nagarsenker. On a test of equality of several exponential survival distributions. *Biometrika*, Vol. 67, No. 2, pp. 475–478, 1980.

[15] H. Nagatsuka, T. Kamakura, and N. Balakrishnan. A consistent method of estimation for the three-parameter Weibull distribution. *Computational Statistics & Data Analysis*, Vol. 58, pp. 210–226, February 2013.

[16] H. Rinne. *The Weibull Distribution: A Handbook*. CRC Press, November 2008.

[17] C. D. Sutton. Computer-Intensive Methods for Tests About the Mean of an Asymmetrical Distribution. *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 802–810, 1993.

[18] D. R. Thoman, L. J. Bain, and C. E. Antle. Inferences on the Parameters of the Weibull Distribution. *Technometrics*, Vol. 11, No. 3, pp. 445–460, 1969.

[19] J. S. White. The moments of log-Weibull order statistics. *Research Laboratories, General Motors Corporation*, 1967.

The Limit Distribution of the Maximum Value Test

Petr Philonenko and Sergey Postovalov
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: `petr-filonenko@mail.ru, postovalov@ngs.ru`

Abstract

The proposed maximum value test is a powerful test between the logrank test and Gehan′s Generalized Wilcoxon test. It is a useful test because these tests are preferable in different alternative hypotheses. To apply the maximum value test in the practice research for two-sample problem testing, a researcher should know a behavior of the test statistics distribution. It is necessary for a researcher to compute the p-value. In this paper, we research the statistics distribution of the maximum value test using the Monte-Carlo method.

Keywords: two-sample problem, lifetime data, logrank test, Gehan′s Generalized Wilcoxon test, maximum value test, Monte-Carlo method.

Introduction

Two-sample problem testing is one of necessary statistical procedures which helps a researcher to solve a problem, for example, to combine two samples or not. By using hypothesis testing, a researcher may face the problem of choosing between parametric tests and nonparametric tests and may face the problem of choosing a test into the groups. The choice in favor of a particular test depends on following factors:

- a distribution-free test (the calculation of the test statistic based on properties of all distributions of data);
- a power of the statistical test (higher this value, higher the probability that the alternative hypothesis will be rejected when the alternative hypothesis is false);
- existence of the limit distribution of test statistics (if the condition is not satisfied, a researcher should simulate the test statistics distribution with certain modeling parameters for hypothesis testing);
- a convergence rate of the test statistics distribution (if the test statistics distribution under a small sample size is not different against the limit statistics distribution, then there is no need to apply computational modelling methods for two-sample problem testing).

In practice, it is necessary to use a distribution-free test, with a high power and a fixed limit statistics distribution (a not simulated distribution with certain modelling parameters).

However, there is not the most powerful test generally, therefore one uses methods of increasing the test power using various strategy, for example:

- using a selector statistic among values of other test statistics;

- using the certain test in the certain alternative hypothesis when a researcher knows the more powerful test;
- etc.

The most commonly encountered in the literature and well-studied tests for two-sample problem testing with lifetime data (the object of observation during the investigation may be failed) are logrank [1] test and generalized Wilcoxon test (for example, Peto & Peto [2], Gehan [3]). In the literature, there are different information about these tests. In Lee [4] states the logrank test is more powerful than the Generalized Wilcoxon tests if the hazard ratio is constant (for example, samples observations are distributed exponentially). On the other hand, the Generalized Wilcoxon tests are more powerful than the logrank test if the hazard ratio is not constant. In other papers [5, 6] states that the logrank test sensitives better differences in late time than the Generalized Wilcoxon tests but the Generalized Wilcoxon tests sensitive better differences in early time than the logrank test and this sensitivity of the tests does not depend on the hazard ratio.

In Section 1, we present the maximum value test and the equation of the limit distribution in general case. In Section 2, we formulate the problem and present results.

1 Statistics

In this paper, we do not present statistics of the logrank test and Gehan′s Generalized Wilcoxon test because these tests have been considered in [6] in detail. We should just know that S_G is a statistic of Gehan′s Generalized Wilcoxon test, S_L is a statistic of the logrank test and both statistics are standard normal distributed with the two-sides critical areas.

1.1 The Maximum Value Test

Thus, using statistics of these tests, we proposed following statistical procedure [7] of two-sample problem testing with lifetime data:

$$S_{MAX} = \max\left(|S_G|, |S_L|\right),$$

where S_{MAX} is a value of the proposed maximum value test, S_G is a value of the Gehan Generalized Wilcoxon test and S_L is a value of the logrank test.

The main idea of the maximum value test is to use a statistic value corresponding to smaller p-value between two tests.

The random variable, that is a maximum of two standard normal absolute values, is distributed with the following probability distribution function:

$$f(x;r) = \varphi(x)\left(\Phi_0\left(x\sqrt{\frac{1-r}{1+r}}\right) + \Phi_0\left(x\sqrt{\frac{1+r}{1-r}}\right)\right),\ x \geq 0, \tag{1}$$

where $\Phi_0(x) = \int\limits_0^x \varphi(t)dt = \frac{1}{\sqrt{2\pi}} \int\limits_0^x e^{-\frac{t^2}{2}} dt$ and r is a correlation value between distributions of S_G and S_L under null hypothesis.

2 Simulation

The proposed statistical test is a distribution-free test because this test is a nonparametric test. Moreover, the test power is close of a maximum power value between the Generalized Wilcoxon test and the logrank test. One can find such results in [7]. Also, statistics distributions of these tests have high convergence rates for their limit distributions [8], respectively. Now, we should make sure that a researcher may freely apply the limit distribution of the maximum value test in the practice research and a researcher should not simulate the limit distribution with certain parameters. Taken into account the fact that the value of S_{MAX} is a maximum of two absolute standard normal distributed values (which are computed by on the same samples), we conclude there is a correlation between these random variables. We have investigated the effect of the alternative hypothesis on a correlation estimation value. For this purpose, we have applied closest alternative hypotheses using various distributions with following probability density functions: the family of Weibull distributions, the family of Gamma distributions, the family of Lognormal distributions and the family of Exponential distributions.

The results of simulation are the correlation estimation values for the various alternative hypotheses, for various sample sizes (20-500 observations), for various censored rates (0-50%), for various censored time distributions (Weibull and Gamma distributions).

Having received the correlation estimation values, we have selected the maximum $r_{max} \approx 0.94$ and minimum $r_{min} \approx 0.86$ values. For both values, we have constructed cumulative probability functions using the equation (1) and then assessed the maximal difference between them. As a result, we have obtained that under the test size $a < 0.15$ the maximal difference is less 0.01. One can see it on Fig 1. Such a result allows us to conclude this equation (1) may be applied practically.

Conclusions

Two-sample problem testing is one of fundamental hypotheses in statistics. Such testing is applied in industry, sociology and so on. The proposed method (the maximum value test) is a powerful test between Gehan′s Generalized Wilcoxon test and the logrank test and, as follows from the paper, the maximum value test has the statistics distribution which is applicable to calculate a p-value for two-sample problem testing without the statistics distribution simulation. The error of p-value calculation is not more than 0.01.

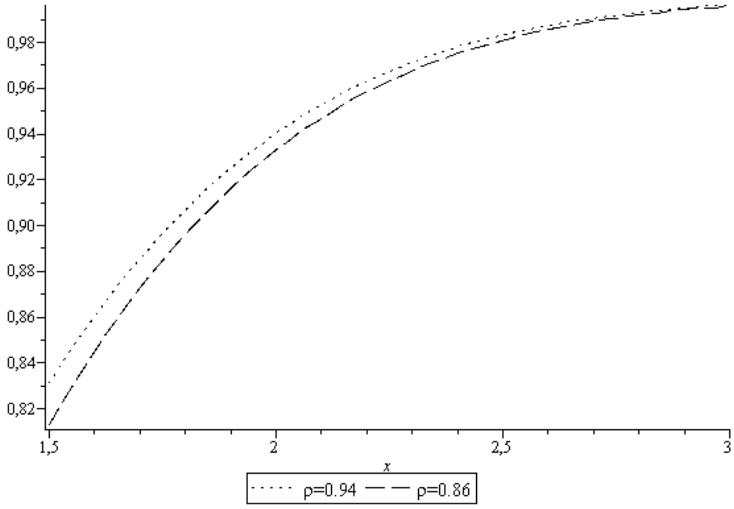


Figure 1: limits distributions of the maximum value test with the maximum r_{max} and the minimum r_{min} simulated correlation values

References

[1] Mantel, N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. Cancer Chemotherapy Reports, 50, 163-170.

[2] Peto, R., and Peto, J. (1972). Asymptotically Efficient Rank Invariant Procedures. Journal of the Royal Statistical Society, Series A, 135, 185-207.

[3] Gehan, E. A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika, 52, 203-223.

[4] Lee ET, Wang JW. (2003). Statistical methods for survival data analysis. 3rd ed. Hoboken (NJ): John Wiley & Sons. doi:10.1002/0471458546.index

[5] Ruvie Lou Maria C. Martinez (2010). A pretest for choosing between logrank and wilcoxon tests in the two-sample problem / Ruvie Lou Maria C. Martinez, Joshua D. Naranjo // International Journal of Statistics,vol. LXVIII, n. 2, 2010. – 111 – 125 pp.

[6] P. Philonenko (2013). A Comparison of Homogeneity Tests for Different Alternative Hypotheses / S. Postovalov, P. Philonenko // Statistical Models and Methods for Reliability and Survival Analysis : monograph. - London : Wiley-ISTE, 2013. – Chap. 12. – P. 177-194. – (Mathematics and Statistics series).

[7] Petr Philonenko and Sergey Postovalov (2014): A new two-sample test for choosing between log-rank and Wilcoxon tests with right-censored data, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2014.941533

[8] Petr Philonenko (2012) Two-sample problem testing with right-censored data. // Nauka. Tekhnologii. Innovatsii – 2012 // Vol. 1. Novosibirsk. In Russian

An Adaptive Method for Selecting an Optimal Bandwidth Parameter in Nonparametric Estimate of the Conditional Reliability Function[1]

Victor A. Demin, Ekaterina V. Chimitova
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: chimitova@corp.nstu.ru

Abstract

In this paper, we have considered the most popular approach to nonparametric estimation of conditional reliability function, proposed by Beran. We give the analysis of dependence of an optimal bandwidth parameter from the value of covariate, for which the conditional reliability function is estimated. An adaptive method for selecting the optimal bandwidth parameter is presented in this paper. The accuracy of the proposed method has been studied depending on the plan of experiment(the sample size and the number of groups).

Keywords: conditional reliability function, nonparametric Beran estimator, optimal bandwidth parameter.

Introduction

One of the most important parts of the statistical analysis of lifetime data in reliability and survival analysis is studying the dependence of the reliability function on the observed explanatory variables, which are also called covariates. Characteristics of objects themselves, such as the type of material or age of patient, or external stresses, such as temperature, pressure or treatment strategy, can be taken as covariates. There are two types of approaches in statistical data analysis: parametric and nonparametric models. The most widely used parametric regression models in reliability and survival analysis are the accelerated failure time (AFT) model and the proportional hazards (PH) model. However, the parametric approach requires knowledge of the functional dependence of reliability function on covariates and the lifetime distribution. In practice, this information is usually absent. In such situations, it is necessary to use nonparametric methods.

One of the most popular nonparametric methods for estimation of conditional reliability function is the Beran estimator [1]. The Beran estimator allows using all information of sample with covariates. The properties of this estimator were investigated by a number of mathematicians. The investigation of properties in the case of random plan, when the values of covariate are not fixed, was presented in [2]-[5]. In [6], the properties of the Beran estimator were studied, when the covariate is random.

The application of the Beran estimator in practice is difficult, because there is no method for selection of bandwidth parameter in the case of determinant plans of

[1] This work is supported by the Russian Ministry of Education and Science (project 2.541.2014K).

experiment. In [7], an asymptotical method for selection of the optimal bandwidth parameter is suggested, but it is impossible to implement this method in practice as it requires to known the form of reliability function. In [10], the bandwidth parameter was chosen empirically. However, this parameter plays the key role in the accuracy of the Beran estimate. In our previous paper [11], we proposed a method for selecting an optimal bandwidth parameter, and in [13], we investigated the statistical properties of the Beran estimator with the usage of this method. However, this approach allows choosing one value of parameter for the whole sample, regardless the value of covariate. At the same time, the behavior of the conditional reliability function can vary greatly depending on the value of covariate. Thus, it is necessary to develop an adaptive method for selecting the optimal bandwidth parameter depending on the value of covariate, for which the conditional reliability function is estimated.

1 Nonparametric Beran estimator

The main feature of lifetime data is the presence of right censored observations, which can be represented as

$$(Y_1, x_1, \delta_1), (Y_2, x_2, \delta_2), \ldots, (Y_n, x_n, \delta_n),$$

where n is the sample size, x_i is the value of covariate for i-th object, Y_i is the failure time or censoring time and δ_i is the censoring indicator, which is equal to 1, if the i-th observation is complete, and 0 if it is censored. The nonparametric Beran estimator allows estimating a conditional reliability function and is defined as follows [1]:

$$\tilde{S}_{b_n}(t|x) = \prod_{Y_{(i)} \le t} \left\{ 1 - \frac{W_n^i(x; b_n)}{1 - \sum_{j=1}^{i-1} W_n^j(x; b_n)} \right\}^{\delta_i} \tag{1}$$

where x is the value of the covariate, for which reliability function is estimated, $W_n^i(x; b_n)$, $i = 1, \ldots, n$ are the Nadaraya-Watson weights, which are defined as follows [5]:

$$W_n^i(x; b_n) = K\left(\frac{x - x_i}{b_n}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x - x_j}{b_n}\right),$$

where $K\left(\frac{x-x_i}{b_n}\right)$ is the kernel function, satisfying to the regularity conditions: $K(y) = K(-y)$, $0 \le K(y) < \infty$, $\int_{-\infty}^{\infty} K(y)dy = 1$, $b_n > 0$ is the bandwidth parameter which satisfies to $\lim\limits_{n\to\infty} b_n = 0$, $\lim\limits_{n\to\infty} nb_n = \infty$.

In our previous papers [11], it was shown that the accuracy of the Beran estimation strongly depends on the bandwidth parameter. Moreover, we proposed the method for selection an optimal bandwidth parameter, which is based on the minimization of the mean deviation of failure times $Y_1, Y_2, ..., Y_n$ from the nonparametric estimate of the inverse reliability function $S_x^{-1}(p)$. According to this method, an optimal bandwidth parameter can be obtained by solving the optimization problem [11]:

$$b_n^{opt} = \arg\min_{b_n} \sum_{i=1}^{n} \delta_i \cdot \left| \hat{g}\left(\hat{p}_i | x_i\right) - Y_i \right|. \tag{2}$$

where $\hat{g}(\hat{p}_i|x_i)$ is the estimate of the inverse reliability function:

$$\hat{g}\left(\hat{p}_i | x_i\right) = \frac{1}{n} \sum_{j=1}^{n} \omega_n^j \left(\hat{p}_i\right) \cdot Y_j, \tag{3}$$

where ω_n^j are the Priestley-Chao weights of the second order [8, 9]:

$$\omega_n^j\left(\hat{p}_i\right) = \left\{\hat{p}_{(i)} - \hat{p}_{(i-1)}\right\} K\left(\frac{\hat{p}_i - \hat{p}_j}{h_{NS}}\right),$$

where h_{NS} is the smoothing parameter, which is estimated by the minimal mean integrated error method [12]:

$$h_{NS} = \left[\frac{8\pi^{1/2} R(K)}{3\mu_2(K)^2 n}\right]^{1/5} \hat{\sigma},$$

where $\mu_2(K) = \int x^2 K(x) dx$, $R(K) = \int K^2(x) dx$, $\hat{\sigma}$ is the robust estimate of the variance:

$$\hat{\sigma}_{robust} = \underset{i=1..n}{med} \left| \hat{p}_i - \underset{j=1..n,\, k=j..n}{med} \left(\frac{\hat{p}_j + \hat{p}_k}{2}\right) \right|.$$

As it is seen from (3), this method allows choosing only one parameter for each conditional reliability function. It doesn't take into account the value of covariate, for which the conditional reliability function is estimated. However, an optimal bandwidth parameter depends on the value of covariate. Let us consider some examples to show this dependence.

The investigation of the properties of a bandwidth parameter is carried out by the Monte Carlo simulations. As the true reliability model, we consider the parametric Cox proportional hazards model [10]:

$$S_x(t) = \left(S_0(t)\right)^{r(x;\beta)}, \tag{4}$$

with the covariate function $r(x; \beta) = \ln(1 + e^{\beta x})$ and the lognormal baseline distribution with the density function

$$f_0(t) = \frac{1}{\sqrt{2\pi}\theta_1 t} \exp\left(-\frac{1}{2\theta_1^2} \ln^2\left(\frac{t}{\theta_2}\right)\right)$$

with parameters $\theta_1 = 21.5$, $\theta_2 = 1.6$.

Let us consider two plans of experiment: the covariate in the first case takes the values from the set $\{0,\ 0.33,\ 0.67,\ 1\}$ and in the second case – from the set $\{0,\ 0.11,\ 0.22,\ 0.33,\ 0.44,\ 0.56,\ 0.67,\ 0.78,\ 0.89,\ 1\}$, the sample sizes are $n = 20, 40, 80, 120, 200, 300, 400$, the number of observations corresponding to different values of the covariate is equal to each other. Let m denotes the number of groups corresponding to different values of covariate.

The true value of the optimal bandwidth parameter is defined as:

$$b^j_{true} = \arg\min_{b^j} \left(\sup_{t<\infty} \cdot \left| \tilde{S}_{b^j}(t|x) - S_{x_j}(t) \right| \right). \tag{5}$$

In Figures 1, 2, the values of the optimal bandwidth parameter (5) are shown for the considered plans of experiment.

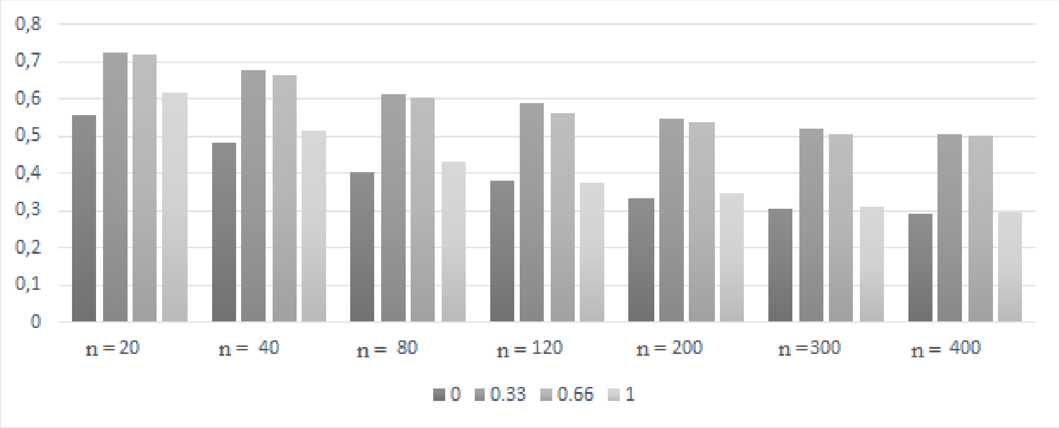


Figure 1: The value of bandwidth parameter b^j_{true} for 4 different values of covariate and different sample sizes.

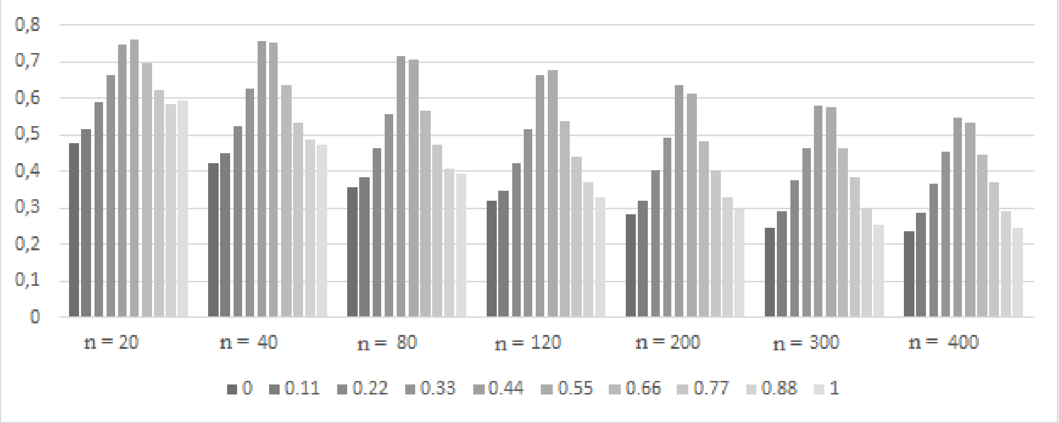


Figure 2: The value of bandwidth parameter b^j_{true} for 10 different values of covariate and different sample sizes.

As it is seen from Figures 1-2, an optimal bandwidth parameter depends on the value of covariate, for which the reliability function is estimated. The values of optimal bandwidth parameter are different for different values of covariate. These differences can be significant for different groups. For example, the optimal bandwidth parameters for extreme values of covariate differ by 80-100% from the optimal bandwidth parameters for average values of covariate. In the case of 4 groups, the difference between optimal bandwidth parameters for adjacent values of covariate can be between 3% and 70%. In the case of 10 groups, this difference can be between 3% and 21%. Moreover, as can be seen from Figure 1-2, for the considered reliability model, the optimal bandwidth parameter is symmetrical relative to the average value of covariate in the case of symmetric plans. Thus, it is sufficient to choose the optimal bandwidth parameters only for a half of the values of covariate, for another half the optimal bandwidth parameters will be the same. The dependence of the optimal bandwidth parameter on the number of groups has been confirmed. For example, for the sample size $n = 80$ and 4 groups, the value of optimal bandwidth parameter is equal to 0.61 for the value of covariate $x = 0.33$, however, for the same sample size and the same value of covariate, but in the case of 10 groups, the optimal bandwidth parameter is equal to 0.55.

So, it is necessary to develop an adaptive algorithm for selecting the optimal bandwidth parameter for the Beran estimator.

2 An adaptive selection of bandwidth parameter

To develop an adaptive algorithm, it is possible to change the old one. The key change is to minimize function (3) only for the failure times of items, observed under the given value of the covariate x. Thus, the optimal bandwidth parameter can be obtained by solving the following optimization problem:

$$b_{opt}^{j} = \arg\min_{b_n} \sum_{i=1}^{n_*} \delta_i \cdot \left| \hat{g}\left(\hat{p}_{b_n}(Y_i|x)\right) - Y_i \right|, \tag{6}$$

where n_* is the number of observations, corresponding to the value of covariate x, $\hat{p}_{b_n}(Y_i|x)$ is the Beran estimate of the conditional reliability function under the covariate x. Nonparametric estimator of the inversed reliability function is obtained by kernel smoothing for all observations:

$$\hat{g}\left(\hat{p}_{b_n}(Y_i|x)\right) = \frac{1}{n} \sum_{i=1}^{n} \omega_i\left(\hat{p}_{b_n}(Y_i|x)\right) \cdot Y_i, \tag{7}$$

where the Priestley-Chao weights are built only for the value of covariate x, but for all observations:

$$\omega_i\left(\hat{p}_{b_n}\right) = \left\{\hat{p}_{(i)} - \hat{p}_{(i-1)}\right\} K\left(\frac{\hat{p}_{(i)} - \hat{p}_{b_n}}{h_n}\right),$$

where h_n is the smoothing parameter.

Let us consider, how the algorithm selects an optimal bandwidth parameter for each group. We compare results of proposed algorithm with the true bandwidth parameter (5). In Figures 3-4, the differences between parameters (5) and estimated parameters (6) in percent are shown. In Figure 3, the distances are shown for the sample sizes n=20, 40, 80, 120, 200, 300, 400 and 4 groups, and in Figure 4 – for sample sizes n=20, 40, 80, 120 and 10 groups.

In this research, we have estimated the optimal bandwidth parameter for the values of covariate $x = 0$ and $x = 0.33$ in the case of the plan with 4 groups and for $x = 0{,}0.11,\ 0.22,\ 0.33,\ 0.44$ in the case of the plan with 10 groups. As the optimal bandwidth parameter is symmetrical relative to the average value of covariate in the case of symmetric plans, we have taken the values of bandwidth parameter for the other groups equal to the obtained estimates corresponding to symmetric groups.

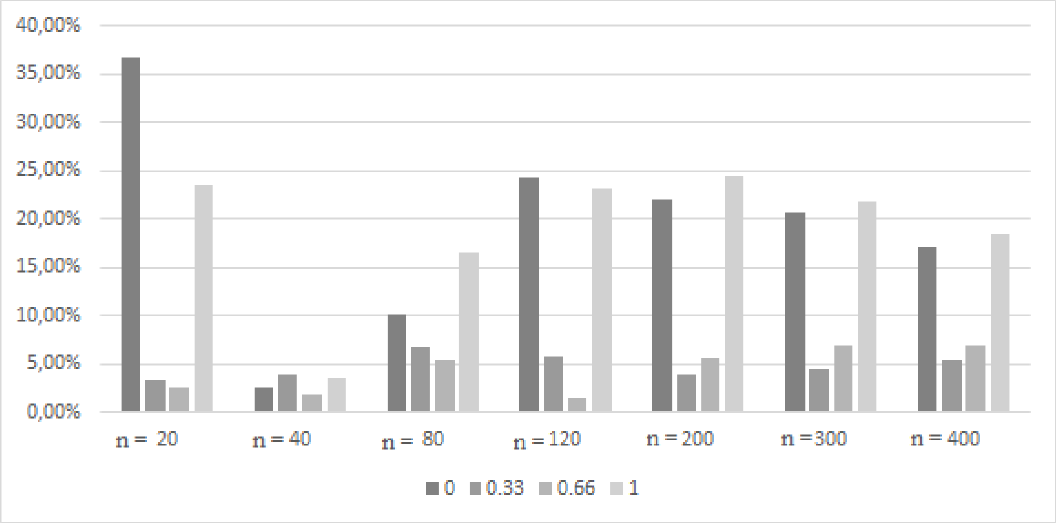


Figure 3: The deviation in percent of estimated parameters (6) from the true parameters (5) for 4 different values of covariate and different sample sizes

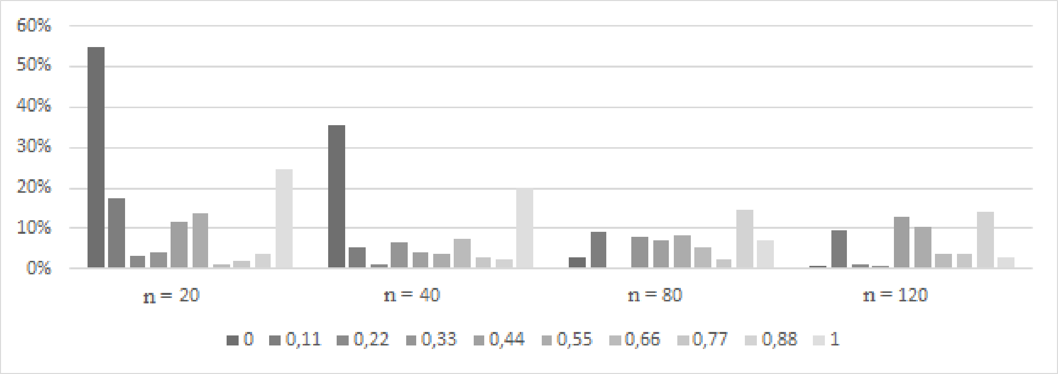


Figure 4: The deviation in percent of estimated parameters (6) from the true parameters (5) for 10 different values of covariate and different sample sizes

As it is seen from Figure 3, the proposed method allows getting sufficiently accurate values of bandwidth parameter. For example, the maximal distance between estimated and true parameter is 37% and the number of observations for one group in this case is only 5. If there are more observations in a group the maximum deviation is not larger than 24% for the extreme groups and not larger than 7% for the central groups.

The large deviations in the extreme groups for the sample of 120-400 observations can be explained by the fact that true values of the optimal bandwidth parameter for these sizes are almost always less or equal than 0.33, which means that the Beran estimation degenerates into the Kaplan-Meier estimation. In this case, any value from the interval (0, 0.3(3)] can be taken as the optimal parameter. In all cases, the optimal bandwidth parameter for the extreme groups and sample sizes $n \geq 120$ turned out to be less than 0.33. It means that the method chooses the best parameter in these cases.

From Figure 4, we can see a similar regularity: the worst estimates of the optimal bandwidth parameter are obtained for extreme groups and the smallest sizes. With the sample size growth, the accuracy of estimates for extreme groups increases. And already for 80 observations in a sample, the deviation of estimated parameters (6) from the true parameters (5) does not exceed 15%.

It is necessary to note, that the sample size growth may not lead to the increase of the accuracy of selecting the optimal bandwidth parameter, since with the sample size growth, not only the number of elements in considered group increases, but also the number of elements in other groups increases too, what “clogs” the sample. This remark explains the fact that for the extreme values of covariate, an accuracy of estimating the optimal bandwidth parameter increases with the sample size growth, since they have less clogging observations.

Conclusions

It has been shown that an optimal bandwidth parameter depends on the value of covariate, for which the conditional reliability function is estimated. In this paper we have proposed an adaptive method for selecting an optimal bandwidth parameter for the Beran estimator in the case of a determinant plan of experiment. This method is a modification of the previously proposed method for bandwidth selection.

It has been shown, that the proposed adaptive method allows getting estimates of optimal bandwidth parameter, which are close to the true optimal parameter for any group of a plan, even for small sample sizes. At this moment the new method has been studied only for symmetric plans. To use proposed algorithm for any plans of experiment, it is necessary to carry out further investigations.

References

[1] Beran R. (1981). *Nonparametic regression with randomly censored survival data.* Technical report. Department of Statistics, University of California, Berkeley.

[2] Dabrowska D.M. (1992). *Nonparametric quantile regression with censored data.* Sankhya Ser. A. P. 252-259

[3] Gonzalez M.W., Cadarso S.C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some application. *J. Nonparametric Statistics*. Vol. **4** pp. 65-78.

[4] McKeague I.W., Utikal K.J. (1990). Inference for a nonlinear counting process regression model . *Ann. Statist.* Vol. **18**, pp. 1172-1187.

[5] Van Keilegom I., Akritas M.G., Veraverbeke N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics and Data Analysis*. Vol. **35**, pp. 487-500.

[6] Akritas M.G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*. Vol. **22**, pp. 1299-1327.

[7] Van Keilegom I., Veraverbeke N. (1998). Nonparametric Estimation of the Conditional Distribution in Regression with Censored Data. *Dissertation*. pp. 7-51.

[8] Racine J.S. (2008). Nonparametric econometrics: a primer. *Quantile*. Vol. **4**, pp. 7-56.

[9] Hardle W., Malyutov M.B. (1993). Applied nonparametric regression. *Mir Publ.* P. 349.

[10] Li G., Datta S. (1999). A bootstrap approach to nonparametric regression for right censored data. *Technical Report*. pp. 1-10.

[11] Demin V.A., Chimitova E.V. (2012). Choice of optimal smoothing parameter for nonparametric estimation of regression reliability model. *Tomsk state university Journal of control and computer science* . Vol. **1(22)**, pp. 50-59.

[12] Rousseeuw P.J., Verboven S. (2002). Robust estimation in very small samples. *Journal Computational Statistics and Data Analysis*. Vol. **40(4)**, pp. 741-758.

[13] Demin V.A., Chimitova E.V., Schekoldin V.Yu. (2014). The research of optimal choice method of bandwidth parameter for nonparametric estimation of reliability regression models. *Tomsk state university Journal of control and computer science* . Vol. **2(27)**, pp. 10-19.

The Construction of Accelerated Life Model for Reliability of a Cutting Tool[1]

Ekaterina V. Chimitova, Mariia A. Semenova, Vitaliy S. Karmanov, and Gennadiy I. Smagin

Novosibirsk State Technical University, Novosibirsk, Russia

e-mail: chimitova@corp.nstu.ru

Abstract

In this paper, we have constructed the parametric accelerated failure time model for the reliability of a drilling tool basing on the results of drill reliability experiment. The second-order polynomial model was considered as the covariate function, and the baseline reliability function was taken from the family of Weibull distributions. The goodness-of-fit of the model was tested with the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling type tests by the sample of residuals. Using the constructed model, we have found the optimal cutting conditions (the feed rate and the rotational speed), which minimize the economical costs. Moreover, the failure-free operation times have been calculated for various probabilities under obtained optimal cutting conditions.

Keywords: accelerated failure time model, reliability of a cutting tool, feed rate, rotational speed, goodness-of-fit testing.

Introduction

In [5], [10] – [12], the application of mathematical models for the reliability of a cutting tool has been discussed by the example of drilling. Such models can be used for the optimization of processing conditions and for the optimal design of experiment. Usually, the lifetime of a drill is characterized by total length of holes, which were made by the tool before its blunting. The lifetime of a drill depends on two factors, such as the feed rate and the rotational speed.

In [5], various mathematical models for the dependence of the drill lifetime on these factors, for example, the second-order polynomial model, the Konig-Depiereux model and the exponential model, have been considered. However, the fact that we need to construct some probabilistic model for reliability, because the drill lifetime is a random variable, was not taken into account in mentioned papers.

The parametric accelerated failure time model is one of the most used probabilistic models in reliability. This model allows to calculate all reliability indices for various values of explanatory variables in the model. In [1, 4], the problems of parameter estimation and testing goodness-of-fit of the parametric accelerated failure time model have been considered for various designs of experiments.

Thus, the main goal of the paper is to define the optimal drilling conditions (optimal values of the feed rate and the rotational speed) and to compute the reliability indices of a cutting tool on the basis of parametric accelerated failure time model.

[1] This work is supported by the Russian Ministry of Education and Science (project 2.541.2014K).

1 Accelerated failure time model

Let us denote by L the non-negative random variable, which defines the total length of holes, made by the tool before failure (blunting). The reliability function is defined in the following form [1]:

$$S(t) = \mathrm{P}\{L > t\} = 1 - F(t), \tag{1}$$

where $F(t)$ is the corresponding distribution function.

In the general case, the object reliability depends on some characteristics of this object and experiment conditions. The effect of these factors on reliability is taken into account by the vector of explanatory variables (covariates) $x = (x_1, x_2, ..., x_m)^T$. The domain of each covariate in the vector is defined by the experiment conditions.

One of the most used reliability models is the AFT-model (Accelerated Failure Time model). In the case of constant covariates, which are not dependent on time, the parametric reliability AFT-model has the form

$$S_x(t) = S_0\left(\frac{t}{r(x;\beta)}\right), \tag{2}$$

where $S_0(t) = 1 - F_0(t)$ is the baseline reliability function and $r(\cdot)$ is the non-negative covariate function.

Let us construct the AFT-model basing on the results of the drill reliability experiment, presented in [5], for various values of the feed rate x_1 and the rotational speed x_2. This sample of lifetimes has 50 elements, which are given in Table 1.

Table 1: The data set of the drill reliability experiment. The lifetime L, mm (drill d4.2 “R6M5”, steel “1X18N9T”, cutting fluid “NGL-205”, take-off drill 10d). The detail has type “bare lattice”, 686 reach-through holes of 20 mm.

x_1, mm/rot	x_2, rot/min				
	750	1098	1447	1795	2145
0.0280	570	1430	3600	1400	430
	390	1370	1800	1200	250
0.0450	5560	8300	4700	4000	700
	8500	6300	5700	3000	1100
0.0621	4040	5800	6130	3330	590
	5640	7800	4230	4070	810
0.0790	3150	3420	2760	1350	470
	3850	4180	3800	1650	690
0.0962	1910	130	100	30	9
	3170	150	140	50	11

It should be noted that each observation in the sample is a failure, i.e. the lifetime sample is complete (we have no censored observations in the sample).

As shown in [5], the most preferable model for the dependence of drill lifetime on considered factors is the second-order polynomial model. Therefore, we use the covariate function in the following form

$$r(x;\beta) = \exp\left\{\beta_1 x_1^2 + \beta_2 x_1 + \beta_3 x_2^2 + \beta_4 x_2 + \beta_5 x_1 x_2\right\}. \tag{3}$$

The baseline distribution was chosen as the Weibull distribution with reliability function

$$S_0(t) = \exp\left\{-\left(\tfrac{t}{\theta_1}\right)^{\theta_2}\right\}, \theta_1, \theta_2 > 0.$$

It is necessary to note here, that the AFT-model with the Weibull baseline distribution is equivalent to the popular proportional hazards Cox model [3].

Table 2 shows the obtained maximum likelihood estimates of parameters of the AFT-model, values of the Wald statistic and corresponding p-values for testing insignificance of regression parameters.

Table 2: The parameter estimation of AFT-model

Parameter	Estimation	Wald statistic	p-value
θ_1 (scale)	0.2085	–	–
θ_2 (form)	1.7382	–	–
$\beta_1(x_1^2)$	-1971.5207	50.92	0.0000
$\beta_2(x_1)$	281.6922	46.51	0.0000
$\beta_3(x_2^2)$	-1.3E-06	8.27	0.0040
$\beta_4(x_2)$	0.0049	11.87	0.0006
$\beta_5(x_1 x_2)$	-0.0414	18.06	0.0001

As it is seen from Table 2, the obtained AFT-model is statistically significant. However, we need to test the goodness-of-fit hypothesis before using this model for computation of reliability indices and optimization of cutting conditions. To test goodness-of-fit we have used the sample of residuals, which can be written as follows

$$R_i = \tfrac{X_i}{r\left(x^i;\hat{\beta}\right)}, i = 1, ..., n,$$

where X_i is the failure time for i-th object, x^i is the values of covarite vector for i-th object, and $\hat{\beta}$ is the vector of maximum likelihood estimates of model regression parameters.

If the tested model is indeed correct, the residuals should fit the baseline distribution. The composite hypothesis of goodness-of-fit of the Weibull distribution by the sample of residuals can be tested with the Kolmogorov, Cramer-von Mises-Smirnov, and Anderson-Darling type tests [2, 4].

The data set has no censored observations, therefore, we can use the obtained in [7, 8, 9] models as the limiting distributions for goodness-of-fit test statistics. In Table 3, the values of statistics of the Kolmogorov, Cramer-von Mises-Smirnov, and Anderson-Darling type tests and corresponding p-values are shown.

Table 3: Goodness-of-fit testing of AFT-model

Test	Statistic	p-value
Kolmogorov	0.498	0.823
Cramer-von Mises-Smirnov	0.042	0.639
Anderson-Darling	0.344	0.499

Thus, Table 3 shows that the hypothesis of goodness-of-fit of the Weibull AFT-model with covariate function (3) is not rejected for significance level $\alpha = 0.05$. Therefore, this model can be used for computation of reliability indices and optimization of cutting conditions.

2 Optimization of cutting conditions

Optimal cutting conditions can be found by solving the following optimization problem:

$$Q(x_1, x_2) \to \min_{x_1, x_2}, \tag{4}$$

where Q is the value of economic costs (in rubles per a detail) and has the following form

$$Q = \frac{C}{x_1 x_2} + \frac{D}{\bar{L}\left(x_1, x_2\right)}, \tag{5}$$

where C, D are values, which depend on chosen optimization criteria and considered costs [5], [10] – [13].

The obtained Weibull AFT-model with covariate function (3) was used to calculate the mean time between failures, which can be written as

$$\bar{L}\left(x_1, x_2\right) = \theta_1 r\left(x; \beta\right) \Gamma\left(1 + \tfrac{1}{\theta_2}\right),$$

where $\Gamma(\cdot)$ is the Euler gamma-function.

The values of C and D from (5) have the following form

$$C = C_3 L_0, \;\; D = C_3 t_u L_0 + \tfrac{C_p L_0}{K} + L_0 C_3' t_3.$$

Table 4 shows the values of parameters, which are necessary for calculation of C and D.

By minimizing (5) with calculated values C and D we obtain the optimal cutting conditions in the form

$$x_1^* = 0.066, \;\; x_2^* = 1539, \;\; Q^* = 254.$$

Here, mean time between failures $\bar{L} = 5281$ mm. Thus, the feed rate equal to 0.066 mm/rot and the rotational speed equal to 1563 rot/min provide the optimal cutting conditions, for which economic costs are minimal and equal to 254 RUB.

Table 4: Economic parameters

Parameter	Description	Value
C_3	salary and overhead costs of driller	1.5 RUB/min (15840 RUB/mon)
C_3'	salary and overhead costs of grinder	1.5 RUB/min (15840 RUB/mon)
L_0	total length of holes for one diameter	12720 mm
t_u	tool-changing time	3 min
t_3	tool-grinding time	7 min
C_p	tool cost	50 RUB
K	number of drill grinding before failure	4

In addition, we have calculated the times before failures for various values of reliability function P using the constructed AFT-model. Table 5 contains the results of this calculation for covariate values $x_1 = 0.066,\ \ x_2 = 1539$, which correspond to optimal cutting conditions.

Table 5: Times before failure of a cutting tool

P	0.25	0.50	0.75	0.90	0.95
L, mm	6251	4194	2528	1418	937

Thus, the obtained values of the drill reliability, presented in Table 5, can be used to predict the tool degradation, to plan replacement of the tools before failures, and, finally, to reduced costs of detail processing.

Conclusions

In comparison with the standard determinate models for the drill lifetime, the application of probabilistic reliability models (AFT-model, for example) allows to compute the reliability indices, to predict the failures and to plan preventive replacements. This possibility is increasingly important with improving quality and costs of cutting tools.

The construction of the AFT-model described in the paper, the goodness-of-fit testing, the optimization of cutting conditions to minimize economical costs, and computation of probability indices can be helpful for optimization some other technological processes. The proposed methods can be also used for other metal working, such as turning, milling, and unrolling.

References

[1] Bagdonavicius V., Nikulin M. Accelerated Life Models. Modeling and Statistical Analysis. Chapman&Hall/CRC, 2002. 360 p.

[2] Balakrishnan N., Chimitova E., Galanova N., Vedernikova M. Testing goodness of fit of parametric AFT and PH models with residuals // Communications in Statistics. B: Simulation and Computation. 2013. Vol. 42, iss. 6. – P. 1352-1367.

[3] Cox D.R., Roy J. Regression models and life tables (with Discussion) // Journal of the Royal Statistical Society, Vol. B, No. 34, 1972. P. 187-220.

[4] Galanova N.S., Lemeshko B.Yu., Chimitova E.V. Using nonparametric goodness-of-fit tests to validate accelerated failure time models // Optoelectronics, Instrumentation and Data Processing. 2012. Vol. 48. Is. 6. – P. 580-592.

[5] Karmanov V.S. Research on mathematical models of resistance of cutting tool // Nauchniy vestnik NSTU, Novosibirsk. 2006. No. 2. – P. 55-64. In Russian.

[6] Konig W. Depiereux W. Methods of optimization of giving and speed of cutting // Ind. Anz. – 1969.

[7] Lemeshko B.Yu. Statistical data analysis, simulation and study of probability regularities. Computer approach – Monograph // NSTU-publisher, Novosibirsk. 2011. 888p. In Russian.

[8] Lemeshko B.Yu., Lemeshko S.B. Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part 1 // Measurement Techniques. 2009. Vol. 52, No. 6. – P. 555-565.

[9] Lemeshko B.Yu., Lemeshko S.B. Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II // Measurement Techniques. 2009. Vol. 52, No. 8. – P. 799-812.

[10] Smagin G.I., Karmanov V.S. Algorithm of norming cut modes of a hard preparing materials by charactical lines and surfaces method used special experimental plans // Nauchniy vestnik NSTU, Novosibirsk. 2011. No. 3. – P. 149-158. In Russian.

[11] Smagin G.I., Karmanov V.S. The method of normalization of optimum cutting conditions for difficult to machine materials // Metalwork. 2004. No. 4. – P. 17-18. In Russian.

[12] Smagin G.I., Karmanov V.S. Characteristics and criteria of tool resistance for roughing and finishing sharpening // Metalwork. 2005. No. 4. – P. 34-35. In Russian.

[13] Smagin G.I., Karmanov V.S., Yakovlev N.D., Fedin I.V. Consideration of power consumption in choosing optimal modes of drilling // Actual problems in machine building. 2015. No. 2. – P. 27-33. In Russian.

Lifetime Data Analysis via Fiabilitis

Evans Gouno and Luc Courtrai
University of South Brittany, Vannes, France
e-mail: evans.gouno@univ-ubs.fr, luc.courtrai@univ-ubs.fr

Abstract

This paper is an overview of the statistical methods that are applied in the software Fiabilitis. Fiabilitis has been developed through students project during several academic years at University of South Brittany, France. The work has been supervised by reseachers from the department of mathematics and from the department of computer sciences. It is a free software available on the internet.

Keywords: Reliability, life testing, censored data model, estimation, Bayesian inference, accelerated life test, design of experiment.

Introduction

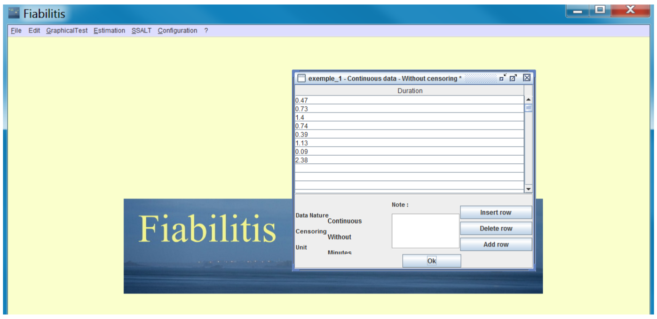
Fiabilitis is a software devoted to statistical analysis of durations samples in the context of reliability. It is developed by researchers from the Department of Mathematics (LMBA) and from the Department of Computer Sciences (IRISA) of the University of South Brittany in Vannes, France. The sofware handles classical models used in the industry but it can be usefull in any fields dealing with durations. Fiabilitis provides graphical goodness of fit tests for parametric models (exponential, Weibull, log-normal). It proposes inferences (estimations, confidence intervals) for the parameters (maximum likelihood, Bayesian methods). It offers the possibility to analyse data from accelerated lifetest (ALT) commonly used in electronics. It also addresses the important matter of designing step stress accelerated lifetest (SSALT). In this paper we describe some of the statistical methods that are implemented in the software.

1 Data and Graphical methods

Fiabilitis deals with differents kind of data. The software treates complete data that can be continuous (exact time to the event of interest) or grouped (belonging to a time interval of the durations). In both case three schemes of censorship are considered: type I, type II and progressive censoring. Data are recorded in a spreadsheet. The figure 1 presents a Fiabilitis file for the following example of complete data (times in hours) ([10] pp. 185):

$$0.47 \quad 0.73 \quad 1.4 \quad 0.74 \quad 0.39 \quad 1.13 \quad 0.09 \quad 2.38$$

The table 1 displays an example of typical grouped data that can be treated with Fiabilitis.



exemple_1 - Continuous data - Without censoring *

```
Duration
0.47
0.73
1.4
0.74
0.39
1.13
0.09
2.38
```

Data Nature: Continuous
Censoring: Without
Unit: Minutes
Note :

Figure 1: Data entry : complete data.

Table 1: Example of grouped data with censoring

Times interval (hours)	[0,168[	[168,500[	[500, 750[	[750,1000]
Number of failure	1	0	2	1
Number of censored data	0	2	4	6

The first step of the analysis is to apply a probability plot to assess the fitting to a theoritical model [12]. In the present time, Fiabilitis suggests test for three classical distributions: exponential, Weibull and log-normale. Let us recall that the graphical test relies on the relationship between transformations of the reliability function and durations. In Fiabilitis, the reliability function is estimated with the Kaplan-Meier estimator [8]. For the probability plot, the following expression is used:

$$R_n = \prod_{x_{(j)}<x} \left(1 - \frac{d_j}{n_j}\right)$$

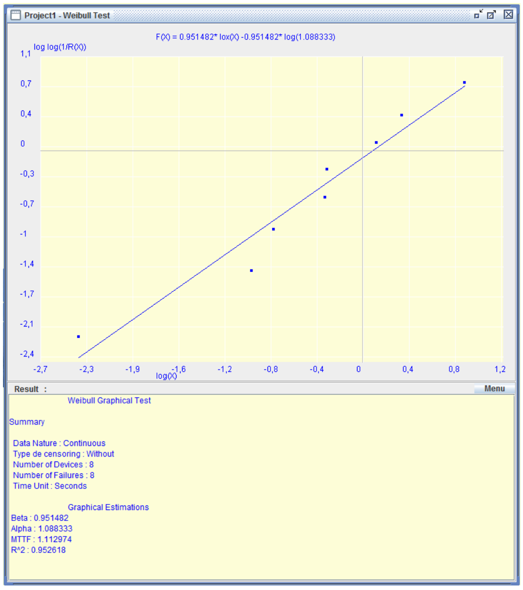
where $x_{(j)}$ is the j-th failure time (not censored), d_j is the number of failures at x_j and n_j is the number of items at risk just before x_j (including censored). For example for the Weibull distribution with parameters (α, β), the relationship is:

$$\log\log 1/\hat{R}(x_{(j)}) = \beta \log x_{(j)} - \beta \log \alpha.$$

The figure 2 displays a Fiabilitis output for a graphical test applied to the previous example of complete data.

2 Estimation

Two approaches for estimating the parameters of the selected model through the probability plot are available: a likelihood approach and a Bayesian approach. Maximum likelihood estimates and confidence intervals for the parameters of exponential,



Weibull Graphical Test

Summary

Data Nature : Continuous
Type de censoring : Without
Number of Devices : 8
Number of Failures : 8
Time Unit : Seconds

Graphical Estimations
Beta : 0.951482
Alpha : 1.088333
MTTF : 1.112974
R^2 : 0.952618

Figure 2: Weibull graphical test applied to the previous example of complete data.

Weibull and log-normale distributions are provided. For the exponential distribution, a closed-form expression is available for the MLE in every case of censorship. Confidence intervals are computed with the Chi-2 and the normal approximation, depending on the nature of the data. Remark that for the Type I censoring confidence intervals could be obtained with the sampling distribution computed by Batholomew [2]. For the Weibull and the log-normale distributions, numerical methods are implemented since no closed-form expressions exist for the MLE. Confidence intervals bounds are obtained using normal approximation. The second approach is Bayesian. Let us consider $f(\underline{x} \mid \theta)$, the distribution of the observation $\underline{x} = (x_1, \ldots, x_n)$ and $\pi(\theta)$, a chosen prior distribution. The posterior distribution is obtained by the Bayes theorem:

$$\pi(\theta) = \frac{f(x \mid \theta)\pi(\theta)}{f(x)} \quad \text{where} \quad f(x) = \int_{\Theta} f(x \mid \theta)\pi(\theta)d\theta.$$

Fiabilitis computes the Bayes estimators of the parameters as the expectation of the posterior distribution (hypothesis of a quadratic loss function). In the next sections details on the computations for the Weibull and for the log-normale distributions are given. The table 2 summarizes the different choices proposed by Fiabilitis.

Weibull

Let us suppose a uniform distribution on β. Let us denote $\eta = \alpha^{\beta}$ and consider a non informative prior of the form $1/\eta^c$, $\eta \in [0, +\infty[$. The posterior distributions are :

$$\pi(\eta \mid x) = \frac{1}{\Gamma(k+c-1)\ \eta^{k+c}} \int_a^b Q(\beta) \exp\left\{-\frac{TTT_{\beta})}{\eta}\right\} d\beta \tag{1}$$

Table 2: Prior and posterior distributions used in Fiabilitis

$f(\underline{x} \mid \theta)$	$\pi(\theta)$	$\pi(\theta \mid \underline{x})$
Exponential	Uniform	Truncated gamma
	Gamma	Gamma
Weibull (α, β)	β known η inverse-gamma	Inverse-gamma
	β uniform η non informative	expression (2) expression (1)
log-normal (μ, σ^2)	$\mu \mid \sigma^2$ normal σ^2 gamma	normal σ^2 inverse-gamma
	non informative $\pi(\mu, \sigma^2) \propto 1/\sigma$	expression (4) expression (5)

and

$$\pi(\beta \mid x) = \frac{Q(\beta)}{\left[TTT_\beta\right]^{k+c-1}} \; \mathbb{1}_{[a,b]}(\beta) \tag{2}$$

where $TTT_\beta = \sum_{i=1}^n x_i^\beta$, $\varphi_{1,0}(x) = \displaystyle\int_a^b \beta^k \prod_{i=1}^n x_i^{(1-\delta_i)(\beta-1)} / \left[TTT_\beta\right]^{k+c-j} d\beta$

and $Q(\beta) = \dfrac{\beta^k \prod_{i=1}^n x_i^{(1-\delta_i)(\beta-1)}}{\varphi_{1,0}(x)}$.

Monte Carlo methods are used to approximate the different integrals involved in the computations of the posterior distributions and their expectations.

Log-normale

A prior distribution of the form $\pi(\mu, \sigma^2) \propto 1/\sigma$ is considered on (μ, σ^2) [13]. The posterior distribution is then:

$$\pi(\mu, \sigma \mid \underline{x}) \propto \frac{1}{\sigma^{n+1}} \; \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\log x_i - \mu)^2\right\}$$

which can be expressed as:

$$\pi(\mu, \sigma \mid \underline{x}) \propto \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{n}{2\sigma^2}\left[(\mu-\overline{\ell})^2 + \overline{\ell^2} - \overline{\ell}^2\right]\right\} \tag{3}$$

with $\overline{\ell} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$ et $\overline{\ell^2} = \frac{1}{n}\sum_{i=1}^{n} (\log x_i)^2$.
The expression (3) can be decomposed into:

$$\pi(\mu \mid \sigma^2, \underline{x}) \propto \exp\left\{-\frac{n}{2\sigma^2}(\mu-\overline{\ell})^2\right\} \tag{4}$$

$$\pi(\sigma^2 \mid \mu, \underline{x}) \propto \frac{1}{(\sigma^2)^{(n-1)/2+1}} \exp\left\{-\frac{n[(\mu-\overline{\ell})^2 + \overline{\ell^2} - \overline{\ell}^2]/2}{\sigma^2}\right\} \tag{5}$$

(4) is a normale distribution with parameters $(\bar{\ell},\ \sigma^2/n)$ and (5) corresponds to an inverse-gamma distribution with parameters (a, b), $a = (n-1)/2$ et $b = n[(\mu-\bar{\ell})^2 + \bar{\ell^2} - \bar{\ell}^2]/2$.
A Gibbs algorithm is applied to obtain realisations of (μ, σ):

```
Initialisation :  $\sigma^{2(0)}$ (the graphical estimate can be used)
At step $(q)$ :

  1. Draw $\mu^{(q+1)} \sim \mathcal{N}(\bar{\ell}, \sigma^{2(q)})$

  2. Draw $z$ according a gamma distribution with parameters $(a, b)$
     where $a = (n-1)/2$ et $b = n[(\mu-\bar{\ell})^2 + \bar{\ell^2} - \bar{\ell}^2]/2$

  3. Compute $\sigma^{2(q+1)} = 1/z$.

  4. return to 1.
```

It is then possible to compute approximations of the expectations and therefore bayesian estimates of μ and σ.

One of the difficuties encountered by the practitioner in applying the bayesian techniques is the choice of the values for the parameters of the prior distribution. Fiabilitis deduces automatically these values from information given by the user. These information are answers to easy questions. For example, it will be asked to the user to propose an order of magnitude for the MTTF and how much he trusts his guess. With these numbers, the values for the prior parameters are computed and the Bayesian estimates are provided.

3 Step-stress Accelerated lifetest

Accelerated life test (ALT) is an experimental strategy to obtain information on the life time of highly reliable products [11]. The material is submitted to higher-than-usual environnental conditions (stress) inducing failures in a shorter time. The stress can be modified several times during the test. This is step-stress accelerated life test

(SSALT) [3]. Many questions arise setting such tests. How many pieces should be put on test? How many step? How long should they last? What should be the level of stress in each step? Another point is the estimation of the parameters characterising the lifetime.

Fiabilitis handles two models of ALT. The first one is the Arrhenius model where the stress is the temperature [6]. The second one is the Peck model where the stress is a combination of relative humidity and temperature [9]. These two ALT models are very commun in electronics. The table 3 displays the ALT models available in Fiabilitis.

Table 3: ALT

ALT	Stress	Model	Parameter θ
Arrhenius	Temperature T	$\exp\left\{-\frac{E_a}{kT}\right\}$	E_a, λ_0
Peck	Relative humidity RH and temperature T	$RH^\eta \; \exp\left\{-\frac{E_a}{kT}\right\}$	η, E_a, λ_0

Fiabilitis can be used to design SSALT and to make inference. Optimal ALT are obtained by minimizing the generalized variance that is to say the determinant of the inverse the information matrix [1], [4], [5]. Two criteria are retained for the designing part.

- For optimal step-stress length : the practitioner gives the number of steps, the level of stress in each steps, a guess on the MTTF and on the activation energy. The Fiabilitis returns the length of the step-stress.
- For optimal level of stress : the practioner gives the number of steps, the length of the steps, a guess on the MTTF and a value for the activation energy. Then Fiabilitis returns the level of stress in each step.

The durations in SSALT are modelised with a piecewise exponential model. The MLE of the parameters are then computed.

Conclusions

The purpose of Fiabilitis is to provide a convenient tool for analysing durations. It is addressed to engineers or any practitioners concerned by lifetime studies. One of the strengths of the software is that it offers a very easy way to deal with the Bayesian approach. The prior parameters are brought out through simple indicators provided by the user. The global structure of the software has been designed in order to allow an easy implementation of new models (distributions, ALT, etc.). This is going to be done in the future. We are also considering the adjunction of the statistical treatments of data from stochastic processes. This new tools will find applications in operational safety of big systems for example.

References

[1] Bai D.S., Kim M.S., Lee S.H. (1989). Optimum simple step-stress accelerated life tests with censoring. *IEEE Transactions on Reliability*. Vol. bf 38, pp. 528–532.

[2] Bartholomew D. (1963). The sampling distribution of an estimate arising in life testing. *Technometrics*. Vol. **5**, pp. 361–374.

[3] Gouno E., Balakrishnan N. (2001). Step-stress accelerated life test. *Handbook of Statistic*. Vol. **20**, pp. 623–638.

[4] Gouno E., Sen A., Balakrishnan N. (2004). Optimal step-stress test under progressive type I censoring. *IEEE Trans. on Reliab.*. Vol.**53**, pp. 388–395.

[5] Gouno E. (2007). Optimum step-stress for temperature accelerated life test. *Qual. Reliab. Engng. Int.*. Vol. **23**, pp. 915–924.

[6] Jensen F. (1985). Activation energies and the Arrhenius equation. *Quality and Reliability Engineering International*. Vol. **1**, pp. 13–17.

[7] Khamis I.H. (1997). Optimum m-step, step-stress test with k stress variables. *Communications in Statistics – Computation and Simulation*. Vol. **26**, pp. 1301–1313.

[8] Kaplan E., Meier P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*. Vol. **53**, pp. 457–481.

[9] Lall P., Pecht (1996). Tutorial: Temperature as an input to microelectronics-reliability models. *IEEE Transactions on Reliability*. Vol. **45**, pp. 3–9.

[10] Lawless J.F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York.

[11] Nelson W. (1980). Accelerated life testing – Step-stress model and data analysis. *IEEE Transactions on Reliability*. Vol. **29**, pp. 103–108.

[12] Nelson W. (2000). Theory and applications of hazard plotting for censored failure data. *Technometrics*. Vol. **42**, pp. 12–25.

[13] Upadhyay S., Peshwani M. (2001). Full posterior analysis of three parameter lognormal distribution using gibbs sampler. *J. Statist. Comput. Simul.*. Vol. **71**, pp. 215–230.

Statistical Inference in Cox Models

Henning Läuter, Hannelore Liero
University of Potsdam, Potsdam, Germany
e-mail: `laeuter@math.uni-potsdam.de, liero@math.uni-potsdam.de`

Abstract

Our procedure of estimating is the maximum partial likelihood estimate which is the appropriate estimate in the Cox model with a general censoring distribution C, covariates $\mathbf{X}$ and an unknown baseline hazard rate $\lambda_0(t)$. We find conditions for estimability and asymptotic estimability. The asymptotic variance matrix of the MPLE is represented and properties discussed.

Keywords: Cox model, estimability, asymptotic variance, optimal design and covariates.

Introduction

The Cox proportional hazards model is the most popular model for analyzing survival data. Let $\lambda(t|\mathbf{x})$ be the conditional hazard function of T given the covariate vector $\mathbf{X}$ has the value $\mathbf{x}$, defined as

$$\lambda(t|\mathbf{x}) = \lim_{\triangle t \downarrow 0} \frac{1}{\triangle t} \mathsf{P}[t \leq T < t + \triangle t | T \geq t, \mathbf{X} = \mathbf{x}].$$

The Cox model assumes the following form (Cox 1972):

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp(\boldsymbol{\beta}_0^T \mathbf{x}), \tag{1}$$

where $\exp(\boldsymbol{\beta}_0^T \mathbf{x}) = \exp(\beta_{01} x_1 + \ldots + \beta_{0p} x_p)$ is the hazard ratio, λ_0 is an unknown baseline hazard function (the hazard function for an individual with $\mathbf{x} = \mathbf{0}$) and $\boldsymbol{\beta}_0 \in \mathbf{R}^p$ is an unknown parameter to be estimated.
The densities of the data are determined by the hazard rate completely. One can describe the model by the hazard rates or in the equivalent way by densities. Sometimes the heuristic interpretation is easier with the hazard rates.
We start with observations, which were realizations of i.i.d. copies $(T_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ of $(T, \delta, \mathbf{X})$. Let $T_i = \min\{T_i^*, C_i\}$ where T_i^* is the individual i'th survival time and $\delta_i = \mathbf{1}(T_i^* \leq C_i)$ is the censoring indicator function ($\delta_i = 1$ if event has occurred, 0 if the lifetime is censored), where C_i is the individual i'th censoring time. The $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ is the individual i'th random covariate.
The parameter $\boldsymbol{\beta}_0$ will be estimated by the maximum partial likelihood estimate which is the appropriate estimate in a Cox model with a general censoring distribution C, covariates $\mathbf{X}$ and an unknown baseline hazard rate λ_0.

Partial likelihood method

Let us derive the partial likelihood function. We denote the observed ordered lifetimes by $t_{(j)}, j = 1, \ldots, d$ where d is the number of observed (uncensored) lifetimes. We start with the presentation of the partial likelihood method as it was introduced by D.R. Cox and we assume that all lifetimes are distinct, in other words there are no ties and we have $t_{(1)} < t_{(2)} < \ldots < t_{(d)}$.

Remark: When ties between event times are found in the data, alternate partial likelihoods have been provided by a variety of authors; see Breslow (1974), Efron (1977) and Cox (1972).

Define the **risk set** $R(t)$ at time t as the set of subjects alive and under observation at time t^-, immediately prior to t:

$$R(t) = \{i : T_i \geq t\}.$$

For the definition of the estimator we need only the risk set at the lifetime $R(t_{(j)})$, however it is defined for all t.

The partial likelihood, based on the hazard function (1) as defined by Cox, is expressed by

$$L_n(\boldsymbol{\beta}) = \prod_{j=1}^{d} \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{(j)})}{\sum\limits_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \tag{2}$$

where $\mathbf{x}_{(j)}$ denotes the covariates associated with the individual whose lifetime is $t_{(j)}$. Cox suggested treating the partial likelihood as a regular likelihood function and making inference on $\boldsymbol{\beta}_0$ accordingly. Then we get the estimate of $\boldsymbol{\beta}_0$, often called MPLE (maximum partial likelihood estimate) by maximizing the partial likelihood and use the minus of the second derivative of the log partial likelihood as the information matrix.

Let $l_n(\boldsymbol{\beta}) = \log L_n(\boldsymbol{\beta})$. Then, we can write $l_n(\boldsymbol{\beta})$ as

$$l_n(\boldsymbol{\beta}) = \sum_{j=1}^{d} \left[\boldsymbol{\beta}^T \mathbf{x}_{(j)} - \log \left\{ \sum_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \mathbf{x}_i) \right\} \right]. \tag{3}$$

The score function $\mathbf{U}_n(\boldsymbol{\beta})$ with components $U_{nk}(\boldsymbol{\beta}) = \dfrac{\partial l_n(\boldsymbol{\beta})}{\partial \beta_k}$, $k = 1, ..., p$ is:

$$U_{nk}(\boldsymbol{\beta}) = \sum_{j=1}^{d} \left[\mathbf{x}_{(j)k} - \frac{\sum\limits_{i \in R(t_{(j)})} \mathbf{x}_{ik} \exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{\sum\limits_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \right],$$

i.e.

$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{j=1}^{d} \left[\mathbf{x}_{(j)} - \frac{\sum\limits_{i \in R(t_{(j)})} \mathbf{x}_i \exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{\sum\limits_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \right]. \tag{4}$$

The MPLE, $\hat{\boldsymbol{\beta}}_n$, can be obtained by solving the given system of equations $\mathbf{U}_n(\boldsymbol{\beta}) = \mathbf{0}$.

The observed information matrix $\mathbf{I}_n(\boldsymbol{\beta}) = \Big(I_{ngk}(\boldsymbol{\beta})\Big)_{p\times p}$ is the negative of the matrix of second derivatives of the log likelihood function and has the elements $I_{ngk}(\boldsymbol{\beta}) = -\dfrac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\beta_k\partial\beta_g}$:

$$\mathbf{I}_n(\boldsymbol{\beta}) = \sum_{j=1}^{d}\left[\frac{\sum\limits_{i\in R(t_{(j)})}\mathbf{x}_i^{\otimes 2}\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{\sum\limits_{i\in R(t_{(j)})}\exp(\boldsymbol{\beta}^T\mathbf{x}_i)} - \left\{\frac{\sum\limits_{i\in R(t_{(j)})}\mathbf{x}_i\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{\sum\limits_{i\in R(t_{(j)})}\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}\right\}^{\otimes 2}\right] \tag{5}$$

with $\mathbf{x}^{\otimes 2} := \mathbf{x}\mathbf{x}^T$.

Under regularity conditions we know the asymptotic behavior of the MPLE $\hat{\boldsymbol{\beta}}_n$. We have

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathrm{D}} \mathsf{N}(0, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_0, \lambda_0))$$

for the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}_n$ and

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \lambda_0) = \mathrm{plim}_{n\to\infty} n^{-1}\mathbf{I}_n(\boldsymbol{\beta}_0)$$

and

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \lambda_0) = \mathrm{plim}_{n\to\infty} n^{-1}\mathbf{I}_n(\hat{\boldsymbol{\beta}}_n).$$

If one likes to estimate $\boldsymbol{\beta}_0$ then the basic property of the estimate is to have a well defined asymptotic variance matrix $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_0, \lambda_0)$.

The observed information matrix $\mathbf{I}_n(\boldsymbol{\beta})$ depends on the $t_1, \ldots, t_n$, $\delta_1, \ldots, \delta_n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This we express in the notation of the form

$$\mathbf{I}_n(\boldsymbol{\beta}) = \mathbf{I}_n(\boldsymbol{\beta}; t_1, \ldots, t_n; \delta_1, \ldots, \delta_n; \mathbf{x}_1, \ldots, \mathbf{x}_n), \tag{6}$$

but this we will write only in the cases when we need this dependence explicitly. In general we use the shorter notation $\mathbf{I}_n(\boldsymbol{\beta})$.

The $\mathbf{X}_i$, $i = 1, 2, ...$ are covariates which characterize conditions of the considered process described by the model. We assume here that these conditions can be controlled as it is given in many technological or medical problems. If we will consider e.g. the failure times of car tires for different countries then we use the countries as covariates and for estimating $\boldsymbol{\beta}_0$ we can choose the countries where we have to measure for estimating $\boldsymbol{\beta}_0$ in an optimal way. This means that the model (1) holds, the parameter $\boldsymbol{\beta}_0$ is to be estimated and the covariates can be chosen. This can be considered as a problem of experimental design, e.g. one chooses the places where one observes the survival times. As usually in experimental design and estimation problems we formulate the assumptions for estimability and derive the criteria for optimal choices of covariates. These problems are discussed in Wichitsa-nguan, Läuter, Liero (2015). An optimal design problem for censored observations was considered by Balakrishnan and Han (2007).

Representation of the observed information matrix

The basis for estimability is determined by the bias and the variance of the estimation. We consider here the estimability of $\boldsymbol{\beta}_0$ with the MPLE $\hat{\boldsymbol{\beta}}_n$ and therefore we look at the properties of the observed information matrix and their limit. In the next Lemma the representation of the observed information matrix gives the similar representation as one knows for the usual variance matrices.

Lemma 1: *Let the points* $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m$ *be elements in the support of* $\mathbf{X}$. *Then the observed information matrix with respect to the MPLE can be written as*

$$\mathbf{I}_n(\boldsymbol{\beta}) = \frac{1}{2} \sum_{r=1}^{m} \sum_{s=1}^{m} \kappa_{nrs}(\boldsymbol{\beta}) \mathbf{w}_{rs} \mathbf{w}_{rs}^T \tag{7}$$

with

$$\mathbf{w}_{rs} = \boldsymbol{\xi}_r - \boldsymbol{\xi}_s, \tag{8}$$

$$R_l(t_j) = \sum_{i:\mathbf{x}_i = \boldsymbol{\xi}_l} Y_i(t_j), \tag{9}$$

$$\kappa_{nrs}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \frac{R_r(t_i) R_s(t_i) \exp(\boldsymbol{\beta}^T(\boldsymbol{\xi}_r + \boldsymbol{\xi}_s))}{(\sum_{l=1}^{m} R_l(t_i) \exp(\boldsymbol{\beta}^T \boldsymbol{\xi}_l))^2}. \tag{10}$$

The proof follows by direct calculations. One recognizes in this representation, that $\mathbf{I}_n(\boldsymbol{\beta})$ is a positive semidefinite matrix and we are able to formulate conditions about the rank of $\mathbf{I}_n(\boldsymbol{\beta})$.
The coefficients $\kappa_{nrs}(\boldsymbol{\beta})$ can be expressed with relative frequencies instead of $R_l(t_i)$. We denote

$$f_{il} = \frac{R_l(t_i)}{\sum_{j=1}^{m} R_j(t_i)}. \tag{11}$$

Then calculations lead direct to

$$\kappa_{nrs}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \frac{f_{ir} f_{is} \exp(\boldsymbol{\beta}^T(\boldsymbol{\xi}_r + \boldsymbol{\xi}_s))}{(\sum_{l=1}^{m} f_{il} \exp(\boldsymbol{\beta}^T \boldsymbol{\xi}_l))^2}. \tag{12}$$

Estimability of $\boldsymbol{\beta}_0$

We are interested in estimating the unknown parameter $\boldsymbol{\beta}_0$ under the model (1) with observations $t_1, \ldots, t_n$, $\delta_1, \ldots, \delta_n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We will find conditions that the MPLE is a unique solution of the partial likelihood equation. This is a fact of identifiability or estimability.

Definition 1 *Let* $t_1, \ldots, t_n$, $\delta_1, \ldots, \delta_n$ *and* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *be the observed values and* $\mathbf{I}_n(\boldsymbol{\beta}; t_1, \ldots, t_n; \delta_1, \ldots, \delta_n; \mathbf{x}_1, \ldots, \mathbf{x}_n)$ *be the observed information matrix. Then* $\boldsymbol{\beta}_0$ *is* *estimable by the maximum partial likelihood estimate if* $\mathbf{I}_n(\boldsymbol{\beta})$ *is nonsingular for all* $\boldsymbol{\beta} \in \mathbb{R}^p$.

Remark 1. *In this case,* $\mathbf{I}_n(\boldsymbol{\beta})$ *is a positive definite matrix. This implies that the log partial likelihood is a concave function of* $\boldsymbol{\beta}$ *and hence there exists a unique maximum, which can be obtained by setting the first derivative of the log partial likelihood, i.e., score function* $\mathbf{U}(\boldsymbol{\beta})$*, to be zero.*

Definition 2 *Let* $t_1, \dots, t_n$*,* $\delta_1, \dots, \delta_n$ *and* $\mathbf{x}_1, \dots, \mathbf{x}_n$ *be the observed values and* $\mathbf{I}_n(\boldsymbol{\beta}; t_1, \dots, t_n; \delta_1, \dots, \delta_n; \mathbf{x}_1, \dots, \mathbf{x}_n)$ *the observed information matrix. Then* $\boldsymbol{\beta}_0$ *is asymptotically estimable by the maximum partial likelihood estimate if* $\boldsymbol{\Sigma}(\boldsymbol{\beta}, \lambda_0)$ *is nonsingular for all* $\boldsymbol{\beta}$ *and* λ_0*.*

Denote $\mathcal{L}(S)$ be the linear space spanned by the elements of the set S.

Theorem 1 *Let* $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_m \in \mathbb{R}^p$ *be given in the support of* $\mathbf{X}$*. Let* $\delta_i = 1$*,* $f_{i1}, \dots, f_{im} > 0$ *for some* $i \in \{1, \dots, n\}$ *,* $\mathbf{w}_{st} = \boldsymbol{\xi}_s - \boldsymbol{\xi}_t$ *and* $\tilde{m} = \dim \mathcal{L}\{\mathbf{w}_{st} | 1 \le s < t \le m\}$ *Then*

$$\operatorname{rank}(\mathbf{I}_n(\boldsymbol{\beta})) = \min(p, \tilde{m}).$$

The assumptions in this Theorem 1 mean that we have at least in one time point uncensored observations for all values $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_m$. Under those assumptions, $\boldsymbol{\beta}_0$ is estimable for $\tilde{m} \ge p$. The next Theorem 3 gives a slightly more general result.

Theorem 2 *Let* $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_m \in \mathbb{R}^p$ *be given in the support of* $\mathbf{X}$*. If for any* r, s *with* $1 \le r < s \le m$ *there exists some* i *with* $\delta_i f_{ir} f_{is} > 0$*, then*

$$rank(\mathbf{I}_n(\boldsymbol{\beta})) = \min(p, \tilde{m})$$

with $\mathbf{w}_{rs}$ *and* $\tilde{m}$ *as in Theorem 1.*

A consequence of the Theorem 2 is that $m \ge p + 1$ is necessary for estimability. For $m \le p$, the $\boldsymbol{\beta}_0$ is not estimable.

Asymptotic estimability

We will use a notation which expresses dependence on all unknown parameters.
We have $\overline{H}(t) = 1 - H(t)$ which depends on $\boldsymbol{\xi_l}$, $\boldsymbol{\beta}_0$ and λ_0, and we write

$$\begin{aligned} \overline{H}(t^- | \boldsymbol{\xi}_l; \boldsymbol{\beta}_0, \lambda_0) &= 1 - H(t^- | \boldsymbol{\xi}_l; \boldsymbol{\beta}_0, \lambda_0) \\ &= (1 - F(t^- | \boldsymbol{\xi}_l; \boldsymbol{\beta}_0, \lambda_0))(1 - G(t^-)). \end{aligned} \tag{13}$$

Sometimes we will use the shorter notation, but in the following theorem we take the full description for clearness.

Theorem 3 *Let the support of* $\mathbf{X}$ *be finite with* $\mathsf{P}(\mathbf{X} = \boldsymbol{\xi}_j) = q_j$ *for* $j = 1, \ldots, m$*;* $q_j > 0$*,* $\sum_{j=1}^m q_j = 1$*. Then*

$$\Sigma(\boldsymbol{\beta}_0, \lambda_0) = \frac{1}{2} \sum_{r=1}^{m} \sum_{s=1}^{m} \nu_{rs}(\boldsymbol{\beta}_0, \lambda_0, \mathbf{q}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m) \mathbf{w}_{rs} \mathbf{w}_{rs}^T \tag{14}$$

with

$$\nu_{rs}(\boldsymbol{\beta}_0, \lambda_0, \mathbf{q}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m) = \int_0^\tau \frac{\overline{H}(\gamma^-|\boldsymbol{\xi}_r; \boldsymbol{\beta}_0, \lambda_0)\overline{H}(\gamma^-|\boldsymbol{\xi}_s; \boldsymbol{\beta}_0, \lambda_0) q_r q_s \exp(\boldsymbol{\beta}_0^T(\boldsymbol{\xi}_r + \boldsymbol{\xi}_s))}{\sum_{j=1}^m \overline{H}(\gamma^-|\boldsymbol{\xi}_j; \boldsymbol{\beta}_0, \lambda_0) q_j \exp(\boldsymbol{\beta}_0^T \boldsymbol{\xi}_j)} \lambda_0(\gamma) d\gamma. \tag{15}$$

The importance of this Theorem consists in the consequences that under general conditions the asymptotic estimability can be proven. We use $\overline{H} = 1 - H$ and assume

$$\overline{H}(t^-|\boldsymbol{\xi}_j; \boldsymbol{\beta}_0, \lambda_0) \neq 0 \quad \text{for} \quad t \in [0, \tau], \tag{16}$$

$$\int_0^\tau (1 - G(\gamma^-)) \lambda_0(\gamma) d\gamma > 0. \tag{17}$$

Theorem 4: *We assume (16) and (17).* $\boldsymbol{\beta}_0$ *is asymptotically estimable if and only if*

$$\dim \mathcal{L}(\{\boldsymbol{\xi}_r - \boldsymbol{\xi}_s, 1 \leq s < r \leq m\}) = p. \tag{18}$$

Asymptotic variance for general covariates

In Theorem 3 a representation of the asymptotic variance is given under the assumption that the support of $\mathbf{X}$ is finite. The support points are $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m$ and we have

$$(\mathbf{X} = \boldsymbol{\xi}_j) = q_j \quad \text{for} \quad j = 1, \ldots, m$$

with $q_j > 0$, $\sum_{j=1}^m q_j = 1$. We denote by Q the measure in $\mathbb{R}^p$, where

$$Q(\boldsymbol{\zeta}) = \begin{cases} q_j, & \text{if } \boldsymbol{\zeta} = \boldsymbol{\xi}_j \\ 0, & \text{otherwise .} \end{cases}$$

Then we have the representations

$$\int_0^\tau \overline{H}(t^-|\boldsymbol{\zeta}; \boldsymbol{\beta}_0, \lambda_0) \exp(\boldsymbol{\beta}_0^T \boldsymbol{\zeta}) dQ(\boldsymbol{\zeta}) = \sum_{j=1}^m \overline{H}(t^-|\boldsymbol{\xi}_j; \boldsymbol{\beta}_0, \lambda_0) q_j \exp(\boldsymbol{\beta}_0^T \boldsymbol{\xi}_j),$$

$$\mathbf{\Sigma}(\boldsymbol{\beta}_0, \lambda_0) = \int\int h(\boldsymbol{\zeta}, \boldsymbol{\rho}, \boldsymbol{\beta}_0, \lambda_0)(\boldsymbol{\zeta} - \boldsymbol{\rho})(\boldsymbol{\zeta} - \boldsymbol{\rho})^T dQ(\boldsymbol{\zeta}) dQ(\boldsymbol{\rho}) \tag{19}$$

with

$$h(\boldsymbol{\zeta}, \boldsymbol{\rho}, \boldsymbol{\beta}_0, \lambda_0) = \int_0^\tau \frac{\overline{H}(\gamma^- | \boldsymbol{\zeta}; \boldsymbol{\beta}_0, \lambda_0) \exp(\boldsymbol{\beta}_0^T \boldsymbol{\zeta}) \overline{H}(\gamma^- | \boldsymbol{\rho}; \boldsymbol{\beta}_0, \lambda_0) \exp(\boldsymbol{\beta}_0^T \boldsymbol{\rho})}{\int \overline{H}(\gamma^- | \boldsymbol{\eta}; \boldsymbol{\beta}_0, \lambda_0) \exp(\boldsymbol{\beta}_0^T \boldsymbol{\eta}) dQ(\boldsymbol{\eta})} \lambda_0(\gamma) d\gamma \tag{20}$$

and

$$\begin{aligned}\int\int h(\boldsymbol{\zeta}, \boldsymbol{\rho}, \boldsymbol{\beta}_0, \lambda_0) dQ(\boldsymbol{\zeta}) dQ(\boldsymbol{\rho}) &= \sum_{r=1}^m \sum_{s=1}^m q_r q_s h(\boldsymbol{\xi}_r, \boldsymbol{\xi}_s, \boldsymbol{\beta}_0, \lambda_0) \\ &= \sum_{r=1}^m \sum_{s=1}^m \nu_{rs}(\boldsymbol{\beta}_0, \lambda_0, \mathbf{q}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m).\end{aligned}$$

In these representations the measure Q is a probability measure in $(\mathbb{R}^p, \mathfrak{B}^p)$, where $\mathfrak{B}^p$ is the Borel-σ-algebra in $\mathbb{R}^p$. We see in (19) that $\mathbf{\Sigma}(\boldsymbol{\beta}_0, \lambda_0)$ depends on the distributions F and G. The dependence on the covariates is described in the probability measure Q. For any Q this representation holds. Therefore one generalizes this representation of $\mathbf{\Sigma}(\boldsymbol{\beta}_0, \lambda_0)$ for any, possibly continuous, probability measures Q.

Definition 5 *Let be $(\mathbb{R}^p, \mathfrak{B}^p)$ a measurable space with the σ-algebra $\mathfrak{B}^p$ of $\mathbb{R}^p$. For a probability measure Q over $(\mathbb{R}^p, \mathfrak{B}^p)$ we call $\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}_0, \lambda_0, Q)$ with (20) and*

$$\mathbf{\Sigma}(\boldsymbol{\beta}_0, \lambda_0, Q) = \int\int h(\boldsymbol{\zeta}, \boldsymbol{\rho}, \boldsymbol{\beta}_0, \lambda_0)(\boldsymbol{\zeta} - \boldsymbol{\rho})(\boldsymbol{\zeta} - \boldsymbol{\rho})^T dQ(\boldsymbol{\zeta}) dQ(\boldsymbol{\rho})$$

the asymptotic variance matrix of the MPLE of $\boldsymbol{\beta}_0$ in the model (1) where the covariates $\mathbf{X}$ have the distribution Q. This matrix will be denoted by $\mathbf{\Sigma}(\boldsymbol{\beta}_0, \lambda_0; Q)$ if we will express the dependence on the distribution of the covariates.

With this asymptotic variance matrix $\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}_0, \lambda_0, Q)$ we are able to characterize the influence of a covariate $\mathbf{X}$ with the induced measure Q. Moreover we can compare two measure Q_1 and Q_2 by comparing $\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}_0, \lambda_0, Q_1)$ with $\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}_0, \lambda_0, Q_2)$. This is the basis for finding optimal covariates.

Conclusions

In this paper the asymptotical estimability in Cox models is discussed. The essential result consists in the explicit representation of the asymptotic variance matrix of the maximum partial likelihood estimate. One recognizes the influence of censoring and covariates. These representations are the basis for statistical inference problems like testing and different estimation problems. The problem of an optimal choice of covariates can be handled now. Some results are contained in Wichitsa-nguan, Läuter, Liero (2015) and will be extended.

References

[1] Balakrishnan, N. and Han, Donghoon: Optimal progressive type-II censoring schemes for nonparametric confidence intervals of quantiles. Comm. Statist. Simulation Comput. 2007

[2] Breslow, N.: Covariance Analysis of Censored Survival Data, Biometrics 1974.

[3] Cox, D.R.: Regression models and life-tables. J. Roy. Statist. Soc. Ser. B, 1972.

[4] Efron, B.: The Efficiency of Cox's Likelihood Function for Censored Data. Jour. Amer. Statist. Assoc. 1977.

[5] Wichitsa-nguan, K., Läuter, H., Liero, H.: Inference in Cox Models. Preprint Univ.Potsdam 2015

A Score-Test for the Time-dependent Coefficient in the Cox Model

Hannelore Liero and Korakot Wichitsa-nguan
Institute of Mathematics
University of Potsdam, Germany
e-mail: liero@uni-potsdam.de, korakot.w@psu.ac.th

Abstract

An extension of the Cox model for describing time-varying coefficients is considered. For the estimation of these parameter functions the method of local constant maximum partial likelihood is applied. For testing the possible parametric form of the time-dependent coefficient a score test is proposed. The results are based on the asymptotic multivariate normality of the score function at different points.

Keywords: Time-dependent coefficients; Cox model; local constant estimation

Introduction

In survival analysis the Cox proportional hazards model plays an important role in exploring the relationship between a survival time T^* and a covariate X. This model assumes that the regression coefficients are constant over time, i.e., the hazard rate

$$\lambda(t|x) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathsf{P}(t \le T^* < t + \Delta t | T \ge t, X = x)$$

is defined by

$$\lambda(t|x) = \lambda_0(t) \exp(\beta_0^T x) \tag{1}$$

where λ_0 is the so-called baseline hazard function, β_0 is the p-dimensional vector of regression coefficients and $x \in \mathbb{R}^p$. However, sometimes this assumption fails and the coefficients vary over the time. In the present paper we consider an extension of the Cox model with time-varying coefficients, i.e.,

$$\lambda(t;x) = \lambda_0(t) \exp(\beta_0(t)^T x),$$

where the components of $\beta_0(t) = (\beta_{01}(t), \dots, \beta_{0p}(t))^T$ satisfy certain smoothness conditions. For the estimation of the coefficient function we apply the local partial likelihood methods introduced by Cai and Sun (2003).

The aim of this paper is to present a test procedure for testing the null hypothesis that the function $\beta_0(\cdot)$ has a prespecified parametric form. For simplicity of presentation we will give the results for the case $p = 1$. Thus we have the test problem

$$\mathcal{H}: \beta_0(\cdot) \in B_{\text{par}} = \{\beta(\cdot, \vartheta),\ \vartheta \in \Theta \subseteq \mathbb{R}^k\} \qquad \mathcal{K}: \beta(\cdot) \notin B_{\text{par}}.$$

The most important special case of this null hypothesis is that $\beta_0(\cdot)$ is constant, that is, that the classical Cox proportional hazards model is true.

The test procedure will be based on the local partial score function. Considering this score function at a finite number of different points we will show that the corresponding quadratic form converges in distribution to the χ^2-distribution.

An extension of the model is to include time-dependent covariates, i.e. $X = X(t)$. In this paper we will concentrate on the statistical inference concerning the time-varying coefficient, thus for simplicity of presentation we consider a time-invariant covariate.

In many applications, the survival times or failure times T^* are not fully observed; instead they are censored. Thus we observe

$$(T_i, \Delta_i, X_i) \qquad T_i = \min(T_i^*, C_i), \qquad \Delta_i = \mathbb{I}(T_i^* \leq C_i) \qquad i = 1, \dots n$$

where $\mathbb{I}(\cdot)$ is the indicator function. The random variables (T_i, X_i, C_i) are n independent copies of (T, X, C) where C is a censoring variable. We assume that conditional on X the survival time T^* and the censoring time C are independent.

It is convenient to introduce the following notations: With the indicator function $\mathbb{I}(\cdot)$ we define the counting processes and the risk processes

$$N_i(t) = \mathbb{I}(T_i \leq t, \Delta_i = 1) \quad \text{and} \quad Y_i(t) = \mathbb{I}(T_i \geq t) \qquad i = 1, \dots, n.$$

The processes N_i and Y_i are observed in some time interval $[0, \tau]$, $\tau < \infty$, such that $\mathsf{P}(T > \tau) > 0$. The history up to time t is given by

$$\mathcal{F}_t = \sigma\{X_i, N_i(u), Y_i(u),\, 0 \leq u \leq t,\, 1 \leq i \leq n\}.$$

The intensity function of the counting process N_i is given by

$$\alpha_i(t) = Y_i(t)\lambda(t; X_i) = Y_i(t)\lambda_0(t)\exp(\beta_0(t)X_i).$$

With the cumulative intensity function $A_i(t) = \int_0^t \alpha_i(s)\mathrm{d}s$ we obtain by the martingale decomposition that the processes M_i defined by

$$\begin{aligned} M_i(t) &= N_i(t) - A_i(t) \\ &= N_i(t) - \int_0^t Y_i(s)\exp(\beta_0(s)X_i)\lambda_0(s)\mathrm{d}s \end{aligned}$$

are local martingales on the time interval $[0, \tau]$.

1 Estimation of the function $\beta_0(\cdot)$

In this section we describe the method for the estimation of the function β_0. The log partial likelihood function in the classical proportional hazards model (1) is given by

$$l(\beta) = \sum_{i=1}^n \int_0^\tau [\beta X_i - \log(\sum_{j=1}^n Y_j(s)\exp(\beta X_j))]\mathrm{d}N_i(s).$$

Now, consider the extended model with a time-dependent $\beta(s)$. For s in a neighborhood of t we have by Taylor expansion

$$\beta(s) \approx \beta_0(t) + \beta'(t)(s-t) = \beta_1 + \beta_2(s-t),$$

where β' is the first derivative of β. Let $h = h_n$ be a bandwidth that controls the size of the local neighborhood and let K be a kernel function, then the local constant log partial likelihood function is given by

$$\ell_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} K_h(s-t)[\beta X_i - \log \sum_{j=1}^{n} Y_j(s) \exp(\beta X_j)] \mathrm{d}N_i(s) \qquad (2)$$

where $\beta = \beta_1$ and $K_h(s-t) = \frac{1}{h} K\left(\frac{s-t}{h}\right)$ and K is a kernel function satisfying some regularity conditions specified later.

The nonparametric local constant estimator of $\beta(\cdot)$ at the grid point t is the maximizer of the function (2). To maximize ℓ_n we consider the corresponding score function in more detail. For this purpose define the sums

$$S_{nk}(\beta, t) \;=\; \frac{1}{n} \sum_{j=1}^{n} Y_j(t) \exp(\beta X_j) X_j^k \qquad k = 0, 1, 2$$

and

$$E_n(\beta, t) = \frac{S_{n1}(\beta, t)}{S_{n0}(\beta, t)}.$$

With this definitions we rewrite the local constant log partial likelihood function

$$\ell_n(\beta) \;=\; \sum_{i=1}^{n} \int_0^{\tau} K_h(s-t)[\beta X_i - \log(n S_{n0}(\beta, s))] \mathrm{d}N_i(s),$$

and the estimate $\widehat{\beta}(t)$ of β_0 at the grid point t is the solution of the score equation $U_n(\beta, t) = 0$, where

$$U_n(\beta, t) = n^{-1/2} h^{1/2} \sum_{i=1}^{n} \int_0^{\tau} K_h(s-t)[X_i - E_n(\beta, s)] \mathrm{d}N_i(s).$$

Note, we have already multiplied the score function by the normalizing factor $\sqrt{h/n}$.

2 Limit distributions

In several papers the pointwise consistency and asymptotic normality of the resulting estimator $\widehat{\beta}(t)$ at a fixed point t were shown. These results are based on the asymptotic normality of the score function $U_n(\beta_0, t)$ at t. Let us consider distinct points $t_1, \ldots, t_d$ and define the vector

$$\boldsymbol{\mathcal{U}_n}(\beta, \underline{t}) = (U_n(\beta(t_1), t_1), U_n(\beta(t_2), t_2), \ldots, U_n(\beta(t_d), t_d))^T.$$

As an extension of the limit theorem at a fixed point t, we prove that the distribution of $\boldsymbol{\mathcal{U}_n}(\beta_0, \underline{t})$ tends to a multivariate normal distribution with zero expectation and covariance matrix $\mathcal{S}(\beta_0, \underline{t})$. To formulate this statement and the consequences we make use of the following assumptions

A1 The coefficient function β_0 is twice continuously differentiable on $[0, \tau]$.

A2 The baseline function λ_0 is twice continuously differentiable on $[0, \tau]$.

B1 There exists a compact set $\mathcal{B}$ in $\mathbb{R}$ that includes a neighborhood of $\beta_0(t)$ for $t \in [0, \tau]$. Further, $s_j(\beta, t) = \mathsf{E} S_{nj}(\beta, t)$ exists for $j = 0, 1, 2$ and

$$|S_{nj}(\beta, t) - s_j(\beta, t)| = O_{\mathsf{P}}(n^{-1/2}) \text{ uniformly in } (\beta, t) \in \mathcal{B} \times [0, \tau]$$

B2 The functions s_j, $j = 0, 1, 2$, and their partial derivatives with respect to β are continuous in $\mathcal{B} \times [0, \tau]$.

B3 The functions $s_j(\beta_0(\cdot), \cdot)$ and $s_j(\beta, \cdot)$ for $j = 0, 1$ are twice differentiable with respect to $t \in [0, \tau]$.

B4 The function s_2 is bounded and s_0 is bounded away from zero.

C1 The function K is a symmetric density with bounded support, say $[-1, 1]$.

C2 The bandwidth sequence satisfies $h = h_n$

$$h_n \to 0 \quad \text{and} \quad n h_n^{1/5} \to \infty.$$

The asymptotic variance of the vector is characterized by the the function

$$v(\beta, t) = \frac{s_2(\beta, t)}{s_0(\beta, t)} - e(\beta, t)^2 \qquad \text{with} \quad e(\beta, t) = \frac{s_1(\beta, t)}{s_0(\beta, t)}.$$

Theorem 1. *Suppose that the assumptions* A1, A2, B1 − B4, C1, C2 *are satisfied. If* $v(\beta_0(t_j), t_j) > 0$ *for all* $j = 1, \ldots, d$. *Then*

$$\boldsymbol{\mathcal{U}_n}(\beta_0, \underline{t}) \xrightarrow{\mathcal{D}} \mathsf{N}_d(0, \mathcal{S}(\beta_0, \underline{t})),$$

where

$$\mathcal{S}(\beta_0, \underline{t}) = \operatorname{diag}(\sigma^2(\beta_0, t_1), \ldots, \sigma^2(\beta_0, t_d))$$

and

$$\sigma^2(\beta_0, t_j) = \kappa^2\, v(\beta_0(t_j), t_j) s_0(\beta_0(t_j), t_j) \lambda_0(t_j)$$

with $\kappa^2 = \int K^2(u) \mathrm{d}u$.

The asymptotic normality is shown for the stochastic parts

$$\tilde{U}_n(\beta_0, t_j) = n^{-1/2}h^{1/2}\sum_{i=1}^{n}\int_0^\tau K_h(s-t_j)[X_i - E_n(\beta_0, s)]\mathrm{d}M_i(s);$$

the smoothness conditions on the underlying functions and the convergence behavior of the bandwidth ensure that the remaining parts

$$n^{-1/2}h^{1/2}\sum_{i=1}^{n}\int_0^\tau K_h(s-t_j)[X_i - E_n(\beta_0, s)]\mathrm{d}A_i(s)$$

can be neglected. Moreover, for $h_n \to 0$ the components of the vector $\boldsymbol{\mathcal{U}_n}(\beta_0, \underline{t})$ are asymptotically independent.

Now, consider the weighted quadratic form

$$\mathcal{T}_n(\beta_0) = \boldsymbol{\mathcal{U}_n}(\beta_0, \underline{t})^T \mathcal{S}^{-1}(\beta_0, \underline{t})\,\boldsymbol{\mathcal{U}_n}(\beta_0, \underline{t}) = \sum_{j=1}^{d} U_n^2(\beta_0(t_j), t_j)\sigma^{-2}(\beta_0, t_j).$$

From the asymptotic normality of the vector $\boldsymbol{\mathcal{U}_n}(\beta_0, \underline{t})$ we obtain the following corollary:

Corollary 1. *Under the assumptions of Theorem 1*

$$\mathcal{T}_n(\beta_0) \xrightarrow{\mathcal{D}} \chi_d^2.$$

The variance matrix $\mathcal{S}$ is unknown. It depends on the unknown limits of the sums S_{nk}, on the function β_0 and on the baseline λ_0. A consistent estimator of $\mathcal{S}(\beta_0, \underline{t})$ is given by

$$\widehat{\mathcal{S}_n(\underline{t})} = \mathrm{diag}(\hat{\sigma}_n^2(t_1), \ldots, \hat{\sigma}_n^2(t_d))$$

with

$$\hat{\sigma}_n^2(t_j) = \kappa^2\,\frac{1}{n}\sum_{i=1}^{n}\int K_h(u-t_j)V_n(\hat{\beta}_n(t_j), u)\mathrm{d}N_i(u)$$

where

$$V_n(\beta, t) = \frac{S_{n2}(\beta, t)}{S_{n0}(\beta, t)} - \left(\frac{S_{n1}(\beta, t)}{S_{n0}(\beta, t)}\right)^2,$$

and $\hat{\beta}_n(t)$ is the estimator of $\beta_0(t)$. Thus, we have the following statement:

Corollary 2. *Under the assumptions of Theorem 1*

$$\widetilde{\mathcal{T}}_n(\beta_0) \xrightarrow{\mathcal{D}} \chi_d^2$$

where

$$\widetilde{\mathcal{T}}_n(\beta_0) = \boldsymbol{\mathcal{U}_n}(\beta_0, \underline{t})^T \widehat{\mathcal{S}}_n^{-1}(\underline{t})\,\boldsymbol{\mathcal{U}_n}(\beta_0, \underline{t}).$$

3 Score test based on the local partial likelihood approach

Consider now the hypothesis that the coefficient function $\beta_0(\cdot)$ has a parametric form, say $\beta_0(\cdot) = \beta(\cdot; \vartheta)$. That is we test

$$\mathcal{H} : \beta_0(\cdot) \in B_{\text{par}} = \{\beta(\cdot, \vartheta),\ \vartheta \in \Theta \subseteq \mathbb{R}^k\} \qquad \mathcal{K} : \beta(\cdot) \notin B_{\text{par}} \tag{3}$$

To estimate ϑ under $\mathcal{H}$ we use the partial likelihood method in the hypothetical model B_{par}

$$\lambda_i(t) = \lambda_0(t) \exp(\beta(t, \vartheta) X_i).$$

The partial likelihood function is given by

$$\tilde{\ell}(\vartheta) = \sum_{i=1}^{n} \int_0^{\tau} [\beta(s, \vartheta) X_i - \log(\sum_{j=1}^{n} Y_j(s) \exp(\beta(s, \vartheta) X_j))] \mathrm{d}N_i(s),$$

let W_n be the corresponding score vector, i.e., its component W_{nr} $r = 1, \ldots, k$ is

$$W_{nr}(\vartheta) = \frac{\partial \tilde{\ell}(\vartheta)}{\partial \vartheta_r} = \sum_{i=1}^{n} \int_0^{\tau} [X_i - E_n(\beta(s, \vartheta), s)] \dot{\beta}_r(s, \vartheta) \mathrm{d}N_i(s)$$

where $\dot{\beta}_r(t, \vartheta)$ is the partial derivative of $\beta(t, \vartheta)$ with respect to ϑ_r. The estimator $\hat{\vartheta}_n$ is the solution of the system of equations

$$W_{nr}(\vartheta) = 0 \quad r = 1, \ldots, k. \tag{4}$$

If the estimator $\hat{\vartheta}_n$ is $\sqrt{n}$-consistent, we can apply $\hat{\vartheta}_n$ in the test procedure. To verify $\sqrt{n}$-consistency we will show that $\hat{\vartheta}$ is asymptotically normal. The proof is based on the following steps: If the partial likelihood function $\tilde{\ell}$ is strictly concave, then the solution to (4) is unique. The consistency follows by showing, that the partial likelihood function converges to a concave function with a unique maximum at the underlying ϑ_0. To obtain the rate of convergence we consider the score function W_n as a local martingale and prove that the matrix of the minus second derivatives converges to a positive definite matrix.

Now, consider the test problem (3) and assume that the hypothesis $\mathcal{H}$ is true, i.e., there exists a ϑ_0 such that $\beta_0(\cdot) = \beta(\cdot, \vartheta_0)$. We estimate ϑ_0 by the maximum partial likelihood method and obtain the following result:

Theorem 2. *Suppose that the hypothetical functions in B_{par} have continuous partial derivatives of second order with respect to ϑ. Further assume that the partial likelihood function is strictly concave. Let the assumptions A2 and B1-B4 be satisfied. Then*

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = O_{\mathsf{P}}(1).$$

Based on Theorem 2, it can be shown

$$\widehat{\mathcal{T}}_n \xrightarrow{\mathcal{D}} \chi^2_d \tag{5}$$

where

$$\widehat{\mathcal{T}}_n = \widehat{\boldsymbol{\mathcal{U}}_{\boldsymbol{n}}}(\beta_{\hat{\vartheta}}, \underline{t})^T\, \widehat{\mathcal{S}}_n^{-1}(\underline{t})\, \widehat{\boldsymbol{\mathcal{U}}_{\boldsymbol{n}}}(\beta_{\hat{\vartheta}}, \underline{t})$$

with

$$\widehat{\boldsymbol{\mathcal{U}}_{\boldsymbol{n}}}(\beta_{\hat{\vartheta}}, \underline{t}) = (U_n(\beta(t_1, \hat{\vartheta}_n), t_1), U_n(\beta(t_2, , \hat{\vartheta}_n), t_2), \ldots, U_n(\beta(t_d, , \hat{\vartheta}_n), t_d))^T.$$

Limit statement (5) implies the following asymptotic test procedure
Reject $\mathcal{H}$, iff

$$\widehat{\mathcal{T}}_n \geq \chi^2_{d;1-\alpha}. \tag{6}$$

Special case As a special case we consider the problem of testing whether the classical Cox proportional hazards model is true, that is, B_{par} is the set of all constants ϑ.

$$\mathcal{H} : \beta_0(\cdot) \equiv \vartheta \;\text{ for some constant } \vartheta.$$

If the assumptions of Theorem 1 are satisfied, then the test procedure is given by (6) with $\widehat{\mathcal{S}}_n^{-1}(\underline{t})$ as defined above and $\widehat{\boldsymbol{\mathcal{U}}_{\boldsymbol{n}}}(\beta_{\hat{\vartheta}}, \underline{t}) = \widehat{\boldsymbol{\mathcal{U}}_{\boldsymbol{n}}}(\widehat{\vartheta}_n, \underline{t})$ with elements

$$U_n(\widehat{\vartheta}_n, t_j) = n^{-1/2} h^{1/2} \sum_{i=1}^n \int_0^\tau K_h(s - t_j)[X_i - E_n(\widehat{\vartheta}_n, s)]\mathrm{d}N_i(s) \quad j = 1, \ldots, d,$$

where $\widehat{\vartheta}_n$ is the maximum partial likelihood estimator in the hypothetical model.

References

Cai Z., Sun Y. (2003). Local Linear Estimation for Time-Dependent Coefficients in Cox's Regression Models. *Scandinavian Journal of Statistics.* 30, 93-111

Reliability Approach in Accelerated Lifetime Modeling for Econometrics Studies in Presence of Time-Depending Covariates

VILIJANDAS BAGDONAVICIUS[1], JIA SHEN[2] AND MIKHAIL NIKULIN[3]

[1] *Department of Statistics, University of Vilnius, Lithuania*

[2] *School of Management, Fudan University, Shanghai, China*

[3] *Université Victor Segalen Bordeaux 2, Bordeaux, France*

Abstract

This paper describes the semiparametric dynamic regression or accelerated life models that are very important in econometric duration analysis in estimation of the risk of defaults, which plays an important role in the pricing and hedging of credit risk see Gouriéroux and Josiak (2005). Dynamic regression models are applied often in economics, reliability and survival analysis, see, for example, Lancaster (1979), Bagdonavicius and Nikulin (2002), Ceci and Mazliak (2004), Martinussen and Scheike (2006), Zeng and Ling (2007), etc. Evident that this approach can be very useful to modelling the default probabilities. Accelerated life models relate lifetime distribution of the default time to the time varying explanatory variables, called in reliability *stresses*, it terms of which is described the past performance of the firms and the banks the information about the current market conditions or about some important economic, political and social factors which influence on the risk of default. These models are used for estimation of the effects of covariates (stresses) over the time on survival and for estimation of survival via its effects on default rates under given covariates values. In terms of the time dependent covariates are described the possible direct and indirect economic (financial) loss for firms, or as one can say, conditional on reasonable available information, which have to be taken in consideration in business risk analysis. For example, the time depending stresses and degradation models can explain the influence of such characteristics as quality, productivity, credibility, profitability of firms, or the dramatic decline in oil price in the market, or the business cyclic effects on default rates.

Keywords: Accelerated life model, accelerated hazards, business risk, Cox model, econometric duration analysis, time varying explanatory variable, risk of defaults, survival data, semiparametric dynamic regression models.

1 Introduction

This paper describes the semiparametric dynamic regression or accelerated life models that are very important in econometric duration analysis in estimation of the risk of defaults, which plays an important role in the pricing and hedging of credit risk see Lancaster (1979), Horowitz (1998), Mosler (2002), Duffie and Singleton (2003), Kiefer (1988), Gouriéroux and Josiak (2005), Ceci and Mazliak (2004), Royston and Parmar (2002), Zhou, Chinnam and Korostelev (2012), Royston and Lambert (2011), etc... Dynamic regression models are applied often in reliability and survival analysis, see, for example, Bedford and Cooke (2001), Bagdonavicius and Nikulin (2002), Ceci and Mazliak (2004), Martinussen and Scheike (2006), Nikulin, Gerville-Reache, Couallier (2007), Bagdonavicius, Kruopis and Nikulin (2011), Voinov, Balakrishnan and Nikulin (2013), etc. Evident that this approach can be very useful to modelling the default probabilities. Accelerated life models relate lifetime distribution of the default time to the time varying explanatory variables, called in reliability *stresses*, it terms of which is described the past performance of the firms and the banks the information about the current market conditions or about some important economic, political and social factors which influence on the risk of default. These models are used for estimation of the effects of covariates (stresses) over the time on survival and for estimation of survival via its effects on default rates under given covariates values. In terms of the time dependent covariates are described the possible direct and indirect economic (financial) loss for firms, or as one can say, conditional on reasonable available information, which have to be taken in consideration in business risk analysis. For example, the time depending stresses and degradation models can explain the influence of such characteristics as quality, productivity, credibility, profitability of firms, or the dramatic decline in oil price in the market, or the business cyclic effects on default rates. The reliability approach based on applications of semiparametric dynamic regression and degradation models provides a basis for some suggestions for further research on statistical estimation and prediction of the default risk and gives an interesting possibility approach to obtain a statistical inference in dependence on situation in the market (see Nikulin et al.,(2009), Couallier et al., (2014)). The considered models are very flexible and are applicable to estimate possible financial losses of different types of firms in the real world economic, financial and politic situations, described in terms of time dependent stresses. Using the terminology of Singpurwalla (1995) we have the possibility to estimate the probability of default risk in *dynamic environments.* The *proportional hazards model* is the most important model in duration analysis. We consider some recent models based on the *Cox model.* The proportional hazards model is generalized by assuming that at any moment the ratio of hazard rates is depending not only on values of time-varying *covariates* (*stresses*) but also on *resources* used until this moment. Relations with generalized multiplicative, modified proportional hazards, *frailty, linear transformation, Sedyakin* are considered. We consider semiparametric models for longitudinal studies the relations between a longitudinal response process and a time-to-event. We consider also the models with *cross-effects of survival functions.* These models are applied for longitudinal studies of the economic and industrial data by Hsieh (2001) , Huber *et al.*, (2007), Wu (2007), Nikulin and Wu (2007), Bagdonavicius and Nikulin (2002), Bagdonavicius, Kruopis and Nikulin (2011), Couallier et al.,(2013) . We discuss also the applications of the so-called *degradation models*, which are very useful in economics and business to make a comprehensive risk analysis when economic damage grow. Such models allow assessing the probability of specific traumatic events and their impact on business (default) process. These models are well adapted for statistical analysis of industrial firms, insurance companies, banks failure data (bankruptcy) in dynamic environments, to qualitatively and quantitavely estimate possible financial and economic losses and damage due to economic, social, politic, etc... changes over the time.

The explanatory variables (stress) may be modelled by stochastic processes, deterministic time functions or constants (possibly different for different individuals). Denote by $x(\cdot) = (x_1(\cdot), ..., x_m(\cdot))^T : [0, \infty) \to \mathbf{R}^m$, a deterministic time function (possibly multidimentional) which is a vector of covariates itself or a realisation of a stochastic process $X(\cdot) = (X_1(\cdot), ..., X_m(\cdot))^T$ when covariates (stresses) are modeled by this stochastic process. We denote $E = E\{x(\cdot)\}$ a set of all possible or admissible stresses. If a stress $x(\cdot)$ is constant in time, $x(t) \equiv x$, then we shall write x instead of $x(\cdot)$. We denote E_1 a set of all constant in time stresses, $E_1 \subset E$.

The distribution of survival under covariates can be defined by the survival, cumulative distribution, or probability density function. Nevertheless, the sense of models is best seen if they are formulated in terms of so-called *hazard rate function.* This notion is used widely in reliability and survival analysis. In econometrics , and in particular in credit analysis, instead of the *hazard rate* function people use the term *forward default rate* function or more simple term *default rate function.*

Denote by T the *time to default.* Then the *probability of surviving function* given stress $x(\cdot)$ is defined as

$$S_{x(\cdot)}(t) = \mathbf{P}\{T > t \mid x(u), 0 \le u \le t\}, \quad t > 0, \quad x(\cdot) \in E,$$

with $S_{x(\cdot)}(0) = 1$ for any stress $x(\cdot)$ from the set E of all admissible stresses. So for any $t > 0$ the value $S_{x(\cdot)}(t)$ denote the probability that the firm will not default for at least t years, if we measure the time in the years, for example,

The *default rate function* or *intensity of default function* under given stress $x(\cdot)$ is defined as

$$\lambda_{x(\cdot)}(t) = \lim_{h \downarrow 0} \frac{1}{h} \mathbf{P}\{T \in [t, t+h) \mid T \ge t, x(u), 0 \le u \le t\} = -\frac{S'_{x(\cdot)}(t)}{S_{x(\cdot)}(t)}.$$

From this definition it follows for any stress $x(\cdot) \in E$ and any $t, t > 0$, the value $\lambda_{x(\cdot)}(t)$ is the rate of default arrival at time t conditional only on survival up to time t. The default rate function is the most important reliability characteristics of survival and its value $\lambda_{x(\cdot)}(t)$ gives the *instantaneous exit rate* per unit of time evaluated at the time t. It is evident also that if the function $\lambda_{x(\cdot)}(\cdot)$ is continuous in t, then under the stress $x(\cdot)$ the probability of default in the interval $[t, t+\Delta]$ for small $\Delta, \Delta > 0$, conditional on survival to t, is approximately equal to $\lambda_{x(\cdot)}(t)\Delta$. We note here that sometimes $\lambda_{x(\cdot)}(\cdot)$ is called also the *forward default rate*, see Lancaster (1972), Duffie and Singleton (2003), the *bankruptcy rate* or *failure rate of instruments*, see Gouriéroux and Josiak (2005). The default rate function is the most important reliability characteristics. Denote by

$$\Lambda_{x(\cdot)}(t) = \int_0^t \lambda_{x(\cdot)}(u)du = -\ln\{S_{x(\cdot)}(t)\}, \quad x(\cdot) \in E,$$

the *cumulative rate of default* under stress $x(\cdot)$. For any $x(\cdot) \in E$ the function $\Lambda_{x(\cdot)}(\cdot)$ is increasing in t, with $\Lambda_{x(\cdot)}(0) = 0$, and $\Lambda_{x(\cdot)}(+\infty) = +\infty$.

Each specified model relates the hazard rate (or survival function) to the explanatory variable in some particular way. From this definition it follows immediately that

$$S_{x(\cdot)}(t) = e^{-\Lambda_{x(\cdot)}(t)} = \exp\{-\int_0^t \lambda_{x(\cdot)}(u)du\}, \quad x(\cdot) \in E.$$

At the end of this section we note that we write $T_{x(\cdot)}$ instead of T to remind that we study the time to default under the stress $x(\cdot)$, and hence the distribution of time to default depends on $x(\cdot), x(\cdot) \in E$. We want to consider here the models on E which are well adapted to study the microstructure of financial markets in terms of observed stresses.

2 The Cox or the proportional default rate model

Under the *proportional default rate* model (traditionally PH model or Cox (1972) model) on E the defaul rate under a stress $x(\cdot)$ has the form

$$\lambda_{x(\cdot)}(t) = r\{x(t)\}\, \lambda_0(t), \quad x(\cdot) \in E, \tag{1}$$

where $\lambda_0(t)$ is a so called *baseline default rate function*, and $r(\cdot)$ is a positive function on E.

The model implies that the ratio $R(t, x_1, x_2)$ of default rates under different fixed constant stresses x_1 and x_2 is constant over time:

$$R(t, x_1, x_2) = \frac{\lambda_{x_2}(t)}{\lambda_{x_1}(t)} = \frac{r\{x_2\}}{r\{x_1\}} = const.$$

In most applications the function r is parametrized in the form

$$r(x) = \exp\{\beta^T x\}, \quad \text{where} \quad \beta = (\beta_1, \cdots, \beta_m)^T$$

is the vector of regression parameters. Under this parametrization we obtain the classical semiparametric Cox model with time-dependent covariables:

$$\lambda_{x(\cdot)}(t) = e^{\beta^T x(t)} \lambda_0(t), \quad t > 0, \quad x(\cdot) \in E. \tag{2}$$

Usually the Cox model is considered as semiparametric: the finite-dimensional parameter β and the baseline hazard function λ_0 are supposed to be completely unknown. Nevertheless, non-parametric estimation procedures when the function r is also supposed to be unknown are sometimes used. Parametric estimation procedures when λ_0 is taken from some parametric class of functions is scarcely used because the *parametric accelerated failure time model* (see in the following sections) is also simple for analysis and more natural. In parametric case we recommend to chose as parametric family for the baseline function the so-called Power Generalized Weibull

(PGW) Family of Distributions, proposed by Bagdonavicius and Nikulin (2002). In terms of the survival functions the PGW family is given by the next formula:

$$S(t,\sigma,\nu,\gamma) = exp\left\{1 - \left[1 + \left(\frac{t}{\sigma}\right)^{\nu}\right]^{\frac{1}{\gamma}}\right\}, \quad t > 0, \gamma > 0, \nu > 0, \sigma > 0.$$

If $\gamma = 1$ we have the Weibull family of distributions. If $\gamma = 1$ and $\nu = 1 = 1$, we have the exponential family of distributions. This class of distributions has very nice probability properties. All moments of this distribution are finite. In dependence of parameter values the hazard rate can be *constant, monotone* (increasing or decreasing), *unimodal* or $\bigcap$-shaped, and bathtub or $\bigcup$-shaped. At the beginning of a firm's life, it has a great risk of failure because of bad market investigation, absence of management experiences, etc. When this initial period known as *birn in period* is passed, the firm has less risks of bankruptcy and win the market. It is a *period of prosperity.* The hazard function $\lambda(\cdots)$ is almost constant which corresponds to the Exponential Distribution. In its end, the firm will undergo competing risks. In this description of its life cycle, its hazard function is U-shaped. The PGW distribution family corresponds to this kind of modelling needs. Another interesting family, is the so-called the *Exponentiated Weibull Family* of distributions, which was proposed by Mudholkar & Srivastava (1995).

The Cox model is not much used analysing failure time regression data in reliability. The cause is that the model is not natural when subjects are aging. Indeed, from (1) it follows that for any t the default rate function under the time-varying stress $x(\cdot)$ at the moment t *does not depend on the values of the stress* $x(\cdot)$ *before the moment* t but only on the value of it at this moment:

$$\mathbf{P}(T \leq t + s \mid T > t) = 1 - e^{-\int_t^{t+s} e^{\beta^T x(u)} \lambda_0(u) du},$$

where λ_0 is the baseline hazard function which does not depend on stress. For this reason we can say that PH model has the *absence of memory property.* Nevertheless, in survival analysis the Cox model usually works quite well, because the values of covariates under which estimation of survival is needed are in the range of covariate values used in experiments. So the use of a not very exact but simple model often is preferable to the use of more adequate but complicated model. It is similar with application of linear regression models in classical regression analysis: the mean of dependent variable is rarely a linear function of independent variables but the linear approximation works reasonably well in some range of independent variable values.

In reliability, accelerated life testing in particular, the choice of a good model is much more important than in survival analysis. For example, in accelerated life testing units are tested under accelerated stresses which shorten the life. Using such experiments the life under the usual stress is estimated using some regression model. The values of the usual stress is not in range of the values of accelerated stresses, so if the model is misspecified, the estimators of survival under the usual stress may be very bad.

If on the bases of graphical analysis or goodness-of-fit tests the PH model is rejected and one has a reason to suppose that the ratios of hazard rates are not constant, other models should be used.

3 Accelerated Failure Time Model

The PH model has the absence of memory propriety: the hazard rate at any moment does not depend on the values of the stress before this moment. It is more natural to suppose that the default rate at any moment t should depend not only on the value of stress at this moment but on the probability to survive up to this moment. Under stress $x(\cdot)$ this probability is $S_{x(\cdot)}(t)$. It characterizes the summing effect of values of stress (of the history) in the interval $[0,t]$ on survival. The equality $\Lambda_{x(\cdot)}(t) = -\ln S_{x(\cdot)}(t)$ implies that the cumulative default rate also characterizes this summing effect. So it can be supposed that the default rate at any moment t is a function of the value $x(t)$ of a stress and the value of the cumulative default rate $\Lambda_{x(\cdot)}(t)$.

The generalized Sedyakin's model namely supposes it (see Sedyakin (1966), Bagdonavičius (1978), Bagdonavičius & Nikulin (1998)):

$$\lambda_{x(\cdot)}(t) = g\left(x(t), \Lambda_{x(\cdot)}(t)\right). \tag{3}$$

This model with g completely unknown is too general to do statistical inference. But if we choose some regression model for constant covariates, the form of the function g can be made more concrete.

Suppose that under different constant covariates $x \in E_0$ the survival functions differ only in scale:

$$S_x(t) = S_0\left(r(x)t\right), \tag{4}$$

If the GS model holds on a set $E, E_0 \subset E$ of covariates then (4) holds on E_0 if and only if the function g has the form $g(x,s) = r(x)q(s)$ (see Bagdonavičius & Nikulin (1998)).

We obtain the following model:

$$\lambda_{x(\cdot)}(t) = r\{x(t)\}\, q\{\Lambda_{x(\cdot)}(t)\}. \tag{5}$$

Solving this differential equation with respect to $\Lambda_{x(\cdot)}(t)$, and using the relation between the survival and the cumulative hazard functions we obtain that the survival function has the form

$$S_{x(\cdot)}(t) = S_0\left(\int_0^t r(x(u))du\right), \tag{6}$$

where the function S_0 does not depend on $x(\cdot)$. The function r changes locally the time-scale.

The model (6) (or, equivalently, (5)) is called *accelerated failure time* (AFT) model.

The function r is often parametrized in the following form:

$$r(x) = e^{-\beta^T x},$$

where $\beta = (\beta_1, \cdots, \beta_m)^T$ is a vector of unknown parameters.

Under the parametrized AFT model the survival function is

$$S_{x(\cdot)}(t) = S_0\left(\int_0^t e^{-\beta^T x(u)} du\right), \tag{7}$$

and the default rate is

$$\lambda_{x(\cdot)}(t) = e^{-\beta^T x(t)}\ \lambda_0\left(\int_0^t e^{-\beta^T x(u)} du\right), \tag{8}$$

and for constant covariates

$$S_x(t) = S_0\left(e^{-\beta^T x}\, t\right).$$

So in the case of constant covariates the AFT model can also be written as a loglinear model, since the logarithm of the failure time T_x under constant covariate x can be written as

$$\ln\{T_x\} = \beta^T x + \varepsilon, \tag{9}$$

where the survival function of the random variable ε does not depend on x and is $S(t) = S_0(\ln t)$. In the case of lognormal failure-time distribution the distribution of ε is normal and we have the standard linear regression model. The equality (8) implies that if the survival function under any constant covariate belongs to parametric families such as Weibull, loglogistic, lognormal, then the survival function under any other constant covariate also belongs to that family.

Differently from PH model, the AFT model is mostly applied in survival analysis as a parametric model: the function S_0 (or the distribution of ε) is taken from some parametric class of distributions and the parameters to estimate are the parameters of this class and the regression parameters β.

In the case of semiparametric estimation the function S_0 is supposed to be completely unknown and the regression parameters as the function S_0 are the parameters to estimate in the model (7). The semiparametric AFT model is much less used in survival analysis then the Cox model because of complicated estimation procedures: modified variants of likelihood functions are not differentiable and even not continuous functions, the limit covariance matrices of the normed regression parameters depend on the derivatives of the probability density functions, so their estimation is complicated.

The parametric AFT model is used in failure time regression analysis and accererated life testing. Under special experiment plans even non-parametric estimation procedures are used. In such a case not only the function S_0 but also the function r in the model (6) would be completely unknown. Among many effective risk analysis models, accelerated life time model presents itself for its good properties in duration analysis in finance market research, for example.

The AFT model is a good choice when the lifetime distribution class is supposed to be known. Nevertherless, it is as restrictive as the PH model. The assumption that the survival distributions under different covariate values differ only in scale is rather strong assumption. So more sophisticated models are also needed.

4 Generalized proportional hazards model

4.1 Definitions

The AFT and PH models are rather restrictive.

Under the PH model lifetime distributions under constant covariates are from the narrow class of distributions: the ratio of the default rates under any two different constant covariates is constant over time.

Under the AFT model the covariate changes (locally, if the covariate is not constant) only the scale.

Generalized proportional hazards (GPH) models allow the ratios of the default rates under constant covariables to be not only constant but also increasing or decreasing. They include AFT and PH models as particular cases.

As was discussed in the previous section, the survival function $S_{x(\cdot)}(t)$ (or, equivalently, the cumulative rate of default function $\Lambda_{x(\cdot)}(t)$) characterizes the summing effect of stress values in the interval $[0,t]$ on survival. So suppose that the default rate function at any moment t is proportional not only to a function of the covariate applied at this moment and to a baseline default rate, but also to a function of the probability of survival until t (or, equivalently, to the cumulative rate of default at t):

$$\lambda_{x(\cdot)}(t) = r\{x(t)\}\; q\{\Lambda_{x(\cdot)}(t)\}\; \lambda_0(t). \tag{10}$$

We call the model (10) the generalized proportional hazards (GPH) model, see Bagdonavičius V. and Nikulin M (1999). Particular cases of the GPH model are the PH model ($q(u) \equiv 1$) and the AFT model ($\lambda_0(t) \equiv \lambda_0 = const$).

Under the GPH model the survival functions $S_{x(\cdot)}$ have the form

$$S_{x(\cdot)}(t) = G\left\{\int_0^t r(x(\tau))d\Lambda_0(t)\right\}, \tag{11}$$

where

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du, \quad G = H^{-1}, \quad H(u) = \int_0^{-\ln u} \frac{dv}{q(v)}.$$

We denote by H^{-1} the function inverse to G.

4.2 Relations with the linear transformations and frailty models

Models of different levels of generality can be obtained by completely specifying q, parametrizing q, or considering q as unknown.

Completely specifying q we obtain rather strict models which are alternatives to the PH model and the field of their application is relatively narrow (see Bagdonavicius and Nikulin (1994)). Under constant stresses such models are the *linear transformation* (LT) models. Indeed, if q is specified and r is parametrized by $r(x) = e^{\beta^T x}$ then under constant stresses the survival functions have the form $S_{x(\cdot)}(t) = G\left\{e^{\beta^T x}\Lambda_0(t)\right\}$ with G specified. This implies that the random variable T_x can be transformed by the function $h(t) = \ln\{H(S_0(t))\}$ to the random variable of the form

$$h(T_x) = -\beta^T x + \varepsilon, \tag{12}$$

where ε is a random error with the parameter-free distribution function $Q(u) = 1 - G(e^u)$. It is the *linear transformation* (LT) model of Dabrowska and Doksum (1988). Examples of the LT models:

1) PH model (G is a Weibull survival function, ε has the extreme value distribution);

2) logistic regresion model (G is a loglogistic survival function, ε has the loglogistic distribution):

$$\frac{1}{S_x(t)} - 1 = r(x)\left(\frac{1}{S_0(t)} - 1\right).$$

3) generalized probit model (G is a lognormal survival function, has the normal distribution):

$$\Phi^{-1}\left(S_x(t)\right) = \ log\left(r(x)\right) + \Phi^{-1}\left(S_0(t)\right),$$

where Φ is the standard normal cumulative distribution function.

The last two models are alternatives to the PH model. They are widely used for analysis of dichotomous data when the probability of "success" in dependence of some factors is analyzed. If application of the PH model is dubious then better is to use a (not very) wider GPH model which is obtained from the general GPH model not by complete specification of the function q but taking a simple parametric model for it.

Let us consider relations between the GPH models and the
frailty models (Hougaard(1986)) with covariates.

The hazard rate can be influenced not only by the observable stress $x(\cdot)$ but also by a non-observable positive random covariate Z, called the *frailty variable.* Suppose that the default rate given the frailty variable value is

$$\lambda_{x(\cdot)}(t|Z = z) = z\, r(x(t))\, \lambda_0(t).$$

Then

$$S_{x(\cdot)}(t) = \mathbf{E}\ exp\{-Z\int_0^t r(x(\tau))\, d\Lambda_0(\tau)\} = G\{\int_0^t r(x(\tau))d\Lambda_0(\tau)\},$$

where $G(s) = \mathbf{E}e^{-sZ}$.

So the GPH model can be defined by specification of the frailty variable distribution. All considered here models are used often in unemployment studies, for example.

4.3 The GPH models with monotone hazard ratios

The following parametrizations of r and q give submodels of the GPH model with monotone ratios of default rates under constant covariates. Using only one parameter and power or exponential functions for function q parametrization several important models are obtained.

4.3.1 The first GPH model

Suppose that $q(0) = 1$ (if it is not so, we can include $q(0)$ in λ_0, which is considered as unknown). Taking a power function $q(u) = (1+u)^{-\gamma+1}$ and $r(x) = e^{\beta^T x}$ we obtain the first GPH model:

$$\lambda_{x(\cdot)}(t) = e^{\beta^T x(t)}(1 + \Lambda_{x(\cdot)}(t))^{-\gamma+1}\lambda_0(t). \tag{13}$$

It coincides with the PH model when $\gamma = 1$. The supports of the survival functions $S_{x(\cdot)}$ are $[0, \infty)$ when $\gamma \geq 0$ and $[0, sp_{x(\cdot)})$ with finite right ends $sp_{x(\cdot)}$, $sp_{x(\cdot)} < \infty$, when $\gamma < 0$. Finite supports are very possible in accelerated life testing: failures of units at different accelerated stresses are concentrated in intervals with different finite right limits.

Suppose that at the point $t = 0$ the ratio $R(t, x_1, x_2)$ of the default rates under constant stresses x_1 and x_2 is greater then 1:

$$R(0, x_1, x_2) = \frac{r(x_2)}{r(x_1)} = c_0 > 1.$$

The ratio $R(t, x_1, x_2)$ has the following properties:

a) if $\gamma > 1$, then the ratio of the default rates decreases from the value $c_0 > 1$ to the value $c_\infty = c_0^{\frac{1}{\gamma}} \in (1, c_0)$, i.e. the hazard rates approach one another when t increases.

b) if $\gamma = 1$ (PH model), the ratio of the default rates is constant.

c) if $0 \leq \gamma < 1$, then the ratio of the default rates increases from the value $c_0 > 1$ to the value $c_\infty = c_0^{\frac{1}{\gamma}} \in (c_0, \infty)$, i.e. the default rates go away one from another when t increases.

d) if $\gamma < 0$, then the ratio of the default rates increases from the value $c_0 > 1$ to ∞, end the infinity is attained at the point $sp_{x_2} = \Lambda_0^{-1}\{-1/((\gamma)r(x_2))\}$. The default rates go away one from another quickly when t increases.

The first GPH model is a generalization of the *positive stable frailty model with explanatory variables*: the GPH model with $\gamma = 1/\alpha > 0$ is obtained taking the frailty variable Z which follows the *positive stable distribution* with the density

$$p_Z(z) = -\frac{1}{\pi z} exp\{-\alpha z + 1\} \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \sin(\pi\alpha k) \frac{\Gamma(\alpha k + 1)}{z^{\alpha k}}, \quad z > 0,$$

where α is a *stable index*, $0 < \alpha < 1$.

4.3.2 The second GPH model

Under the first GPH model the support of the survival functions is infinite when $\gamma \geq 0$ and finite when $\gamma < 0$. The limit is $\gamma = 1$. So it is interesting to take a model with the following parametrization: $q(u) = (1 + \gamma u)^{-1}$. We obtain the second GPH model:

$$\lambda_{x(\cdot)}(t) = e^{\beta^T x(t)}(1 + \gamma\Lambda_{x(\cdot)}(t))^{-1}\lambda_0(t), \quad (\gamma \geq 0). \tag{14}$$

It also coincides with the PH model when $\gamma = 0$. The supports of the survival functions $S_{x(\cdot)}$ are $[0, \infty)$.

The ratio $R(t, x_1, x_2) = \lambda_{x_2}(t)/\lambda_{x_1}(t)$ has the following properties:

a) if $\gamma > 0$, then the ratio of the default rates decreases from $c_0 > 1$ to the value $\sqrt{c_0} \in (1, c_0)$, i.e. the default rates approach one another when t increases.

b) if $\gamma = 0$ (PH model), the ratio of the default rates is constant.

The second GPH model equivalent to the *inverse gaussian frailty model with explanatory variables*: the GPH model with $\gamma = (4\sigma\theta)^{1/2} > 0$ is obtained taking the frailty variable Z which follows the *inverse gaussian distribution* with the density

$$p_Z(z) = \left(\frac{\sigma}{\pi}\right)^{1/2} e^{\sqrt{4\sigma\theta}} z^{-3/2} e^{-\theta z - \frac{\sigma}{z}}, \quad z > 0.$$

4.3.3 The third GPH model

Taking the exponential function $q(u) = e^{-\gamma u}$ and $r(x) = e^{\beta^T x}$ we obtain the third GPH model:

$$\lambda_{x(\cdot)}(t) = e^{\beta^T x(t) - \gamma\Lambda_{x(\cdot)}(t)}\, \lambda_0(t). \tag{15}$$

It coincides with the PH model when $\gamma = 0$. The supports of the survival functions $S_{x(\cdot)}$ are $[0, \infty)$ when $\gamma \geq 0$ and $[0, sp_{x(\cdot)})$ with finite right ends when $\gamma < 0$.

Suppose that $R(0, x_1, x_2) = r(x_2)/r(x_1) = c_0 > 1$.

The ratio $R(t, x_1, x_2)$ has the following properties:

a) if $\gamma > 0$, then the ratio of the default rates decreases from the value $c > 0$ to 1, i.e. the default rates approach one another and meet at infinity.

b) if $\gamma = 0$ (PH model), the ratio of the default rates is constant.

c) if $\gamma < 0$, then the ratio of the default rates increases from the value $c_0 > 1$ to ∞, end the infinity is attained at the point $sp_{x_2} = \Lambda_0^{-1}\{-1/(\gamma r(x_2))\}$. The default rates go away one from another quickly when t increases.

The third GPH model is a generalization of the *gamma frailty model with explanatory variables*: the GPH model with $\gamma = 1/k > 0$ is obtained taking the frailty variable Z which follows the *gamma distribution* with the density

$$p_Z(z) = \frac{z^{k-1}}{\theta^k \Gamma(k)} e^{-z/\theta}, \quad z > 0.$$

All the three GPH models are considered as semiparametric: finite-dimensional parameters β and γ and unknown baseline function Λ_0 are the unknown parameters.

4.4 Regression models with cross-effects of survival functions

Let us consider models for analysis of data with cross-effects of survival functions under constant covariates.

4.5 First model with cross-effects of survival functions

The first model with cross-effects of survival functions (CE model) can be obtained from the first GPH model considered in the previous section replacing the scalar parameter γ by $e^{\gamma^T x(t)}$ in the formula (13), where γ is m-dimensional (see Bagdonavičius and Nikulin (2002)):

$$\lambda_x(t) = e^{\beta^T x(t)}\{1 + \Lambda_x(t)\}^{1-e^{\gamma^T x(t)}}\,\lambda_0(t), \quad \gamma = (\gamma_1, ..., \gamma_m)^T. \tag{16}$$

Suppose that at the point $t = 0$ the ratio of the default rates

$$R(t, x_1, x_2) = \lambda_{x_2}(t)/\lambda_{x_1}(t)$$

under constant covariates x_1 and x_2 is greater then 1:

$$R(0, x_1, x_2) = e^{\beta^T(x_2 - x_1)} = c_0 > 1 \quad \text{and} \quad \gamma^T(x_1 - x_2) < 0.$$

In this case the ratio $R(t, x_1, x_2)$ decreases from the value $c_0 > 1$ to 0, i.e. the hazard rates intersect once. The survival functions S_{x_1} and S_{x_2} also intersect once in the interval $(0, \infty)$ (more about see in Bagdonavičius and Nikulin (2002).)

Other CE models can be obtained using the same procedure for the second and the third GPH models.

4.6 Second CE-model

Hsieh (2001) considered the following model with cross effects of the survival functions generalization of the PH model

$$\Lambda_x(t) = e^{\beta^T x(t)}\{\Lambda_0(t)\}^{e^{\gamma^T x(t)}}. \tag{17}$$

It is a generalization of the PH model taking the power $e^{\gamma^T x(t)}$ of $\Lambda_0(t)$ instead of the power 1.

Note that the difference between this second model and the first CE model is the following. In the case of the second CE model the ratios of the default rates and even the ratios of the cumulative rate of defaults go to ∞ (or 0) as $t \to 0$. In the case of the first CE model these ratios are defined and finite at $t = 0$. This property of the first CE model is more natural and helps avoid complications when seeking efficient estimators.

4.7 Changing shape and scale models

Natural generalization of the AFT model (4) is obtained by supposing that different constant stresses x influence not only the scale but also the shape of survival distribution, see Mann et al (1974):

$$S_x(t) = S_0\left\{\left(\frac{t}{\sigma(x)}\right)^{\nu(x)}\right\},$$

where σ and ν some positive functions on E_1. Generalization of this model to the case of time-variale covariates is the *changing shape and scale* (CHSS) model, Bagdonavičius and Nikulin (1999):

$$S_{x(\cdot)}(t) = S_0\left(\int_0^t r\{x(u)\}u^{\nu(x(u))-1}du\right). \tag{18}$$

In this model the variation of stress changes locally not only the scale but also the shape of distribution.

In terms of the default rate functions the model can be written in the form:

$$\lambda_{x(\cdot)}(t) = r\{x(t)\}\; q(\Lambda_{x(\cdot)}(t))\, t^{\nu(x(t))-1}, \tag{19}$$

where $q(u) = \lambda_0(\Lambda_0^{-1}(u)),\ \Lambda_0(t) = -\ln S_0(t),\ \ \lambda_0(t) = A_0'(t)$.

If $\nu(x) \equiv 1$ then the model coincides with the AFT model with $r(x) = 1/\sigma(x)$. The CHSS model is not in the class of the GPH models because the third factor at the right of the formula (19) depends not only on t but also on $x(t)$.

The CHSS model is parametric, if S_0 is taken from some parametric class of survival functions and the functions r and ν are parametrized, usually taking $r(x) = e^{\beta^T x}$, $\nu(x) = e^{\gamma x}$. The model is semiparametric, if the function S_0 is considered as unknown and the functions r and ν are parametrized:

$$\lambda_{x(\cdot)}(t) = e^{\beta^T x(t)}\; q(\Lambda_{x(\cdot)}(t))\, t^{e^{\gamma^T x(t)}-1}, \tag{20}$$

For various classes of S_0 the CHSS model includes cross-effects of survival functions under constant covariates. For example, it is so, if the survival distribution under constant covariates is Weibull, loglogistic ($\Lambda_0(t) = t, \ln(1+t)$, respectively).

Parametric analysis can be done using the method of maximum likelihood. Semiparametric analysis is more complicated because the same problems as in the case of AFT semiparametric model arise: modified variants of likelihood functions are not differentiable and even not continuous functions, the limit covariance matrices of the normed regression parameters depend on the derivatives of the probability density functions.

5 Models with time-dependent regression coefficients

5.1 PH model with time dependent regression coefficients

Flexible models can be obtained by supposing that the regression coefficients β in the PH model (2) are time-dependent, i.e. taking

$$\lambda_{x(\cdot)}(t) = e^{\beta(t)^T x(t)} \lambda_0(t), \tag{21}$$

where

$$\beta^T(t)\, x(t) = \sum_{i=1}^{m} \beta_i(t) x_i(t).$$

If the function $\beta_i(\cdot)$ is increasing or decreasing in time then the effect of the ith component of the explanatory variable is increasing or decreasing in time.

The model (21) is the PH model with time-dependent regression coefficients.

Usually the coefficients $\beta_i(t)$ are considered in the form

$$\beta_i(t) = \beta_i + \gamma_i g_i(t), \quad (i = 1, 2, ..., m),$$

where $g_i(t)$ are some specified deterministic functions as $t, \ln t, \ln(1+t), (1+t)^{-1}$, for example, or realizations of predictable processes. In such a case the PH model with time dependent coefficients and constant or time dependent explanatory variables can be written in the usual form (2), where the role of the components of the "covariables" play not only the components $x_i(\cdot)$ but also $x_i(\cdot) g_i(\cdot)$. Indeed, set

$$\theta = (\theta_1, \cdots, \theta_{2m})^T = (\beta_1, \cdots, \beta_m, \gamma_1, \cdots, \gamma_m)^T,$$

$$z(\cdot) = (z_1(\cdot), \cdots, z_{2m}(\cdot))^T = (x_1(\cdot), \cdots, x_m(\cdot), x_1(\cdot) g_1(\cdot), \cdots, x_m(\cdot) g_m(\cdot))^T. \tag{22}$$

Then

$$\beta^T(u) x(u) = \sum_{i=1}^{m} (\beta_i + \gamma_i g_i(t))\, x_i(t) = \theta^T z(u).$$

So the PH model with time dependent regression coefficients of above given form can be written in the form

$$\lambda_{x(\cdot)}(t) = e^{\theta^T z(t)} \lambda_0(t), \quad t \geq 0. \tag{23}$$

We have the PH model with time-dependent "covariables" and constant "regression parameters". So methods of estimation for the usual PH model can be used. Note that the introduced "covariables" have time-dependent components even in the case when the covariable x is constant over time.

Alternative method is to take $\beta_i(t)$ as piecewise constant functions with jumps as unknown parameters. In such a case the PH model is used locally and the ratios of the default rates under constant covariates are constant on each of several time intervals.

5.2 AFT model with time dependent regression coefficients

Similarly as in the case of the PH model flexible models can be obtained by supposing that the regression coefficients β in the AFT model (7) are time-dependent, i.e. taking

$$S_{x(\cdot)}(t) = S_0 \left\{ \int_0^t e^{-\beta^T(u)x(u)} du \right\}, \quad t \geq 0, \tag{24}$$

where

$$\beta^T(t)\, x(t) = \sum_{i=1}^m \beta_i(t) x_i(t).$$

As in the case of the PH model with time-dependent coefficients, the model (24) with $\beta_i(t) = \beta_i + \gamma_i g_i(t)$ can be written in the form of the usual AFT model

$$S_{x(\cdot)} = G \left\{ \int_0^t e^{-\theta^T z(u)} du \right\}. \tag{25}$$

where θ and z are defined by (22).

Alternative method is to take $\beta_i(t)$ as piecewise constant functions with jumps as unknown parameters. It is evident that now we have many interesting possibilities to use different models with time depending stresses to studies the microstructure of financial markets.

6 Additive hazards model and its generalizations

An alternative of the PH model is the *additive defaults* or *hazards* (AH) model:

$$\lambda_{x(\cdot)}(t) = \lambda_0(t) + \beta^T x(t), \tag{26}$$

where β is the vector of regressor parameters. If the AH model holds then the difference of default rates under constant covariates does not depend on t. As the PH model this model has the absence of memory property: the default rate at the moment t does not depend on on the values of the covariate before the moment t.

Usually the AH model is used in the semiparametric form: the parameters β and the baseline default rate λ_0 are supposed to be unknown.

Both the PH and AH models are included in the *additive-multiplicative hazards* (AMH) model (Lin and Ying (1996)) :

$$\lambda_{x(\cdot)}(t) = e^{\beta^T x(t)} \lambda_0(t) + \gamma^T x(t). \tag{27}$$

Even this model has the absence of memory propriety so rather restrictive.

A modification of the AH model for constant covariates is the *Aalen's additive risk* (AAR) model (Aalen (1980)): the default rate under the covariate x is modeled by a linear combination of several baseline rates with covariate components as coefficients:

$$\lambda_x(t) = x^T \alpha(t). \tag{28}$$

where $\alpha(t) = (\lambda_1(t), \cdots, \lambda_m(t))^T$ is an unknown vector function.

Both AH and AAR models are included in the *partly parametric additive risk* (PPAR) model (McKeague and Sasieni (1994)):

$$\lambda_x(t) = x_1^T \alpha(t) + \beta^T x_2, \tag{29}$$

where x_1 and x_2 are q and p dimensional components of the explanatory variable x, $\alpha(t) = (\lambda_1(t), \cdots, \lambda_q(t))^T$, $\beta = (\beta_1, \cdots, \beta_p)^T$ are unknown.

Analogously as in the case of the PH model the AH model can be generalized by the *generalized additive hazards* (GAH) model:

$$\lambda_{x(\cdot)}(t) = q\{\Lambda_{x(\cdot)}(t)\}(\lambda_0(t) + \beta^T x(t)), \tag{30}$$

where the function q is parametrized as in the case of GPH models.

Both the GPH and the GAH models can be included into the *generalized additive-multiplicative hazards* (GAMH) model (Bagdonavicius and Nikulin (1997)):

$$\lambda_{x(\cdot)}(t) = q\{\Lambda_{x(\cdot)}(t)\} \left(e^{\beta^T x(t)} \lambda_0(t) + \delta^T x(t) \right). \tag{31}$$

In both GAH and GAMH models the function q is parametrized as in the GPH models: $q(u) = (1+u)^{-\gamma+1}$, $(1+\gamma u)^{-1}$, $e^{-\gamma u}$, and the GAH1, GAH2, GAH3 or GAMH1, GAMH2, GAMH3 models are obtained.

7 Remarks on parametric and semiparametric estimation

The literature on parametric and non-parametric estimation for the above considered models is enormous. Methods of estimation depend on experiment plans, censoring, covariate types, etc. We do not give here all these methods but give two general methods of estimation (one for parametric and other for semiparametric case) which work well for all models.

If the models are considered as parametric then the maximum likelihood estimation procedure gives the best estimators.

Let us consider for simplicity right censored survival regression data which is typical in survival analysis (more complicated censoring or truncating schemes are considered similarly):

$$(X_1, \delta_1, x_1(\cdot)), \cdots, (X_n, \delta_n, x_n(\cdot))),$$

where

$$X_i = T_i \wedge C_i, \quad \delta_i = \mathbf{1}_{\{T_i \leq C_i\}} \quad (i = 1, \cdots, n),$$

T_i and C_i and are the failure and censoring times, $x_i(\cdot)$-the covariate corresponding to the ith object, $T_i \wedge C_i = min(T_i, C_i)$, $\mathbf{1}_A$ is the indicator of the event A.

Equivalently, right censored data can be presented in the form

$$(N_1(t), Y_1(t), x_1(t), t \geq 0), \cdots, (N_n(t), Y_n(t), x_n(\cdot), t \geq 0),$$

where

$$N_i(t) = \mathbf{1}_{\{X_i \le t, \delta_i = 1\}}, \quad Y_i(t) = \mathbf{1}_{\{X_i \ge t\}}.$$

In this case for any t, $t > 0$

$$N(t) = \sum_{i=1}^{n} N_i(t) \quad \text{and} \quad Y(t) = \sum_{i=1}^{n} Y_i(t)$$

are the number of observed failures of all objects in the interval $[0, t]$ and the number of objects at risk just prior the moment t respectively.

Suppose that survival distributions of all n objects given $x_i(\cdot)$ are absolutely continuous with the survival functions $S_i(t, \theta)$ and the default rates $\lambda_i(t, \theta)$, specified by a common possibly multidimensional parameter $\theta \in \Theta \subset \mathbf{R}^s$.

Denote by G_i the survival function of the censoring time C_i. We suppose that the function G_i and the distributions of $x_i(\cdot)$ (if they are random) do not depend on θ.

Suppose that the multiplicative intensities model is verified: the compensators of the counting processes N_i with respect to the history of the observed processes are $\int Y_i \lambda_i du$. The likelihood function for θ estimation is:

$$L(\theta) = \prod_{i=1}^{n} \lambda_i^{\delta_i}(X_i, \theta)\, S_i(X_i, \theta)$$

$$= \prod_{i=1}^{n} \left(\int_0^\infty \lambda_i(u, \theta)\, dN_i(u) \right)^{\delta_i} \exp\left\{ -\int_0^\infty Y_i(u) \lambda_i(u, \theta)\, du \right\}$$

The maximum likelihood (ML) estimator $\hat{\theta}$ of the parameter θ maximizes the likelihood function. It verifies the equation:

$$U(\hat{\theta}) = 0,$$

where U is the score function:

$$U(\theta) = \frac{\partial}{\partial \theta} \ln L(\theta) = \sum_{i=1}^{n} \int_0^\infty \frac{\partial}{\partial \theta} \log \lambda_i(u, \theta) \{ dN_i(u) - Y_i(u) \lambda_i(u, \theta) du. \qquad (32)$$

The form of the default rates λ_i for the PH, AFT, GPH1, GPH2, GPH3, CE, CHSS, AH,AMH, AAR, PPAR, GAH, GAMH are given by the formulas (2),(7),(13),(14), (15), (16),(20),(26),(27),(28), (29), (30), (31). The parameter θ contains the regression parameter β, the complementary parameter γ (for some models) and the parameters of the baseline rate function λ_0, which is taken from some parametric family.

Let us consider a general approach (Bagdonavičius and Nikulin (2002)) for semiparametric estimation in all given models when the baseline default function λ_0 is supposed to be unknown. The martingale property of the difference

$$N_i(t) - \int_0^t Y_i(u) \lambda_i(u, \theta) du \qquad (33)$$

implies an "estimator" (which depends on θ) of the baseline cumulative hazard Λ_0. Indeed, all the above considered models can be classified into three groups in dependence on the form of $\lambda_i(t,\theta)dt$. It is of the form

$$g(x_i(s),\Lambda_0(s),0\le s\le t,\theta)d\Lambda_0(t)$$

(for PH, GPH, CE models), and $d\Lambda_0(f_i(t,\theta))$ (for AFT, CHSS models) or

$$g_1(x_i(s),\Lambda_0(s),0\le s\le t,\theta)d\Lambda_0(t)+g_2(x_i(s),\Lambda_0(s),0\le s\le t,\theta)dt$$

(for AH, AMH, AR, PPAR, GAH, GAMH models), Λ_0 possibly multi-dimensional for the AR and PPAR models). We remind that the estimation for the PH and AFT models with time-dependent regression coefficients and time-dependent or independent covariates is analogous to the estimation for the PH and AFT models with constant regression coefficients and properly chosen time-dependent "covariates".

For the first group the martingale property of the difference (33) implies the recurrently defined "estimator":

$$\tilde\Lambda_0(t,\theta)=\int_0^t\frac{dN(u)}{\sum_{j=1}^n Y_j(u)g(x_j(v),\tilde\Lambda_0(v,\theta),0\le v<u,\theta)}.$$

For the second group

$$\tilde A_0(t,\theta)=\sum_{i=1}^n\int_0^t\frac{dN_i(h_i(u,\theta))}{\sum_{l=1}^n Y_l(h_l(u,\theta))},$$

where $h_i(u,\theta)$ is the function inverse to $f_i(u,\theta)$ with respect to the first argument.

For the third group (AH, AMH, GAH, GAMH models)

$$\tilde\Lambda_0(t,\theta)=\int_0^t\frac{dN(u)-\sum_{i=1}^n g_2(x_i(v),\Lambda_0(v),0\le v<u,\theta)du}{\sum_{j=1}^n Y_j(u)g_1(x_j(v),\Lambda_0(v),0\le v<u,\theta)}.$$

A little more complicated situation is with AR and PPAR models. The "estimator" $\tilde\Lambda_0$ is obtained in the following way (McKeague and Sasieni (1994)): let us consider a submodel

$$\lambda_0(t)=\alpha(t)+\eta\varphi(t),$$

in which η is a one-dimensional parameter and φ,α are m-vector of functions.

The score function obtained from the parametric likelihood function for the parameter η (AR model) is

$$U(\eta)=\sum_{i=1}^n\int_0^\infty\frac{\varphi^T(t)x^{(i)}(t)}{\lambda_i(t)}(dN_i(t)-Y_i(t)(x^{(i)}(t))^Td\Lambda_0(t)),$$

and the score functions for the parameters η and β (PPAR model) are:

$$U_1(\eta,\beta)=\sum_{i=1}^n\int_0^\infty\frac{\varphi^T(t)x_1^{(i)}}{\lambda_i(t)}(dN_i(t)-Y_i(t)(x_1^{(i)})^Td\Lambda_0(t)-\beta^Tx_2Y_i(t)dt)=0,$$

$$U_2(\eta,\beta)=\sum_{i=1}^{n}\int_0^\infty \frac{x_2^{(i)}}{\lambda_i(t)}(dN_i(t)-Y_i(t)(x_1^{(i)})^T d\Lambda_0(t)-\beta^T x_2^{(i)}Y_i(t)dt)=0. \qquad (34)$$

If Λ_0 is unknown and we want to estimate it, the estimator should be the same for all φ. Setting $U(\eta)=0$ (AR model) or $U_1(\eta,\beta)=0$ (PPAR model) for all functions φ implies that for all t

$$\frac{x^{(i)}(t)}{\lambda_i(t)}(dN_i(t)-Y_i(t)(x^{(i)}(t))^T d\Lambda_0(t))=0,$$

or

$$\frac{x^{(i)}}{\lambda_i(t)}(dN_i(t)-Y_i(t)(x_1^{(i)})^T d\Lambda_0(t)-\beta^T x_2^{(i)}Y_i(t)dt)=0,$$

which implies the "estimators" (AR model):

$$\tilde\Lambda_0(t)=\sum_{j=1}^{n}\int_0^t\left(\sum_{i=1}^{n}x^{(i)}(u)(x^{(i)}(u))^T Y_i(u)(\lambda_i(u))^{-1}\right)^{-1}x^{(j)}(u)\,(\lambda_j(u))^{-1}\,dN_j(u)$$

or (PPAR model)

$$\tilde A(t)=\sum_{j=1}^{n}\int_0^t\left(\sum_{i=1}^{n}x_1^{(i)}(x_1^{(i)})^T Y_i(u)(\lambda_i(u))^{-1}\right)^{-1}x_1^{(j)}\,(\lambda_j(u))^{-1}\,(dN_j(u)-\beta^T x_2^{(j)}Y_j(u)du).$$

Note that for PH, GPH1, GPH2, GPH3 models

$$g(x(s),\Lambda_0(s),0\le s\le t,\theta)=e^{\beta^T x(t)},\quad e^{\beta^T x(t)}(1+\gamma\int_0^t e^{\beta^T x(u)}d\Lambda_0(u))^{\frac{1}{\gamma}-1},$$

$$e^{\beta^T x(t)}(1+2\gamma\int_0^t e^{\beta^T x(u)}d\Lambda_0(u))^{-\frac{1}{2}},\quad e^{\beta^T x(t)}(1+\gamma\int_0^t e^{\beta^T x(u)}d\Lambda_0(u))^{-1},$$

respectively. For the CE model

$$g(x(s),\Lambda_0(s),0\le s\le t,\theta)=e^{\beta^T x(t)}\{1+\Lambda_{x(\cdot)}(t)\}^{1-e^{\gamma^T x(t)}},$$

where the function $\Lambda_{x(\cdot)}$ is defined by the equation

$$\int_0^t e^{\beta^T x(u)}\{1+\Lambda_{x(\cdot)}(u)\}^{1-e^{\gamma^T x(u)}}d\Lambda_0(u)=\Lambda_{x(\cdot)}(t).$$

If x is constant in time then for the CE model

$$g(x,\Lambda_0(s),0\le s\le t,\theta)=e^{\beta^T x}\{1+e^{(\beta+\gamma)^T x}\Lambda_0(t)\}^{e^{-\gamma^T x}-1}.$$

For the AFT and CHSS models

$$f_i(t,\theta)=\int_0^t e^{-\beta^T x(u)}du,\quad \int_0^t e^{-\beta^T x(u)}u^{e^{\gamma^T x(u)}-1}du.$$

For the AH, AMH, AR, PPAR, GAH and GAMH models

$$g_1(x_i(s), \Lambda_0(s), 0 \le s \le t, \theta) = 1, \quad e^{\beta^T x(t)}, \quad x^T, \quad x_1^T$$

and

$$g_2(x_i(s), \Lambda_0(s), 0 \le s \le t, \theta) = \beta^T x(t), \quad \beta^T x(t), \quad 0, \quad \beta_2^T x(t),$$

respectively. For the GAMH1 model (formulas are analogous for the GAMH2, GAMH3, GAH1, GAH2, GAH3 models):

$$g_1(x_i(s), \Lambda_0(s), 0 \le s \le t, \theta) = e^{\beta^T x(t)}\, g(x_i(s), \Lambda_0(s), 0 \le s \le t, \theta),$$
$$g_2(x_i(s), \Lambda_0(s), 0 \le s \le t, \theta) = \delta^T x(t)\, g(x_i(s), \Lambda_0(s), 0 \le s \le t, \theta),$$

where

$$g(x_i(s), \Lambda_0(s), 0 \le s \le t, \theta) = \left(1 + \gamma(\int_0^t e^{\beta^T x(u)} d\Lambda_0(u) + \delta^T \int_0^t x(u) du)\right)^{\frac{1}{\gamma}-1}.$$

For the PH, GPH and CE models the weight $\frac{\partial}{\partial\theta} \log \lambda_i(u, \theta)$ in (32) is a function of $x_i(\cdot)(v), \Lambda_0(v), 0 \le v \le u$ and θ. So the modified score function is obtained replacing Λ_0 by its consistent estimator $\tilde{\Lambda}_0$ in the parametric score function (32).

In the case of the AFT, CHSS, AH, AMH, AR and PPAR models the weight depends not only on Λ_0 but also on λ_0 and (or) λ_0'. But the more important thing is that $\lambda_i(u)du$ do not depend on λ_0 and λ_0'. So construction of the modified likelihood function can be done by two ways. The first way is to replace Λ_0 by $\tilde{\Lambda}_0$ and λ_0 and λ_0' by nonparametric kernel estimators which are easily obtained from the estimator $\tilde{\Lambda}_0$. The second, much more easy way is to replace λ by 1, λ' by 0 and Λ_0 by $\tilde{\Lambda}_0$ in the score function (32) (or (34) for the PPAR model, in the case of the AR model there are no parameters left to estimate). The efficiency loses very slightly in this case of such simplified weight.

Computing of the modified likelihood estimators is simple for the PH, GPH and CE models. It is due to the remarkable fact that these estimators can be obtained by another way: write the partial likelihood function

$$L_P(\theta) = \prod_{i=1}^n \left[\int_0^\infty \frac{g\{x_i(v), \Lambda_0(v), 0 \le v \le u, \theta\}}{\sum_{j=1}^n Y_j(u) g\{x_j(v), \Lambda_0(v), 0 \le v \le u, \theta\}} dN_i(u)\right]^{\delta_i}, \qquad (35)$$

and suppose at first that Λ_0 is known. Replacing Λ_0 in the score function by $\tilde{\Lambda}_0$ exactly the same modified score function is obtained as going from the full likelihood! So computing the estimator $\hat{\theta}$ the score equation is not needed. Better maximize the modified partial likelihood function which is obtained from the partial likelihood function (35) replacing Λ_0 by $\tilde{\Lambda}_0$. The general quasi-Newton optimization algorithm (given in Splus) works very well seeking the value of θ which maximizes this modified function (Bagdonavičcius, Hafdi, Himdi and Nikulin (2002)).

The most complicated case is the case of AFT and CHSS models: the modified score functions are not differentiable and even continuous. So the modified maximum likelihood estimators are the values of θ which minimize the distance of the modified score function from zero. Computational methods for such estimators are given in Lin and Geyer (1992).

8 References

1. Aalen, O. (1980) A model for nonparametric regression analysis of counting processes. In. *Mathematical Statistics and Probability Theory*, Lecture Notes in Statistics, **2**, (Eds. W. Klonecki, A. Kozek and J. Rosinski), New York: Springer Verlag, 1-25.

2. Andersen, P.K. (1991). Survival analysis 1981-1991: The second decade of the proportional hazards regression model. *Statistics in Medicine*, **10**, # 12, 1931-1941.

3. Andersen, P.K., Borgan, O., Gill, R.D.& Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer.

4. Bagdonavičius, V. (1978) Testing the hyphothesis of the additive accumulation of damages. *Probab. Theory and its Appl.*, **23**, No. 2, 403-408.

5. Bagdonavičius, V. and Nikulin, M. (1994). Stochastic models of accelerated life , Advanced Topics in Stochastic Modelling (ed. J.Gutienez, M.Valderrama), *World Scient.*, Singapore.

6. Bagdonavičius, V. and Nikulin, M. (1999). Generalized Proportional Hazards Model Based on Modified Partial Likelihood, *Lifetime Data Analysis*, **5**, 329-350.

7. Bagdonavičius V., M.Hafdi and Nikulin M. (2002).The Generalized Proportional Hazards Model and its Application for Statistical Analysis of the Hsieh Model." In : Proceedings of " The Second Euro-Japanese Workshop on Stochastic Risk Modelling for Finance, Insurance, Production and Reliability," September 18-20, Chamonix, France, (Eds. T.Dohi, N.Limnios, S.Osaki), p. 42-53.

8. Bagdonavičius V., Hafdi, M., El Himdi, K. and Nikulin M. *Analyse du modèle des hazards proportionnels généralisé. Application sur les donnés du cancer des poumons.* Preprint 0201, I.F.R. "Santé Publique", (2002).

9. Bagdonavičius V., Hafdi, M., El Himdi, K. and Nikulin M. *Analysis of Survival Data with Cross-Effects of Survival Functions. Applications for Chemo and Radiotherapy Data.* Preprint 0202, I.F.R. "Santé Publique", (2002).

10. Bagdonavičius V. and Nikulin M. (2002). *Accelerated Life Models: Modeling and Statistical Analysis.* Boca Raton: Chapman and Hall/CRC.

11. Bagdonavičius V., Kruops J., and Nikulin M. (2011). *Non-parametric tests for censored data.* London: ISTE & J.WILEY.

12. Bedford,T. and Cooke,R. (2001). Probability Risk Analysis. Fondation and Methods. Cambridge: Cambridge University Press.

13. Ceci Z., Mazliak L. (2004). Optimal design in nonparametric life testing. *Statistical Inference for Stochastical Processes*, **7**, 305-325.

14. Couallier,V., Gerville-Reache,L., Huber, C., Mesbah, M. (2013), *Statistical Models and Methods for Reliability and Survival Analysis*, ISTE/WILEY:London.

15. Cox, D.R. (1972). Regression models and life tables, *J.R.Statist.Soc.*, B, **34**, 187-220.

16. Cox, D.R. (1975) Partial likelihood. *Biometrika*, **62**, 269-276.

17. Cox, D.R., and Oakes, D. (1984). Analysis of Suvival Data, *Methuen (Chapman and Hall)*, New York.

18. Dabrowska, D.M., Doksum, K.A. (1988). Partial likelihood in Transformations Models with Censored Data, *Scand. J. Statist.* **15**, 1-23.

19. Duffie,D. and Singleton,K.J. (2003). *Credit Risk. Pricing, Measurement, and Management.* Prinseton University Press: Prinseton and Oxford.

20. Gouriéroux, Ch. and Jasiak, J. (2005). Duration. In: A Companion to Theoretical Econometrics, (Ed. Badji H. Baltagi), Beijing: Peking University Press

21. Hougaard, P. (1986) Survival models for heterogeneous populations derived from stable distributions, *Biometrika*, **73**, 3, 387-396.

22. Hsieh, F. (2001). On heteroscedastic hazards regression models: theory and application. *Journal of the Royal Statistical Society,* Series B **63**, 63-79.

23. Huber, C., Limnios,N., Mesbah, M., Nikulin, M.S. (Eds.) (2007). *Mathematical Methods in Survival Analysis, Reliability and Quality of Life*, ISTE/Wiley: London.

24. Kiefer, N. (1988). Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, Vol. 26, No.2. pp. 646-679.

25. Kleinbaum, D, (1996). *Survival Analysis: A Self-Learning text.* New York: Springer-Verlag.

26. Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis*, New York: Springer.

27. Lancaster, T. (1979). Econometric Methods for the the Duration of Unemployment. *Econometrica*, 47(4), 939-56.

28. Lawless, J.F. (1982). Statistical Models and Methods for Lifetime Data, *Wiley*, New York.

29. Lin, D.Y., Geyer, C.J. (1992) Computational methods for semiparametric linear regression with censored data. *Journal Comput. and Graph. Statist.,*, **1**, 77-90.

30. Lin, D.Y. and Ying, Z. (1996) Semiparametric analysis of the general additive-multiplicative hazard models for counting processes. *The Annals of Statistics*, **23**, 5, 1712-1734.

31. Mann, N.R., Schafer, R.E. and Singpurwalla, N. (1974) *Methods for Statistical Analysis of Reliability and Life Data.* New York: John Wiley and Sons.

32. McKeague, I.W., Sasieni, P.D.(1994) A partly parametric additive risk model. *Biometrika*, **81**,#3, 501-514.

33. Nikulin,M., Limnios,N., Balakrishnan, N., Huber, C., Kahle W. (Eds.), (2009). *Advances in Degradation Modeling: Applications to Reliability, Survival Analysis and Finance.* Birkhauser: Boston.

34. Royston, P., Palmar, M. (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data with application to prognostic modelling and estimation of treatement effects. *STATASTICS in MEDICINE*, **21**, 2175-2179.

35. Royston, P., Lambert, P. (2011) *Flexible Parametric Survival Analysis Using Strata: Beyond the Cox Model.* Stata Press Publications, StataCorp LP, College Station, Texas, 347p.

36. Sedyakin, N.M. (1966) On one physical principle in reliability theory. *Techn. Cybernetics*, **3**, 80-87.

37. Stablein, D. M., Koutrouvelis, I. A. (1985). A two sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics* **41**, 643-652.

38. Voinov,V., Nikulin,M.S., Balakrishnan, N. (2013).*Chi-Squared Goodness of fit Tests with Applications.* Academic Press/Elsevier: Amsterdam.

39. Zeng, D., Lin, D.Y.(2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J.R.Statist. Soc.*, **B 69**, 509-564.

40. Zhou, C., Chinnam, R.B., Korostelev, A. (2012) Hazard rate models for early detection of reliability problems using information from warranty databases and upstream supply chain. *International Journal of Production Economics*, **139**,180-195.

Estimating of the Russian Banks Bankruptcy Probability Using Improved Classification Algorithms

V.S. Timofeev and A.A.Sanina
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: v.timofeev@corp.nstu.ru, anastas.sanina@gmail.com

Abstract

In this article we consider the problem of predicting Russian Bank bankruptcy prior to the actual occurrence of this event. A mathematical model based on the officially published data characterizing the banks activity is developed and it is used to estimate the probability of a Russian Bank bankruptcy. In Conclusion, the main results concerning the possible further application of the developed method are formulated for predicting bank bankruptcy in practice.

Keywords: Bank, bank bankruptcy, Discriminant Function Analysis, Logit model, Probit Model, likelihood function, classification problem, factors, two-valued dependent variable.

Introduction

Banking system has an important function in the Nation's Economy. Consequently it requires very careful attention. A sudden bank failure leads to loss of confidence in the entire banking system and it causes the reduction in private savings and the inefficient allocation of funds that in turn does not contribute to the strengthening of the economy. So, it is natural that we want to have some more or less universal technics in order to estimate banking institutions and predict a bank failure. This could be an additional tool for the Central bank to revoke the license as well as a good useful instrument for an independent rating agency's work.

The official statistic data concerned the banking institutions is usually reported on public websites and at first glance it is only official dry information that does not carry much meaning and especially it does not contain any information about a possible bank bankruptcy. However, it is not so if we consider this information in detail and analyse it.

Currently, binary choice models are used to solve such problems. They are Logit and Probit Models and the Discriminant Function Model which is only a particular case of the previous models [1, 2, 3, 4, 5, 7]. It is logical that at some point a natural question will arise which of the models are preferred for solving this problem. We chose two main criterion to select a model. They are the "unpretentiousness" of a model to the input data and the classification quality. The Discriminant Function Model is obviously loser in this case [2, 3, 4]. According to the model it is supposed that data satisfies the main assumptions of the Discriminant Function Analysis such as continuity and independence for input factors and their normal distribution. It should be noted that these assumptions make the model non-working for real conditions [8, 9]. Logit and Probit Models apply less restrictions on input data and therefore they are more flexible [8, 9]. In addition, it is known that the considered models are based on the Logistic and Normal distribution respectively. So, it is reasonably to think about possibility of constructing a model based on any other distribution. The new model was tested before and was compared with the existing models in terms of classification quality on the model data [10]. In this work we study the classification quality of the new model and compare it with results obtained by using already known models to solve the problem of predicting the Russian banks bankruptcy probability.

1 Problem Definition and Metodology

We denote by y a bankruptcy indicator for a bank to solve this problem. The output dependent variable y takes one of two possible values. They can be 1 if a bank is bankrupt or 0 when a bank is not bankrupt and financially stable organization with a good reputation on the financial services market respectively. The Russian banks were selected for statistical analysis and part of them was declared bankrupt in 2013. Two sets of data characterizing the work of financial institutions in 2011 and 2012 were considered. The total number of banks was 37 ($m = 37$) and 10 of them were declared bankrupt (this was approximately 27% of the total volume of the sample).

The following indicators based on the officially published statistics on the banking institutions were calculated [6]:

- AU_1 is the ratio of the cash amount on the balance-sheet of the credit institution to the total amount of assets;
- AU_2 is the ratio of equity capital to the total amount of liabilities of the credit institution;
- AU_3 is the ratio of the total amount of own funds revaluation reserves/fixed assets (and intangible assets, tangible reserves and similar values which are written as a single line of the assets bank balance) / to the equity capital (aggregated sum of own funds) of the credit institution;
- AU_4 is the ratio of the current year retained earnings total sum (at liability side of the balance-sheet) to the aggregated value of bank liabilities;
- AU_5 is the ratio of the previous years retained earnings total sum (except for the retained earnings of the current year, which is written as a separate line) to the total value of the bank assets;
- AM_2 is the ratio of the credit institution own funds revaluation reserves to the total value of liabilities;
- AM_3 is the ratio of the credit institution equity capital (the sum of own funds) to the total value of bank assets;

- AM_4 is the ratio of the current year and previous years retained earnings sum of liability side of the balance-sheet to the credit institution total value of liabilities.

We will consider a vector of variables
$x_i = (AU_{i1}, AU_{i2}, AU_{i3}, AU_{i4}, AU_{i5}, AM_{i2}, AM_{i3}, AM_{i4})$ as an input factors vector for each i-th bank. We denote the variables AU_1-AM_4 for $X_1 - X_8$ for the convenience of data processing and obtain the vector $x_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6}, X_{i7}, X_{i8})$ where $x_{ij} \in R$ is a value of the j-th factor for the i-th observation, $i = \overline{1,\, 37}$, $j = \overline{1,\, 8}$.

It is easy to see when we can establish a relation between the occurrence or non-occurrence of a bank bankruptcy and the main factors describing the banks activity it will be possible to forecast these events. We build a model to estimate a probability of bank bankruptcy $(y = 1)$. Since what y is a binary variable it is logical to use the Logistic Model. The basic equation of the model is

$$P\left\{y_i = 1 \,|\, x_i\right\} = F\left(z_i\right),$$

where $F\left(z\right)$ is a cumulative distribution function for the Standard logistic distribution describing the probability of the specified event from values of input factors, z_i is defined as a linear combination of the input factors

$$z_i = \theta x_i^T = \theta_1 x_{i1} + \ldots \theta_n x_{in},$$

where $\theta = (\theta_1,\, \theta_2, \ldots,\, \theta_n)$ are unknown coefficients. The $\theta_1,\, \theta_2, \ldots,\, \theta_n$ parameters are fitted based on the independent variable values and the corresponding values of the dependent variable y. The maximum likelihood method is usually used to estimate θ parameters so, that they maximize the value of the likelihood function. However, it is common to use an equivalent logarithmic expression for calculating the likelihood function:

$$\ln L\left(\theta\right) = \sum_{i=1}^{m} y_i \ln F\left(\theta x_i^T\right) + \left(1 - y_i\right) \ln\left(1 - F\left(\theta x_i^T\right)\right).$$

It is easy to see that theoretically it can be taken any distribution function as $F\left(z\right)$ which is not equal to 0 or 1 on the entire argument domain. The Logistic or Normal distribution are traditionally chosen as $F\left(z\right)$ for the Logit and Probit models, respectively. In this this paper the Laplace distribution is proposed as an alternative distribution function [11]:

$$F\left(z\right) = \begin{cases} \frac{1}{2}\exp^{\alpha(z-\beta)}, & x \leq \beta \\ 1 - \frac{1}{2}\exp^{-\alpha(z-\beta)}, & x > \beta \end{cases},$$

where α and β are unknown parameters $(\alpha > 0,\, -\infty < \beta < \infty)$.

Let $Err\left(\arg\max_{\theta}\left(\ln L\left(\theta,\, \alpha,\, \beta\right)\right)\right)$ is the magnitude of the classification error (that is the part of incorrectly classified observations) obtained by any model. Since the distribution function depends on the parameters they can fitted by a special way in order to minimize this error:

$$\left(\hat{\alpha},\,\hat{\beta}\right)=\underset{(\alpha,\,\beta)}{\arg\min}\left(Err\left(\hat{\theta},\,\alpha,\,\beta\right)\right) \tag{1}$$

Since we agreed that the probability of bankruptcy or not bankruptcy for a selected bank will be described by the function of the Laplace family distribution we try to build a model with the optimal coefficients and the law parameters values (1).

It should be noted that there are some conditions when the model described above does not work at all due to the fact that the value of $F(z)$ function argument can take on a big value or vice versa with the certain factors and coefficients values. The distribution function takes its on extreme values which "break" the likelihood function. In this case, it is a good practice to perform a preliminary normalization of the input factors. The classification accuracy of the proposed method was already discussed earlier [10]. Next, we consider the the accuracy of the method using the previous data.

2 Experimental Results

Table 1 shows the values of the *Err* indicator when solving the classification problem with estimated values of the unknown coefficients and the Laplace distribution parameters values for both data sets. There is a nomenclature in the next tables where Logit means that the model is built based on the Logistic distribution, Probit means that the model is built based on the Normal distribution, Laplace1 means that the model is built based on the Laplace distribution with fixed parameter values ($\alpha = 1,\ \beta = 0$), Laplace2 is the (1)-st task solution.

Table 1: The *Err* values for the model included all input factors

Year	Logit	Probit	Laplace1	Laplace2	Laplace1 / Laplace2
2011	0.135	0.135	0.054	0.027	2
2012	0.162	0.108	0.027	0.027	1

Table 2: The *Err* values for a model with factor variables

Year	Logit	Probit	Laplace1	Laplace2	Laplace1 / Laplace2
2011	0.054	0.162	0.054	0.027	2
2012	0.083	0.083	0.055	0.055	1

Since the obtained results are very unstable the model should be simplified, that is to analyze the data and reduce the number of explanatory variables. The variable inclusion and exclusion methods show that only AU_1 and AU_3 variables are significant when building a linear model. However, if we consider the pair correlation matrix in detail then it is clear to see that there are not any compelling reasons for exception the rest variables. The factor analysis was performed for this reason and three new variables were selected. Next, we perform the classification procedure once again and compare its classification quality with the results obtained above. The 2-nd, 3-rd and the 4-th tables show the classification quality, estimated coefficients values and the Laplace law parameters for solving the Laplace2 problem.

Table 2 shows that the model based on the Laplace distribution is constantly better in terms of classification quality at all the test sets. An additional parameters fitting procedure for the Laplace distribution family improves the obtained good result by up to 2 times. It amounts to only 1 incorrectly classified case compared with 3 cases for another models. The new obtained result is significantly more stable than the result for the original model and does not almost depend on initial estimates for unknown coefficients and law parameter values.

Table 3: The coefficient values for a model with factor variables (data of 2011)

2011	Logit	Probit	Laplace1	Laplace2$*$
const	-1.140	-1.327	-0.353	-0.345
FAC_1	6.070	2.069	5.538	5.348
FAC_2	7.043	2.358	6.440	6.407
FAC_3	-1.387	-1.750	-0.262	-0.222

$*\alpha = -0.031,\ \beta = 1.038$

Table 4: The coefficient values for a model with factor variables (data of 2012)

2012	Logit	Probit	Laplace1	Laplace2$*$
const	-1.145	-0.633	-0.921	-0.837
FAC_1	1.237	0.751	0.887	0.806
FAC_2	1.597	0.834	1.296	1.178
FAC_3	3.446	1.883	2.739	2.490

$*\alpha = 0,\ \beta = 1$

Conclusions

The investigation shows that the proposed method is effective and give good results when solving the classification problem not only on the model data [10] but also on the real data in practice. The new method works really well and improves the classification quality up to 2 times. It is a good result especially if we speak about such an important event which is to declare a bank bankrupt before it happens in actual fact.

Acknowledgements

This research has been supported by the Ministry of Education and Science of the Russian Federation as part of the state task no 2014/138 (project No 1689).

References

[1] Ajvazyan S.A., Buhshtaber V.M., Enukov I.S., Meshalkin L.D. (1989). *Prikladnaya statistica: Classificaciya I snizhenie razmernosti*. Finance and Statistics, Moscow.

[2] Alvin C. Rencher (2002). *Methods of Multivariate Analysis*. Brigham Young University, Provo.

[3] Forsyte J., Malkolm M., Mouler K. (1980). *Machine Methods of Mathematical Calculations*. Mir, Moscow.

[4] Judd Ch., McCleland G. (1989). *Data Analysis*. Harcourt Brace Jovanovich, USA.

[5] Kendall M. G., Stuart A. (1976). *Multidimensional Statistical Analysis and Time Series*. Nauka, Moscow.

[6] Kokova E.V. (2014). Metody otsenki veroyatnosti bankrotstva kommercheskogo banka *Vserossiiskaya studencheskaya olimpiada po statistike*. MESI, Moscow. pp. 62-68.

[7] Naresh K. Malhotra (2002). *Marketing Research: An Applied Orientation*. Williams, Moscow.

[8] Pohar M., Blas M., Turk S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study *Metodolovski zvezki journal*. Vol. **1**, pp. 143-161.

[9] James Press S., Wilson S. (1978). Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the America Statistical Assotiation*. Vol. **73**, pp. 699-705.

[10] Sanina A.A. (2015). Improving the quality classification using multiple choice linear models. *Sbornik Nauchnyh Trudov of Novosibirsk State Technical University*. Vol. **1**, pp. 23-32.

[11] Zolotukhin I.V. (2012). New Class of Multivariate Generalized Laplace distribution *Vestnik of Lobachevsky State University of Nizhnii Novgorod*. Vol. **68**, pp. 60-64.

Specific of Dairy Entrepreneurship in Period of Globalization

Marina U. Arkhipova[1] and Kirill V. Arkhipov[2]
[1] *National Research University – Higher school of economics, Moscow, Russia*
[2] *Moscow state university of economics statistics and informatics (MESI) Moscow, Russia*
e-mail: marhipova@hse.ru, KVArkhipov@mesi.ru

Abstract

In this paper we describe the tendencies on dairy market in Russia: transition from separate independent dairy plants supplying small regions, to national units and then to international enterprises. Globalization helps to implement world's best practices in R&D and gain competitive advantage with suppliers due to consolidation of purchasing volumes of raw and pack. Such global companies need at least several plants to cover national consumers demand in Russia. One of the most important topics for transnational dairy companies is defining optimal number of plants, places of their location and building efficient supply chain scheme. Optimization model was created using genetic methods of search. It showed the optimal network of production plants on the territory of the country. Results of the model showed that the optimal scenario differs from existing model and requires significant changes to be implemented. Model can be implemented to the whole national dairy industry branch as part of government policy, or by a private enterprise to optimize its logistic and manufacturing expanses.

Keywords: supply optimization, dairy products production, sourcing models.

Introduction

Globalization of world trade and production processes brings specific features to supply chain development. In Russia now most dairy plants that used to be state or independent entities are now parts of transnational corporations.It requires the new way of managing their production lines utilization and products portfolio. New approach that helps to optimize finish products manufacturing and transportation expenses is sourcing management. This part of supply chain helps company with several production plants to manage their production lines utilization and find optimal logistic routs from factory to clients (or hub warehouses). In short term that means daily production scheduling between several plants depending on demanded consumption in each territory, logistic costs from plants to customers, free capacity in plants, difference in production expenses between factories and materials availability. In long term sourcing management helps to estimate needs and payback from capital investments in constructing of new plants, installing of new production lines, their reallocation between different plants and demand in transport and warehouse capacity.

Sourcing management enables enterprises with several production sites optimize their supply chain operations, purchasing costs for raw & pack and production expenses. It is especially important in current economic situation, when from one hand competitiveness has increased on the dairy market and from the other hand the country is in crises period when production companies must pay the primary attention to all expenses and control the price of its products on the shelf.

According to current result of the research done by [11] logistic, warehousing and production expenses play a significant role in finished good price (15-25%) for dairy products in Russia, their optimization is very important for company competitiveness. Significant results in their optimization can be achieved first of all if the company clearly understands consumers' needs, can prepare reliable forecast of sales and then can convert it into production plan.

In the article bellow we studied how to convert the long term (3-5 years) forecast of sales into production plan. The analyses showed that current number and location of dairy plants in Russia is far from optimal. The model was created that helped to define the optimal number of dairy plants taking into account logistic and production costs.

The model is important for the country, because food supply including dairy production is one of the top priority in the state policy. This branch of economics is under counter-sanctions now, meaning restriction for import from European Union. Now it causes growths in production volumes, but at the same time makes it very important to have a special long term state policy of dairy branch support.

1 Dairy products classification

Dairy products are very important part of human nutrition and in Russia is very popular. 68% of population in the country [15, 17] consume at least 3 dairy products a day. According to current research [18] key indicators for dairy products in Russia that consumers take into account are price and freshness. For milk production companies that means primary importance to build optimal and efficient supply chain, determine number, types and places of location for their plants.

We start our research with dairy products classification, depending on the features important for supply chain and operations planning. Such as their shelf life, consistence and type of packaging. Dairy companies determine several product categories with different approach for their production and sourcing management:

- products with long shelf life (more than 2 months): UHT milk and milk cocktails, condensed milk, skimmed milk powder, etc.;
- products with medium shelf life (4-8 weeks): butter, spoonable yogurts, drinkable yogurts, cheese, whey drinks, etc.;
- products with short shelf life (3 weeks and less) baby dairy, pasteurized milk, etc..

Study of consumers behavior and actual trends on dairy market shows that clients become more interested in local products, they associate local traditional products with naturality and freshness, while traditional products that comes from other regions seem to consumers unnatural and not healthy. According to national survey made in 2013, 59% of consumers of dairy products would like to know not only the production plant, but also farm and even cow that provided milk for that exact SKU.

This brings new challenges for dairy companies in Russia. On the one hand, there are many milk plants left from soviet period, but on the other hand, the expanses are too high for their modernization and regular maintenance.

2 Specific features of dairy production in Russia

Boisterous growth in volumes of sales that Russia faced in 00th practically in every sphere of goods and services was caused mainly by extensive development and growth of distribution to the remote places of the country. In that period main attention of companies' management was concentrated on regional expansion and development (www.rbc.ru). In such conditions operational expenses were not in focus. By 2010 many companies, especially on FMSG market, reached the upper level of possible distribution and double digit growth changed to insignificant decrease or flat sales trend. In current conditions, transnational corporations are not interested in high investments in plants modernization, while in Russia nearly 100 dairy factories didn't have any reconstruction in the last 40 years.

Main specific of supply processes in Russia is poor development of transport infrastructure and low reliability of local suppliers. Average deviation in lead time when using rail way is more than 8 days and average speed of materials flow inside the country is 17,8 km per hour [19]. Another Russian specific is long distance between production sites and customers and low concentration of population in most of the regions. Huge distance in the country makes companies operating on milk market to have several production sites, because on average, the requirements from retail chains is to deliver finished milk products with at least 70% freshness and have 14 days shelf life at maximum. Spread of production volumes between several plants has a negative impact on materials turnover: 31 days in Russia, 12-16 days in USA and Europe (www.danone.com). Suppliers' service level that shows a percentage of orders fulfilled on time, in the exact quantity and perfect quality is also lower, than average in the developed countries (only 78% of orders in 2012 in Russia satisfy parameters mentioned above, while in US and Europe this KPI was close to 98%).

The main indicator of supply chain efficiency is customers' service level that shows the percentage of orders from the clients that were fulfilled at once. In values of this indicator Russia is close to developed countries level. But in Russia there is no frozen horizon for orders from clients that explains difference between average accuracy of forecast of sales (table 1).

Russia was one of the world's leader in growth of milk products consumption till 2012, nevertheless further industry development is limited by weak state support of farmers. Milk production companies now suffer from shortage of fresh milk from

Table 1: Indicators, characterizing materials supply of milk factories in 2012 (www.danone.com)

Indicators	Russia	Europe	USA
Average distance from clients to milk plant, km	868	229	276
Average distance from suppliers to milk plants, km	637	289	309
Average speed of materials flow across the country, km/hour	17,4	33,5	31,8
Materials turnover, days	31	12	16
Suppliers service level (SSL), % of perfect orders	78	96	94
Customers service level (CSL), %	94	97	97

farms and concentrate in profitability projects. One of such project is implementation of hub model for materials flow management. Great variety of materials which have different logistic parameters requires their multidimensional classification analyses to be executed firstly.

3 Multidimensional classification of dairy plants in Russia

As of 2014, in Russia 348 dairy plants was operating. Multidimensional analysis was done on the next step of the research to determine structure homogeneous groups of plants by their industrial parameters and volumes of production. The results of the analysis will help to determine leading plants and those who need serious reconstruction or closing. Following indicators were included:

- x1 - capacity utilization, %;
- x2 - number of years since last reconstruction, years;
- x3 - fished goods produced during last 12 months, tons;
- x4 - average weighted distance to milk suppliers, km;
- x5 - number of people living in 500km distance from plant, number.

Materials testing for anomalous observation using Grubbs criteria and criteria of Tietjen-Moore (Ayvazyan and others, 2001) helped to determine such observations, mainly for $x3$ indicator. Most observations which were detected as anomalous were closed for reconstruction (they were excluded from further analysis) or work seasonally. Such dairy plants formed a special cluster “Seasonal plants” (cluster four).

For normalized values of indicators multidimensioned cluster analysis were done, that helped to divideplants into three structurally homogenous clusters by values of selected indicators:

First cluster “modern plants” (14% of total dairy plants) included sites with high capacity utilization (74% average), 2-10 years since last modernization. Low distance to milk farms and high number of people living not farer than 500km from that plants show that these plants are efficient and optimal from sourcing point of view

Second cluster “medium utilized plants” (58% of total dairy plants) contains sites that have medium level of their capacity utilization, located close to consumers but far from the milk suppliers. For such plants due to economic changes it is reasonable to relocate some of their capacities closer to farms and organize milk processing factories that will supply people of that region with fresh milk with production farm reference. That reflects modern tendencies and consumers’ expectations. Dairy products that can be produced at farm: pasteurized milk, thermostatic sour cream and curds.

Third cluster “not efficient plants” (28% of total plants) was formed by sites with low capacity utilization, last reconstruction done more than 30 years ago. Their number is big, with more than 50 thousand employees, so optimization of their operations, their relocation or modernization should be part of direct state support program, otherwise in few years taking into account current tough economic conditions that plants are in high risk of closing. That will have a negative social and finance impact on nation economy.

To create such state support program and determine transition steps, it is very important to analyze and forecast consumers’ needs from one hand, and optimize dairy products national sourcing from the other hand.

4 Sourcing optimization model

In the third cluster of plants “not efficient plants” were plants that are rather close to consumers, but far from milk farms, with high depreciation of fixed assets. Without state support and special program that plants are likely to be closed, so it is important to estimate is it worth to modernize them, change their production portfolio, close them or relocate their production capacities to other plants.

To estimate sourcing model the following assumptions were done [9]:

- fives scenarios for plants of cluster #3 are considered: modernization for current production portfolio, modernization for new portfolio, capacity relocation to another plants, capacity relocation to new location (construct new plant in other location using production lines from plant of cluster #3) or plant closure;

- plants and production portfolio of plants from cluster #1, 2 & 4 are considered as constants;

- consumers demand, places of their location and transportation costs are considered as constants;

- One kind of finished products and fresh milk from farms are considered, no option to mix different milk or dairy products producers in one truck;

- horizon to calculate efficiency is 5 years;

- milk farms to plants of cluster 3 will be assigned automatically based on nearest farms with extra milk supply capacities;
- consumers are assigned automatically to plants based on unsatisfied demand left after supply from plants of cluster #1, 2, 4;
- As places for plants relocation are regarded big cities (with population over 300 thousand people) and existing milk farms.

Thus, to solve the problem, the following objective function should be estimated:

$$J = \sum_{t=1}^{5}\sum_{i=1}^{M}\sum_{j=1}^{N} m_{i,j} + \sum_{j=1}^{N}\sum_{k=1}^{L}\sum_{t=1}^{5} c_{j,k} + \sum_{j=1}^{N} R_j \to min \tag{1}$$

where t - number of years to analyze optimization efficiency; M - number of milk farms; N - number of plants from cluster #3; $m_{i,j}$ - cost of milk delivery from farm i to plant j; L - number of clients (consumers of dairy products); $c_{j,k}$ - cost of dairy products delivery from plant j to client k R_j - cost of plant reconstruction or closure.

The problem has a great variety of possible solutions (more, than 4.6 × 1034), so to find an efficient supply scheme a genetic algorithm was used. It allows to start a process of directed search of possible solution that will have the best value of the objective function. Following steps were done according to rules of genetic algorithms usage:

1. created a way for solution codding that uniquely determine the value of objective function;
2. created sets of generations of possible solutions (so called populations);
3. determined the rules of population evolution;
4. started continuous process of new generation;
5. created a rule of algorithm stoppage.

Using genetic algorithms to solve economic problems gain accelerating interest and become an object of scientific researches in Russia [7] and in other countries of the world. Often genetic algorithms are considered as a part of imitation models. For economic problems with great number of different possible solutions, genetic algorithm can provide several choices that are very close to each other in the value of objective function, but provide absolutely different results. This gives a possibility to decision-makers to choose the best option also taking into account another factors that were not included in the initial model. To start the evolution process of directed search of optimal supply scheme possible solutions were presented in the coded way (table 2).

Where N – number of plants from cluster#3, The whole matrix containing all materials for each plant is regarded as one solution. On the next step of research

Table 2: Principles of solution codding for genetic algorithms

Plant number	Type of change	New portfolio type (for change type #2) / Plant for capacity relocation (for change type #3) / Place of relocation (town number for change type #4)	Code of solution
1	1	0	1-1-0
2	5	0	2-5-0
3	2	3	3-2-3
...	...	...	...
N		16	

two populations of possible solutions were generated, each 1000. For each solution the value of objective function was calculated. Based on Russian roulette principle solutions were selected and theytook part in new generation creation. After 1000 iterations, solutions from population #1 and population #2 were merged that helped to further improve the value of the objective function. Dynamic of the average values for each generation is presented in figure 1.

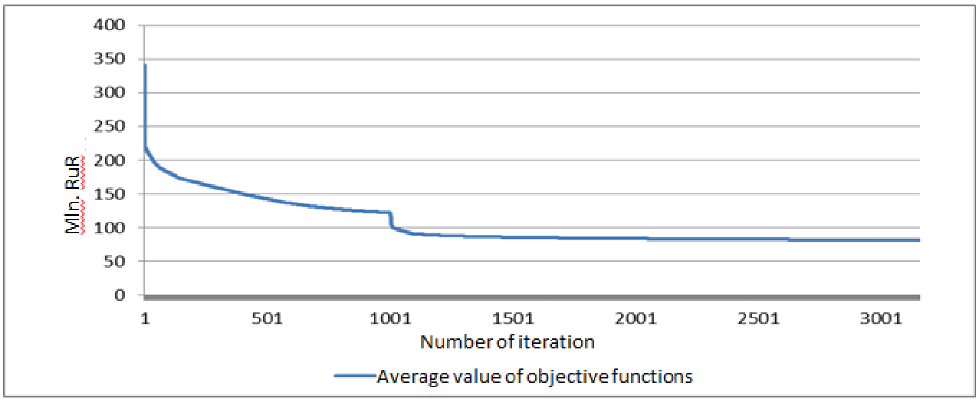


Figure 1: Average value of the objective function during genetic modeling

After 2250 iterations improvement of the objective function from generation to generation became insignificant, thus the algorithm was stopped. As a result of the modelling the solution was selected that had the lowest value of objective function. This choice offers supply scheme where some of the existing plants from cluster #3 should be closed, some of them relocated to farms (as milk processing plants), some should change their portfolio of products.

5 Results of the model, its stability and comparison of statistic indicators

After applying genetic algorithms of directed search to the problem of determination sourcing scheme for plants of cluster “not efficient plants”, the best solution provided results that suggest:

- modernization of capacities for current production portfolio – 19 dairy plants (20% of plants from cluster #3);
- change of portfolio type for 6 dairy plants (mostly that supposes transition from sour cream and pasteurized milk to butter and cheese production, which are in deficit after ban from deliveries from EU);
- plants closure with production lines relocation to existing factories - 20;
- plants closure with production lines relocation to milk farms (new milk processing plants construction) –37;
- plants closure – 14.

So, according to model result, only 14 plants should be closed as not efficient, other factories could be either modernized or relocated to another places with expected payback period less than 5 years.

Table 3: Values of Indicators before and after (planned) supply model implementation in Russia for plants of cluster #3

Indicator	Real monthly value 2012	Theoretical monthly value, suggested by model
Average weighed distance from milk farms to dairy plants, km	486	260
Average weighed distance from dairy plants to consumers, km 189 220 Production capacity utilization, %	36%	61%
Number of employee involved, people	51 380	36 720
Initial rate of return, %	3.6%	8.2%

As described in table 3, calculated results for most indicators improved significantly after implementation of sourcing optimization model. Average weighed distance from milk farms to dairy plants decreased significantly due to the fact, that model suggested to relocate many plants to milk farms. But at the same time distance to consumers increased by 31 km because in as is model all plants are located in cities, while in to be model many plants are outside of the towns.It also positively reflected production lines utilization and initial rate of return.

Suggested sourcing scheme is not very sensitive to market changes, because changes in consumers' demand have long trends, and model should be recalculated at least once a year to determine state policy and directions of supply for this important branch of national economy.

Conclusions

Determine and implement efficient state support is very important for all branches of economy, but it is essential for food products manufacturing industries. After ban for some categories of food from EU, that used to be a significant supplier especially on dairy market, it is important to have a specific long-term program that will help to create a strong, modern and competitive dairy industry in Russia.

In the article current dairy industry was analyzed: specific Russian features of supply, number of dairy plants, places of their location, current production portfolio and consumers' expectations and trends in their demand. As a result of a many-dimensions cluster analyses it was proved, that a big number of plants in Russia (97 plants, 28% of all dairy factories) are not efficient now with low utilization of production capacities, and low initial rate of return.

In the research a special approach offered to determine optimal number of dairy plants in Russia, and for plants of cluster #3 “not efficient plants” suggested one of five options: modernization for current production portfolio, modernization for new portfolio, capacity relocation to another plants, capacity relocation to new location (construct new plant in other location using production lines from plant of cluster #3) or plant closure. According to results obtained using genetic algorithms, 20% of not efficient plants should be closed, 6% should change their production portfolio, 21% should close with their production line relocation to another existing factories, 39% of plants should be closed with new milk processing sites constructed close to milk farms using their existiong production lines, and 15% of plants should be closed.

Model is not sensitive to changes of the input data and needs to be recalculated at least once a year to maintain state policy depending on changes in demand, consumers' expectations and other parameters. This method of research is not unique and can be easily modified for using and implementing in any other branch of economy that has several production plants. Further globalization of trade will make such models more important for transnational corporations.

References

[1] Arkhipova M., Arkhipov K. Supply process optimization using hubs for materials. *Proceeding of the 9th European Conference on Innovation and Entrepreneurship*. Edited by B.Galbraith. University of Ulster Business School and School of Social Enterprises Ireland Belfast, UK, 2014. p.43-50

[2] Arkhipova M., Arkhipov K. Optimal regional system design for a trade company in Russia Chapter X in Monograph *"In Marketing and logistics problems in the management of organization"*. Bielsko-Biala, Poland, 2011. – p. 214-229

[3] Arkhipov K.V. Model cost optimization for the supply of the head office to the branches *Scientific journal "Economics of Contemporary Russia."* – 2012 – No 4 (59);. – No 6, 2013;

[4] Arkhipova M. Modelling of innovative activity of manufacturing industries // *Applied Statistics* No 3, 2006. p.9-16;

[5] Arkhipova M., Arkhipov K. *Defining optimal regional logistic system for a trade company in Russia*. Platforms and innovations: In search of efficiency and effectiveness. EuroMot 2011, Finland;

[6] Ayvazyan SA, Mkhitaryan VS sl Applied statistics and econometrics bases. Textbook for high schools. – M .: UNITY, 2014;

[7] Ballou Ronald H. sl Business Logistic Management. – Prentice-Hall International, Inc., 1999;

[8] Coyle John J., Bardi Edward J., Langley John Jr. *The management of business logistics.A supply chain perspective.* — South-Western devise of Thomson Harming, 2003;

[9] Dybskaya V.V. sl Logistics for practitioners. Effective solutions in warehousing and cargo handling. – M .: VINITI, 2002;

[10] Global supply chain nonprofit consulting organization – www.supply-chain.org

[11] Langley John Jr., Jphn J. Coyle sl Managing Supply Chains: a logistic approach. – Cengage Learning Canada, 2008;

[12] National trade and marketing research agency – www.marketcenter.ru;

[13] Orekhov NA, Levin AG, Gorbunov EA *Mathematical methods and models in economics: Textbook. manual for schools.* M .: UNITY-DANA, 2004;

[14] Russian business consulting – www.rbc.ru;

[15] Russian national statistical agency – www.gks.ru;

[16] Shapiro J.F. *Modelling the Supply Chain.* Duxbury / Thomson Leading, 2001;

[17] World leader in milk products manufacturing – www.danone.com.

[18] The Nilson Report newsletter is the most trusted source of global news and statistics about the payment industry – http: www.nilsonreport.com

[19] Marketing research & consulting www.marketcenter.ru

Predicting Changes in the Number of Students for Making Management Decisions[1]

Ekaterina A. Khailenko and Oleg E. Avrunev
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: ekavka@yandex.ru, avrunev@ciu.nstu.ru

Abstract

The model for determining the economic efficiency of implementation of the curriculum for higher education programs is proposed. The regression model using the Generalized Lambda-distribution of the results of the Centralized Testing students as explanatory factors is applied to predict changes in the number of students and as consequence forecasting the effectiveness of implementation of the educational program. Groups of educational programs with similar changes in the student's contingent in the learning process were separated by using clustering. The results of investigation of the proposed algorithm were presented.

Keywords: Educational process, Generalized Lambda-distribution, clustering, regression model.

Introduction

Nowadays, researchers are directing their attention to estimating quality of accommodated educational service. One of these kind indicators is known as job placement of graduates [1]. However, this indicator permits to estimate only final result, but does not give information about educational process, which needs to making management decisions about correction of curriculums, programmes of work and others aspects of educational activities.

An important role in normative per capita financing [2] plays number of students index for separate educational programs. Imbalance parameters of curriculums and number of students lead to economic inefficiency of educational process and, as a result, to degradation education quality.

In this work new approach is proposed, where dependence between number of students index for separate educational programs and results of Centralized Testing (after school) as entrance examination is considered. This approach admits to estimate complexity of separate program and to forecast number of students.

The practice needs of these kind investigations making instrument to support management decisions, which could be built in University's information system for users who has not enough knowledge in statistical analysis methods.

[1] The report study was supported by the Russian Foundation for Basic Research, research project no. 14-07-31171 mol-a.)

1 Problem definition

The algorithm for making prognoses of economical efficiency of realization educational process is needed to develop. This algorithm lets to take estimation parameters of future number of students for separate educational programs.

Economical efficiency model of educational process

Characteristics of students' contingent are number of education program $i = 1...n$, n – quantity of educational programs in University, number of entrance year $j = 1...m$ and year $k = 1...t_{ij}$. Characteristics of curriculums for educational programs in respective entrance year can be different. The number of students in a particular course curriculum is let as s_{ijk}.

The educational program will be characterized by the total amount of hours on the course k teaching load (h_{ijk}):

$$h_{ijk} = a^{'}_{ijk} + a^{''}_{ijk}\left(\frac{s_{ijk}}{b''}\right) + a^{'''}_{ijk}[\frac{s_{ijk}}{b'''}] + a^{''''}_{ijk}s_{ijk}, \tag{1}$$

where $a^{'}$ is volume of lection hours, $a^{''}$ – volume of practice hours, $a^{'''}$ – volume of laboratory work, $a^{''''}$ – standard time for individual work with the student, $b^{''}$ – standard number of student in the educational group, $b^{'''}$ – standard number of student in the class.

In the context of normative per capita financing to save an acceptable payment one hour and avoid overloading teachers is needed to load volume per academic year for a particular educational program does not exceed the value directly depends on the number of students: $\alpha_i s_{ijk}$.

To solve the problem of obtaining the forecast of educational process efficiency we will use the minimum number of students in the educational process which is cost-effective. This quantity, which denoted by $s^{'}_{ik}$, can be taken by solving following equation:

$$s^{'}_{ik} = a^{'}_{ijk} + a^{''}_{ijk}\left(\frac{s_{ijk}}{b''}\right) + a^{'''}_{ijk}\left(\frac{s_{ijk}}{b'''}\right) + a^{''''}_{ijk}s_{ijk}.$$

Model of changing number of students' contingent

The model of changing students' contingent is proposed. The main factor in this model is number of students distribution depended on results of Centralized Testing as entrance examination. This approach also allows to implicitly consider the factor of interaction between students.

Let the number of students distribution according to the scores on the exam has universal Lambda distribution (GL-distribution), which depends on four parameters $\lambda_1, .., \lambda_4$ and is defined in terms of distribution quantiles [3,4]:

$$Q\left(u, \lambda_1, \lambda_2, \lambda_3, \lambda_4\right) = \lambda_1 + \frac{1}{\lambda_1}\left(\frac{u^{\lambda_3} - 1}{\lambda_3} - \frac{(1-u)^{\lambda_4} - 1}{\lambda_4}\right), 0 \leq u \leq 1.$$

The probability density function is:

$$f(x) = \frac{\lambda_2}{u^{\lambda_3 - 1} + (1-u)^{\lambda_4 - 1}}, 0 \le u \le 1, x = Q(u, \lambda_1, \lambda_2, \lambda_3, \lambda_4).$$

Depending on the values of the function GL-distribution describes a class of distributions, such as normal, exponential, student, chi-square, gamma, logistic, beta and others.

We denote as $X^i = \{X_1^i, X_2^i, ..., X_s^i\}$ sample of students' results of Centralized Testing for each i-th educational program, which quantity equal to n. Coefficient of change in the number of students in i-th the educational program for j-th admitted year between courses k and $k+1$ is denoted by $u_{ijk} = \frac{s_{ijk+1}}{s_{ijk}}$. On the assumption of the existence of groups of educational programs, for each of which there is a dependence of the next course contingent and distribution of the results of the exam on the current course, we can write this relationship as:

$$u_{ijk}^l = \theta_l^T L_{ijk} + \varepsilon_i \tag{2}$$

where l – number of education program group, L_{ijk} – set of GL-distribution parameters, i – number of education program, j – entrance year, k – course, ε_l – random component.

Thus, the algorithm needs to determine the similarity of both groups change in the distribution of the results of Centralized Testing, and the number of students for each educational program from course to course, and parameter estimation of (2) in these groups.

2 Adaptive algorithm for predicting changes in the number of students

The following algorithm to solve the problem is proposed by authors:

1. Step 1. Identification of the GL-distribution of students according to the results of the Centralized Testing for each educational program, each admitted year and each course is completed. Since the GL-distribution is completely described by the first initial and the second, third, fourth central moments, parameters of the GL-distribution are estimated by using the method of moments [4]. The number of students in each course of educational program is denoted by s_{ijk}.

2. Step 2. Set of data to determine the similarity of changes in the distribution groups is obtained for $j = j_0$, where j_0 corresponds to the least admitted year:

$$d_k = \left\{ \frac{s_{ijk}}{s_{ij1}}, \frac{\lambda_{1ij1} - \lambda_{1ijk}}{\lambda_{1ij1}}, \frac{\lambda_{2ij1} - \lambda_{2ijk}}{\lambda_{2ij1}}, \frac{\lambda_{3ij1} - \lambda_{3ijk}}{\lambda_{3ij1}}, \frac{\lambda_{4ij1} - \lambda_{4ijk}}{\lambda_{4ij1}} \right\}, k = 2...4.$$

3. Step 3. For set of data from step 2 clustering using k-means method [5] with the number of clusters l from 2 to the number of educational programs is run.

The ratio inter-cluster to the intracluster variances for a set of data from step 2, which appropriate set of next admitted year, is computed to determine the optimal number of clusters after the clustering for the next l.

4. Step 4. Estimation model (2) parameters $\hat{\theta}_l$ is taken for each educational program group from step 3 by using adaptive algorithm [6]. This algorithm allows to obtain estimates for violations of the assumptions of normality of the random variable error ε_l.

5. Step 5. Predicting of violation of the educational program effectiveness for the next course for each educational program is performed by using obtained estimates previously. The program is effective if the following inequality $\hat{\theta}_l^T L_{ijk} s_{ijk} \geq s'_{ik}$ is true.

2.1 Results of investigations

For research the results of the adaptive algorithm work the data set of the student contingent of the Novosibirsk State Technical University (NSTU) and curriculum parameters for 2011-2014 admitted year was used. These data were obtained from the information system NSTU. Number of education program is 55, number of students stream is 226, number of students is 10249.

The example of changing identification of the GL-distribution of students according to the results of the Centralized Testing for educational program "Computer Science and Engineering" from 2011 entrance year, which was taken from step 1 of proposed algorithm, is presented on figure 1.

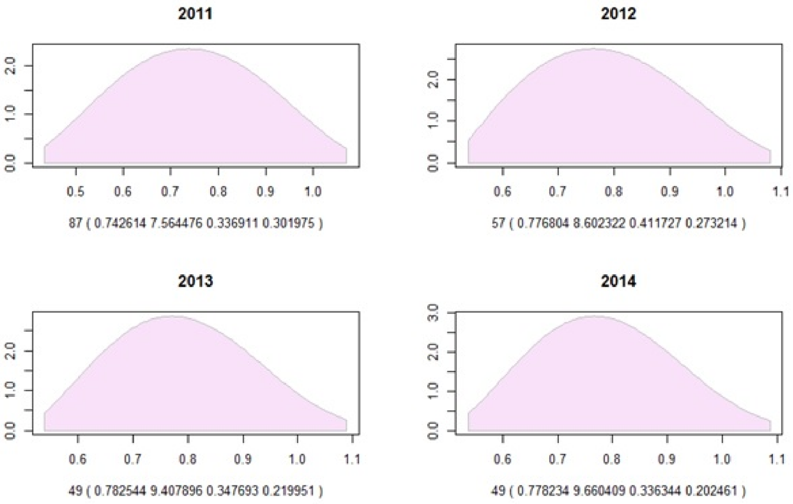


Figure 1: Changing identification of the GL-distribution of students according to the results of the Centralized Testing from 2011 entrance year

As seen in Figure 1, the distribution of the General Testing results at entrance to university (in 2011) is symmetric, as evidenced by the values of the parameters

$\lambda_3 \approx \lambda_4$. Further, after sending down part of the students, the distribution becomes asymmetrical, $\lambda_3 \neq \lambda_4$.

After the clustering of data (steps 2 and 3) the optimum number of clusters of educational programs was determined as 4. Figures 2 and 3 is shown diagrams of range changes in the number of student within clusters obtained and the entire set of data as a whole.

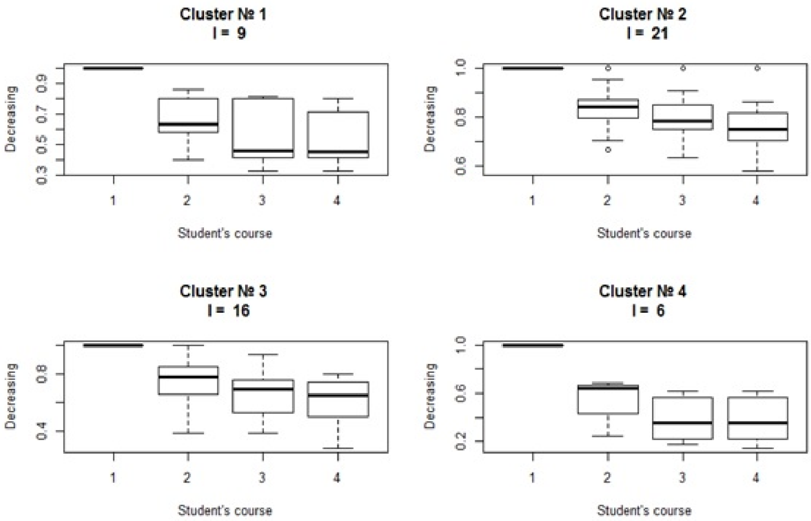


Figure 2: Diagrams of range changes the number of students for 4 clusters of educational programs, l - number of educational programs in the cluster

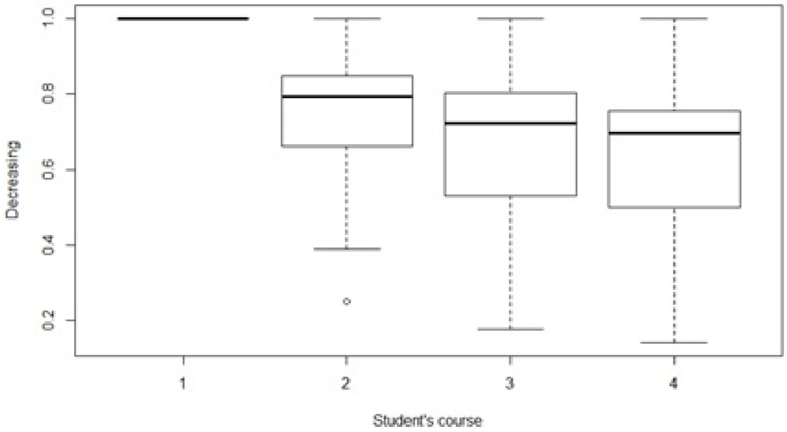


Figure 3: Diagram of range changes the number of students for all educational programs

It is seen from figures 2 and 3 that within the framework of clusters range of changing in number of students is much smaller than that observed for all educational programs.

We calculate the parameters of a regression model (2) using the least squares (LS) method and adaptive method based on GL-distribution. The values of the obtained coefficients are shown in table 1.

Table 1: Results of estimation regression model parameters

Method	Parameters	Cluster No.1	Cluster No.2	Cluster No.3	Cluster No.4
Adaptive	$\hat{\theta}_1$	0,927	0,766	1,051	1,139
	$\hat{\theta}_2$	0,006	0,004	0,006	0,006
	$\hat{\theta}_3$	0,025	0,039	0,019	0,144
	$\hat{\theta}_4$	0,435	0,057	-0,050	-0,057
LS	$\hat{\theta}_1$	0,807	0,984	1,135	1,143
	$\hat{\theta}_2$	0,005	0,005	0,006	0,005
	$\hat{\theta}_3$	-0,001	0,074	0,028	0,141
	$\hat{\theta}_4$	0,467	0,050	-0,026	-0,068

We obtain forecasts effectiveness of the achievement of educational programs using estimation of obtained previously parameters and expression (3). Following results are taken by comparison known and predicted number of students' contingent and shown in table 2.

Table 2: Results of effectiveness forecast

Perform data clustering	Estimation method	Incorrect forecasts ineffectiveness,%
Not perform	LS	11.61
Not perform	Adaptive	17.74
Perform, 4 clusters	LS	12.25
Perform, 4 clusters	Adaptive	10.97

As seen from Table 2, the implementation of the provisional application of clustering and adaptive estimation method gives the smallest forecast error effectiveness of the implementation of the educational program.

Conclusions

In this work the adaptive algorithm for predicting changes in number of students has been proposed. Results of proposed algorithm investigation are discussed. The advantage of the algorithm is simplicity. Instead of predicting the probability of sending down of individual students analysis is taken for groups of students in general. At the same time the use of pre-clustering identifies groups of educational programs which demonstrates changes in the characteristics of students.

The algorithm can be extended for using as a regressors distribution parameters of the student contingent on other quantitative factors. Also, this algorithm can be modify for effectiveness predicting using penalty function $f(x)$:

$$\sum_t \left(\sum_{i=n_t \ldots n_{t+1},j,k} h_{ijk}\left(s_{ij\hat{k}+1}\right) - \alpha_i s_{ij\hat{k}+1} \right)$$

where t - faculty number, $s_{ij\hat{k}+1} = \hat{\theta}_l^T L_{ijk} s_{ijk}$. This approach allows to take into account contiguity of educational programs and compensation for insufficient number of students one program for sufficient amount the other, for example within faculties.

References

[1] Timofeev V.S., Borisova A.A., Avrunev O.E. (2014). The readiness of students to succed in their profession: a start of professional formation *Izvestiya of Irkutsk State Economics Academy*. Vol. **3**, pp. 53-62 (in Russian).

[2] Ministry of Education and Science of the Russian Federation. Switch to normative per capita financing educational program of higher education http://минобрнауки.рф/проекты/нормативно-подушевое-финансирование

[3] Karian Z.A., Dudewicz E.J. (2000). *Fitting statistical distributions: the Generalized Lambda Distribution and Generalized Bootstrap methods*. CRC Press LLC, New York.

[4] Lakhany A., Mausser H. (2000). *Estimation the parameters of the Generalized Lambda Distribution. ALGO research quarterly*. Vol. **3**, pp. 27-58.

[5] Oldenderfer M. S., Bleshfild S. K., Kim Dzh.-O., Myuller Ch. U. (1989). *[Factor, discriminant and cluster analysis*. Finansi i statistika Publ. Moscow (in Russian)

[6] Timofeev V.S., Khailenko E.A. (2010). Adaptive estimation of regression models parameters using universal Generalaized Lambda-Distribution *Proceedings of the Russian Higher School Academy of Sciences* Vol. **2(15)**, pp. 25-36

Estimating Polychoric Correlations for Mixed Data of Graduate Employment Survey

Anastasiia Yu. Timofeeva and Alena A. Borisova
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: a.timofeeva@corp.nstu.ru, bborisova2012@yandex.ru

Abstract

The problem of choosing methods for correlation estimation has been solved for mixed data pertaining to the college graduate's employment. Analysis of such data often reveals the relationship between indicators. In such case it is difficult to implement resource-monitoring and to interpret results. In order to choose the necessary and sufficient set of indicators the correlation analysis is usually used. A distinctive feature of studies of the graduate situation in the labor market is a combination of different measurement scales of indicators. This leads to the problem of the correlation analysis for mixed data. Using bootstrapping the properties of sample estimates of Spearman's rank coefficient and polychoric correlation coefficient have been investigated obtained by the maximum likelihood, least squares, least absolute deviation and minimum chi-square methods. A comparative analysis of the estimates has been conducted. The study is focused on problem cases with significant difference of the estimate's distributions which are identified by the sign test. Recommendations have been given regarding their applicability for the analysis of mixed data.

Keywords: polychoric correlation, ordered categorical data, count data, employment, monitoring, graduate.

Introduction

In recent years, in order to analyze the efficiency of higher education institutions the students' opinions, success stories, information about their situation in the labor market are increasingly involved. Thus, as a rule, to assess the parameters of graduate employment a number of self indicators measured by Likert scale are used [1]. Many such indicators are correlated, thus their number can be reduced to provide a less expensive monitoring. The problem of reducing the number of indicators is solved on the basis of their correlation matrix using factor analysis. Depending on the measurement scale, there are different measures of association between variables. Likert scale is ordered categorical, so polychoric correlation coefficient is most suitable.

To reduce the distortion of employment parameters' measure due to subjective judgment, as opposed to the standard approach, this study is focused on objective measures of employment (value of salary, workweek, etc.). Consequently, the mixed data are obtained, that is measured in different scales. Therefore it is necessary to choose the correlation measure which allows to handle such data. Further, the possibility of using for this case polychoric correlations has been studied. Let us take a closer look at their definition and methods of estimation.

1 Problem definition

Consider the relationship between two indicators of graduate employment, for example, the management level and the amount of salary. Suppose that these indicators are dimensionless, described as continuous random variables ξ_1 and ξ_2 with bivariate standard normal distribution. Due to the imperfection of the measurement scale (for example, use of an ordinal scale, Likert scale, rounding values of wages, etc.) during the survey researchers can not obtain continuous values of indicators. There are discrete random variables x_1 and x_2 obtained by grouping, i.e. the partition of the range of values of random variables ξ_1 and ξ_2 into intervals. Let the number of such groups (possible values x_1 and x_2) be n_1 and n_2. It is assumed that x_1 takes values from 1 to n_1, x_2 — from 1 to n_2. If the bounds of these intervals are α_{i1}, $i = 0, 1, \ldots, n_1$, α_{j2}, $j = 0, 1, \ldots, n_2$, then the probability of the indicator values measured in the survey is

$$P(x_k = i) = P(\alpha_{(i-1)k} < \xi_k < \alpha_{ik}) = \Phi(\alpha_{ik}) - \Phi(\alpha_{(i-1)k}),\ k = 1, 2$$

where $\Phi(\cdot)$ is standard normal distribution function, $\alpha_{0k} = -\infty$, $\alpha_{n_k k} = +\infty$. The probability of each combination of random variables x_1 and x_2 is defined as

$$\begin{aligned} p_{ij} = P(x_1 = i,\ x_2 = j) = P(\alpha_{(i-1)1} < \xi_1 < \alpha_{i1},\ \alpha_{(j-1)2} < \xi_2 < \alpha_{j2}) = \\ = \Phi_2(\alpha_{i1}, \alpha_{j2}, \rho) - \Phi_2(\alpha_{(i-1)1}, \alpha_{j2}, \rho) - \Phi_2(\alpha_{i1}, \alpha_{(j-1)2}, \rho) + \Phi_2(\alpha_{(i-1)1}, \alpha_{(j-1)2}, \rho), \end{aligned} \tag{1}$$

where $\Phi_2(z_1, z_2, \rho)$ is bivariate standard normal distribution function with correlation ρ between random variables ξ_1 and ξ_2.

During survey the proportion of answers, or relative frequency, d_{ij} is observed with i-th value of the random variable x_1 and j-th value of x_2. Usually such proportions are represented in the form of contingency tables. The problem is to estimate the unknown parameters of the bivariate distribution of random variables x_1 and x_2 based on observed values d_{ij}. Estimate of ρ in this model is called the polychoric correlation coefficient. In current study we consider a two-step approach [5]. The first step is to find estimates for the interval boundaries α_{ik} as quantile of corresponding marginal empirical distributions:

$$\hat{\alpha}_{i1} = \Phi^{-1}\left(\sum_{l=1}^{i}\sum_{j=1}^{n_2} d_{lj}\right),\ i = 1, \ldots, n_1 - 1,$$

$$\hat{\alpha}_{j2} = \Phi^{-1}\left(\sum_{l=1}^{j}\sum_{i=1}^{n_1} d_{il}\right),\ j = 1, \ldots, n_2 - 1.$$

In the second step estimates of the interval bounds are substituted into (1) and estimation of correlation coefficient $\hat{\rho}$ is obtained. The current estimation methods have a number of restrictions for mixed data processing. Let us analyze possibilities of methods and justify the ways to overcome restrictions.

2 Methods of correlation estimation

The most popular method of polychoric correlation estimation is maximum likelihood (ML). For the joint discrete distribution of random variables x_1 and x_2 under the assumption of independence of observations the log-likelihood of the sample [5] is

$$\ln L = C + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{ij} \ln p_{ij} \tag{2}$$

where C is a constant. In this case each d_{ij} is a fixed value for the concrete sample. Obviously, if during survey some combination of the random values x_1 and x_2 is not observed, then the corresponding contingency table cell will be zero, and theoretical probability value (1) will not affect the value of the function (2). In other words, the log-likelihood function is not sensitive to the value of a random variable with zero frequency. At the same time, it will be sensitive to those frequencies such that theoretical probabilities are close to zero. That is on the tails of joint distribution. If the probability p_{ij} will be close to zero at nonzero frequency d_{ij}, weight of such terms in the ratio (2) tends to $-\infty$. Even much worse case is estimation based on the method of minimum chi-square [3] ($\min \chi^2$). Here, the loss function

$$\chi^2 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(p_{ij} - d_{ij})^2}{p_{ij}}$$

is minimized. Consequently, squares of residual with almost zero probability will have the greatest weights. This problem is somewhat weakened by using other measures of differences between the observed and theoretical values of probabilities [2]. The most natural would be to take the distance between the vector of observed frequencies $D = (d_1, d_2, \ldots, d_{n_1})$ with $d_i = (d_{i1}, \ldots, d_{in_2})$ and the vector of corresponding probabilities $P = (p_1, p_2, \ldots, p_{n_1})$ with $p_i = (p_{i1}, \ldots, p_{in_2})$. To determine the distances L_p-norm is normally used and defined for any vector Z with elements z_i as $\|Z\|_p = (\sum_i z_i^p)^{1/p}$. Then we obtain the loss function of the form

$$\Delta = \|P - D\|_p. \tag{3}$$

The case of minimizing (3) by $p = 2$ corresponds to the method of least squares (LS), by $p = 1$ we obtain the method of least absolute deviation (LAD). It is expected that the use of LAD results to less sensitivity to non-zero frequencies on the tails of the joint distribution.

As alternative of polychoric correlations the Spearman's coefficient is considered, which is defined as the Pearson correlation coefficient between ranks of the values of the variables. Its advantage is the ease of computation. Also through the use of ranks instead of the raw data (for example, salary) skewness of distributions and heavy tails must not distort the estimation results. However, it is known the method do not work very well with ties (duplicate values).

The described techniques have been implemented in the statistical environment R. To calculate the values of bivariate standard normal distribution function the algorithm have been used which is proposed in [4]. To speed up the calculations based on this algorithm the vectorized function was built that returns a matrix of probabilities p_{ij}. A one-parameter optimization was carried out using a basic function optimize{stats} in the interval $\rho \in (-1, 1)$.

3 Empirical data

As an empirical base the data from a graduate survey at Novosibirsk and Irkutsk have been used. The sample size was 640 valid observations. The graduate employment parameters on the first place of employment was investigated. 18 variables were selected that are measured in different scales. The variables are divided into groups depending on the number of categories (levels).

- Ordered categorical (up to 5 levels): indicator of wage growth, form of employment, nationality of ownership, level of management's position, type of labor contract, frequency of attraction to work overtime, number of staff in organization, frequency of career promotion.
- Count (5 to 30 levels): number used in the job search channels of employment, duration of job search, duration of adaptation in the organization, measure of using a professional capacity, failure rate of employers in finding jobs.
- Continuous with the rounded values (more than 30 levels): workweek, salary on and after probation, duration of work in organization, maximum wage value.

Although continuous indicators with rounded values seem to be less problematic for the correlation analysis but the situation gets worse in the presence of outliers. Figure 1 illustrates the problem of heavy-tailed distribution on the example of wages. Since the data are discrete, then it is necessary to display on the scatter plot the observed frequencies of various combinations of values. For this purpose different markers are used which become larger and darker with increasing value of the relative frequency. Bounds of relative frequency intervals, expressed as a percentage, are given in the legend to Figure 1. Although in general both substantive and from Figure 1 a close linear relationship is evident, but it is expected that outliers in salary after probation should increasingly affect the correlation estimation by minimum chi-square and less the estimation of Spearman's coefficient because of its rank character.

During the analysis of relationships between ordinal variables the problem of the dominance of a certain category occurs. For example, 64.2% of the respondents concluded indefinite labor contract with the organization and 82.5% of graduates work full-time. Contingency table between the type of labor contract and form of employment is shown graphically in Figure 2. Here area of the rectangle is proportional to the frequency. It can be seen that the proportion of full-time workers is reduced with a decrease of the contract term. However the Spearman's coefficient will probably indicate a weak correlation due to problems of ties (many identical values).

Moreover the problem of leptokurtic distribution occurs, e.g. for the workweek: 67.3% of the respondents have a 40-hour day. At the same time, the range of work

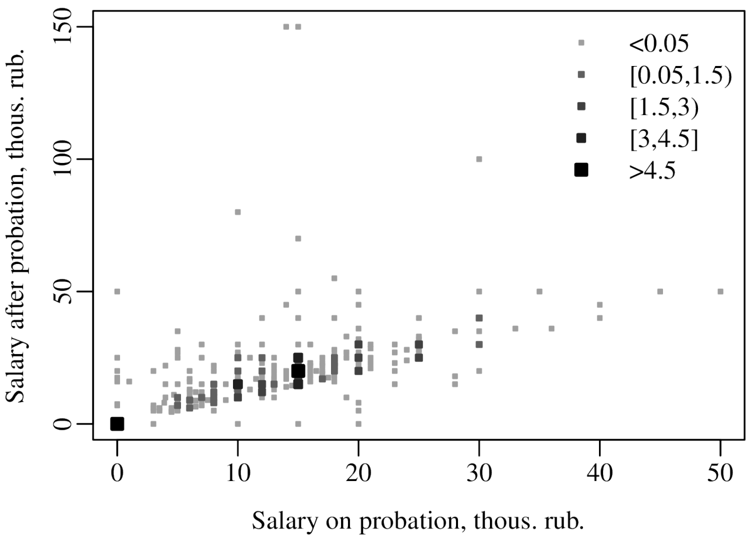


Figure 1: The scatter plot of salary on and after probation

hours is very large, from 0 to 80 hours a week. And if we analyze the relationship between this indicator and a categorical variable with the dominant level (for example, a form of employment), we have difficulty by using a number of methods for correlation estimating due to their sensitivity to non-zero frequencies on the tails of the joint distribution of the indicators. Although this relationship is obvious for economic reasons, but a number of polychoric correlation estimates will probably underestimate it (for example, estimates on the basis of $\min\chi^2$ and ML).

4 Comparative analysis of the results

The considered methods of correlation estimation was analysed using statistical technique of bootstrapping. For this purpose a random samples with replacement are drawn from the available data with the same sample size. The number of samples was $N = 500$. For each sample the correlation matrix between the 18 analyzed variables were estimated by different methods. As a result, the empirical distributions of estimates were recovered. On their basis, the 95% confidence intervals were constructed for each of the estimates. If zero was not included in the confidence interval, the conclusion about the significance of the correlation coefficient was made. Table 1 shows the fraction of significant coefficients $\hat{\rho}^*$ of all the possible 153 coefficients. The largest number of significant coefficients is contained in the correlation matrix estimated by ML, the smallest is by LAD. It suggests that LAD-estimates have large standard errors. In general, they are about 1.5-2 times higher than by other methods.

For application of factor analysis the significance of coefficient estimates is not so much important as their absolute values, because if they are bigger, then the explained variance of extracted factors will be greater. It allows the use of a smaller number of indicators for the diagnosis of employment. In order to characterize how different

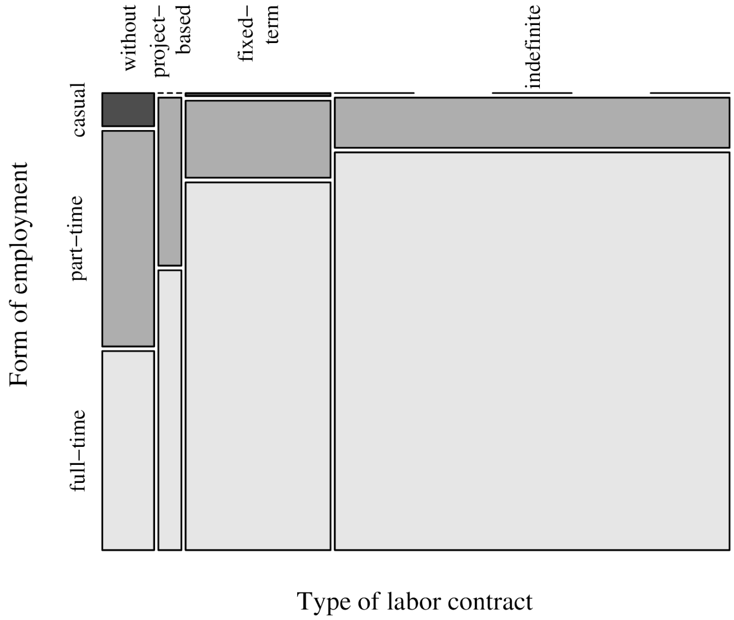


Figure 2: The contingency table of the type of labor contract and form of employment

Table 1: Characteristics of correlation matrices estimated by different methods

	Spearman	ML	LS	LAD	$\min \chi^2$
Fraction of $\hat{\rho}^*$	0.379	0.399	0.301	0.222	0.373
λ_{max}	2.81 (0.10)	3.11 (0.14)	3.44 (0.16)	3.51 (0.15)	2.74 (0.18)
$\|R\|$	0.051 (0.008)	0.009 (0.003)	0.0003 (0.0005)	0.0003 (0.0006)	0.035 (0.013)

elements of the correlation matrix would have to be, the maximum eigenvalue λ_{max} of correlation matrix R is used. It is proportional to percentage of explained variance of the principal component. Table 1 summarizes the characteristics of the middle value (median) and dispersion (MAD, mean absolute deviation, is given in parenthesis) of λ_{max}. The correlation matrices obtained by LAD and LS are characterized by maximum variance accounted for by the first factor, i.e. correlation estimates are on the average larger in absolute value. Minimum values of λ_{max} are obtained by the methods the minimum chi-square and Spearman, i.e. these methods underestimate the correlation on the mixed data.

However, the use of LAD and LS leads to the fact that the correlation matrix loses the property of positive semi-definite. It makes impossible to perform a factor analysis. Table 1 shows the median (MAD is given in parenthesis) of correlation matrix determinants $|R|$. The results of Spearman's coefficient and minimum chi-square are least problematic, where negative values rarely occur (median is much more MAD). For ML-estimates negative values of $|R|$ are observed in 4.4% of cases, it is also quite acceptable. The use of methods based on minimizing (3) leads to the fact that in the 20-25 % of samples $|R| < 0$. So it is recommended to compute the nearest correlation matrix in the weighted Frobenius norm using a function nearPD{Matrix}. However, it should be emphasized that the values of estimates could significantly distort.

To compare the results of estimation by different methods the hypothesis of the equality of correlation estimate distributions were tested. Since the estimation by different methods was performed on the same samples, the estimates can not be considered as independent. And to test the hypothesis sign test was used. It is based on the difference

$$Z_k^{(ij)} = \hat{\rho}_k^{(i)} - \hat{\rho}_k^{(j)},\ k = 1, \ldots, N,\ i, j = 1, \ldots, 5,\ j > i$$

where $\hat{\rho}_k^{(i)}$ is the correlation coefficient estimated for k-th sample by i-th method. The test statistic is the normalized sum of such differences. In case of significant deviation from zero the inequality of distribution concludes.

The largest share of estimated coefficients with unequal distributions is obtained by ML and $\min \chi^2$ (97.4%). Closest estimates are obtained by LS and LAD, although the proportion of estimates with unequal distributions is also very high (0.843). To characterize these differences a location estimate is calculated as the median of values $Z_1^{(ij)}, \ldots, Z_N^{(ij)}$. Table 2 shows the maximum positive and negative values of the location. In the upper triangle of the matrix i corresponds to the method specified in the string, j — in the column, at the bottom — vice versa. In many cases values of location are very large. They just correspond to problem situations described above.

Table 2: Comparison of estimation methods by the sign test

	Spearman	ML	LS	LAD	$\min \chi^2$
Spearman	-	0.143	0.142	0.159	0.263
ML	-0.200	-	0.145	0.177	0.305
LS	-0.473	-0.418	-	0.129	0.739
LAD	-0.467	-0.414	-0.069	-	0.735
$\min \chi^2$	-0.201	-0.174	-0.285	-0.322	-

Thus most significant estimates of location (greater than 0.3 or less than -0.4) appear by analyzing the relationship between the form of employment and the workweek. The estimate by the minimum chi-square was not significant, Spearman's and ML estimates were on average about 0.33 and 0.38, estimates obtained by LS and LAD were more than 0.8. Consequently, as expected, the method minimum chi-square was the most sensitive to nonzero frequencies at the tails of joint distribution. ML is less sensitive, but still underestimates the correlation.

About the problem with the predominance of certain categories (Figure 2), the largest location is characteristic of Spearman's coefficient compared with estimates obtained by minimum chi-squared method and ML, it is about -0.2. Finally, estimates of correlation between wages on and after the probation obtained by the minimum chi-square are clearly underestimated (average is 0.48). Spearman's coefficient and ML are more robust and give an average estimate of 0.74. Methods based on the minimization of (3) completely ignore outliers and provide the highest values of coefficients.

Conclusions

The obtained results allow us to give some recommendations about the applicability of correlation estimation methods for mixed data. Thus, Spearman's correlation coefficient and polychoric coefficient estimated by the minimum chi-square method have the advantage of positive semi-definite correlation matrix. However, by using their the low absolute values of correlation estimates are usually obtained. As a result factor analysis on such correlation matrices leads to low explained variance of the principal component. Maximum likelihood method gives acceptable correlation matrix in terms of the properties of the positive semi-definite. In general, the estimated correlation coefficients are sufficiently large in absolute values which provides a high percentage of explained variance of the principal component. But ML gives failures in the heavy-tailed distribution and dominance of one category over the other. LAD and LS are less sensitive to such problems. They provide the highest proportion of explained variance of the principal component. However, in every fifth case positive semi-definiteness of the correlation matrix is not ensured.

Therefore, for the analysis of mixed data it is necessary to search a compromise between the ML and methods based on minimization of the distance from the observed frequencies to theoretical probabilities. This version should provide a positive semi-definiteness of the correlation matrix and the low sensitivity to non-zero frequencies in the tails of the joint distribution. This should be the subject of further research into the correlation analysis on the mixed data.

References

[1] Cheruiyot T.K., Tarus D.K. (2015). Modeling Employee Social Responsibility as an Antecedent to Competitiveness Outcomes. *SAGE Open*. Vol. **5**.

[2] Ekström J. (2011). A generalized definition of the polychoric correlation coefficient. *Department of Statistics, UCLA*.

[3] Kendall M., Stuart A. (1961). *The Advanced Theory of Statistics: Inference and relationship*. Charles Griffin and Co., London.

[4] Meyer C. (2013). Recursive Numerical Evaluation of the Cumulative Bivariate Normal Distribution. *Journal of Statistical Software*. Vol. **52**, pp. 1-14.

[5] Olsson U. (1979). Maximum Likelihood Estimation of the Polychoric Correlation Coefficient. *Psychometrica*. Vol. **44**, pp. 443-460.

The Influence of Educational Environment on the Student's Knowledge Level based on the PISA Survey Data

Svetlana Eremina
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: ereminasa@rambler.ru

Abstract

Education system is one of the most important parts of socialization and society's life. It's also affected the economic potential and the ability to generate social or material capital. Modern education system is effective, but in any case we can suggest that there are some perspectives for its development. The main idea of this study was to show what do we can change in educational process to make it more successful. The aim of the study was to determine factors that can influence on the students' educational effectiveness. To achieve this aim international PISA survey data were used in which more than 60 countries (including Russia) participated. This article presents how educational environment can influence on perspectives of changes in the knowledge degree on the base of PISA data.

Keywords: students' knowledge degree, PISA survey, educational environment, teacher's strategy, interest in education, educational process.

Introduction

It is generally agreed today that education is the important process of our life. This was the reason that in last decade we could notice a lot of researches about tendency of changes in this sphere. As one of the most significant project was the Programme for International Student Assessment (PISA study). It has been holding since 2000 in more than 60 counties and engaged about million 15-years students. Now we have different approaches based on this study, but most of them consider general tendency and are not focused on results in one country (e.g. Russia)[4,5,7].

The aim of this study was to determine which school environmental factors can influence on the knowledge degree. To achieve this aim following tasks were set:

- emphasize all indicators which can be important for changes of knowledge degree;
- determine how exactly chosen factors can influence the knowledge degree.

The general hypothesis was that one of the most important factors which can define effectiveness of the educational process is the teacher's strategy (how teacher interact with students in the classroom).

1 Materials and methods

To confirm proposed hypothesis PISA survey data were used which is available to download on the OECD website [2]. The survey questionary consists of 54 questions from different areas, such as living standard, social status of the student's family, school environment and student's knowledge degree. In regard to knowledge degree there are questions about three general directions: mathematic, reading and science. The analysis presented in this article based on questions about mathematical knowledge, because in PISA survey this sphere was shown more detailed than other [4].

For further analysis survey data for 2012 year was taken. The total sample of 485 490 participants includes 5 231 students from Russia (mean age 16 years, 2 617 males and 2 614 females). Sample of the Russian students allows us to emphasize main tendency in Russian educational system. For this purpose from all variables question which may characterizes students' knowledge degree was chosen. This question includes such statements as: "I get good grades in mathematics", "I learn mathematics quickly", "I have always believed that mathematics is one of my best subjects" etc.

2 Main factors that ensure the level of knowledge

To identify which statements about knowledge degree can influence the level of knowledge polychoric correlation and factor analysis (proportion of variance explained = 58%) were used to formed one principal component based on initial variables. Next we should indicate all links between dependent variable and other items. To achieve it partial correlation analysis with Spearman's coefficient in SPSS program was applied that allow us to show in one table all correlations that exist between different items. Thus the list of significant factors, except items which didn't have any links with dependent variable was obtained.

As significant following factors were emphasized:

Interest in the learning process (Spearman's corr. 0.2-0.3, p=0.001). This group includes such items as: "I enjoy reading about mathematics", "I do mathematics because I enjoy it" etc. Besides, students noticed that mathematics can be useful for their future career: "Learning mathematics is worthwhile for me because it will improve my career", "I will learn many things in mathematics that will help me get a job". Thereby terminal and instrumental values of the learning process have positive correlation with students' knowledge in mathematics.

Influence of the social environment. By this we mean positions of the parents and friends on the educational item: in case when student's parents or friends like mathematics, he or she demonstrates high knowledge degree (for parents position Spearman's corr. 0.2, p=0.000, for friends position Spearman's corr. 0.12-0.2, p=0.01).

Next two factors are the time, which students spend on their homework and preparation for exams (Spearman's corr. 0.21-0.37, p=0.000), and how students may handle with big volume of information or can easily link facts together (Spearman's corr. 0.17-0.27, p=0.000).

Except above factors the links with dependent variable also exist for such items as: "I participate in a mathematics club", "I play chess". Despite those correlations these questions cannot be used as factors because it's hard to determine direction of the influence (participating in a mathematic club influence the knowledge degree or straight conversely). In other words we have bilateral causality between such type of questions an dependent variable.

This way we had the list of factors which correlated with knowledge degree, but their number is still too big to use them for creating a regression model. Besides, each group includes several questions, and some of them reflect similar sense (furthermore, it may be the reason of multicollinearity or deformation of regression model). As one of solutions in this case dimension reduction procedure (factor analysis) was applied. Another reason why we can use this procedure is that it forms one variable with quantitative scale (which is convenient for regression analysis) instead several question with ordinal scale.

Both factor analysis and regression model were performed using a free software environment for statistical computing R. Selected variables were joined in several groups on the ground of their means and questions that they belong. Further one or two integrated factors which reflect level of severity current index were formed. They are present below:

F1 - interest in the learning process (proportion of var. 58%);

F2 - ability to process information (proportion of var. 53%);

F3.1; F3.2 - influence of the social environment (friends and parents position, cumulative var. 63%);

F4 - time which students spend on the preparation for classes (proportion of var. 55%).

3 Regression analysis

Thus five factors which reflected different spheres of the students' school live were formed. These indicators were used for creating linear regression model with stepwise method of variable selection (that allow us to exclude insignificant variables). As a result following model was received:

$$\hat{Y} = -0.01 + 0.33F_1 + 0.18F_2 + 0.34F_4.$$

$\hat{Y}$ - the level of knowledge (dependent variable)

t-value for factors: $t_1 = 16,1$; $t_2 = 9,4$; $t_4 = 15,7$

Adjusted R-squared presented model: 0.497, F-statistics 535.6, model is significant at the 99.9% confidence level.

Regression model allows us to determine how current factors can influence on changes of dependent variables. In this case we can say that the knowledge degree depends on: interest in mathematics, time which students spend on the classes and homework, the ability to process information. Link between knowledge degree and ability to handle a lot of information may be explained by students' learning capacity and lack of interest in further analysis. With regard to other factors such as interest in learning and time which students spend on preparation for classes we can use them to form certain educational space. Moreover, those factors can be useful as instrument of students' motivation. Influence of the social environment (friends and parents position) was excluded from the regression model as variable with low significance.

4 Cluster analysis

To understand how we can influence such complex factor as interest in education process another type of classification (cluster analysis) was used. This analysis allows us to combine respondents in groups with similar opinion or level of rank of selected indicators. As those indicators items about students' interest in mathematics ("I enjoy reading about mathematics" etc.) were used. As a result all cases were divided on three main groups, which are: students who are interested in mathematics as science (n = 1138), those who think that mathematics is important for their future career (n = 1196), and students who don't interest in mathematics at all. Thereby all links between group which students belong and teacher's strategy (by this we mean different instruments which teachers use during classes) were defined. Also cross-tabs analysis was used to show how evaluations of variables can change, as well as the correlations and links between them.

Thus, with this data we can suggest that low but systematic links between students' interest in learning and teacher's educational strategy (Spearman's corr. 0.15-0.22, chi-square 47-93) exist. Further there are some elements of this strategy:

- The teacher shows an interest in every student's learning (Spearman's corr. 0.15);
- The teacher continues teaching until the students understand (Spearman's corr. 0.21);
- The teacher tells how to get better and gives feedback on students' strengths and weaknesses in mathematics (Spearman's corr. 0.2);
- Students work in small groups (Spearman's corr. 0.2) ;
- The teacher asks students to decide on own procedures for solving problems and presents problems in different contexts (0.16).

In addition to teachers' strategy, interest in education associated with position of parents and friends (Spearman's corr. 0.2, p = 0.001). Parents' position is the complex factor which is also linked with knowledge degree and time spending on education. As to time on preparation for classes this factor also correlates with teachers' strategy, especially with his/her feedback on students' strengths and weaknesses (Spearman's corr. 0.125, p = 0.01) and task with multiple solutions (Spearman's corr. 0.13, p = 0.005).

Conclusions

Excluding factors on which we cannot influence, we have the list of factors that can be useful for increase of knowledge degree. They are: interest in education, students' motivation to learn and complex task with multiple solutions.

The hypothesis which we suggested in the beginning of this study (the knowledge degree depends on the teachers' strategy) wasn't confirmed. In fact this factor shows indirect influence through interest in education. Thus, besides student's ability, classes' effectiveness and teachers' qualification are important to create environment that will prosper to increase students' interest in education.

References

[1] Kovaleva G., (2013). What is the actual level of financial literacy of Russian students? (according to a study PISA-2012). *Center OKO*. URL http://www.centeroko.ru/download/Present-FL-PISA2012.zip

[2] OECD. About PISA. *PISA Overview*. URL http://www.oecd.org/pisa/aboutpisa/.

[3] OECD, (2015). Education Policy Outlook 2015: Making reforms happen *PISA Overview*. URL http://www.mecd.gob.es/dctm/inee/eag/e-book-education-policy-outlook-2015.pdf?documentId=0901e72b81bdc851.

[4] OECD, (2014). PISA 2012 Results in Focus: What 15-year-olds know and what they can do with what they know. *PISA Overview*. URL http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

[5] Schneider M., (2009).The International PISA Test. *Education next*. Vol.9 , pp. 69-74.

[6] Tsukerman G., Kovaleva G.,Kuznetsova G., (2014).Evolution of Reading Literacy, or The New Adventures of the Push Me Pull You. *Voprosy obrazovaniya*. Vol.5 , pp. 284-300.

[7] Tucker M., (2013).PISA On the Teacher as Professional *Education Week*. URL http://blogs.edweek.org/edweek/top-performers/2013/12/pisa-on-the-teacher-as-professional.html.

On Sequential Estimation of a Periodic Signal Distorted by an Autoregressive Noise

Tatiana V. Emelyanova and Victor V. Konev
National Research Tomsk State University, Tomsk, Russia
e-mail: tv_em@mail.ru, vvkonev@mail.tsu.ru

Abstract

This paper considers the estimation problem for a trigonometric signal in a discrete time from observations with an additive noise described by a stationary autoregressive process with unknown parameters and unknown distribution. We propose a one-step sequential procedure to estimate signal coefficients, which provides a given mean-square accuracy of estimates for any values of the nuisance parameters. An asymptotic formula for the mean duration of the procedure has been obtained.

Keywords: sequential estimation, prescribed mean-square accuracy, trigonometric regression, stopping time, autoregressive noise.

Introduction

There is a large literature devoted to the development of efficient methods for estimating parameters in a discrete time regression scheme with a signal modeled by a trigonometric polynomial [1]-[3]. This problem has been thoroughly investigated when the noise is a sequence of independent identically distributed random variables. The estimation problem of a regression with dependent noises having unknown spectral densities has not been solved yet and many questions still remain open.

There is a large literature devoted to the estimation of parameters in deterministic regression schemes. In application one of the most popular is the least squares method (LSM). The least squares estimates have been investigated in many papers and their properties are well understood for the regression models with known properties of noises. The estimation problem becomes less tractable if the noises are dependent with the spectral density depending on unknown parameters.

The paper [4] provides an example of estimating the mean in an autoregression of the first order with unknown autoregressive parameter, which shows that the LSM does not ensure a given accuracy for any fixed number of observations in the presence of the nuisance parameters. One of the ways to overcome these difficulties is to apply the sequential analysis approach.

In [5] a sequential procedure for estimation of periodic signal in autoregressive noise with unknown parameters is proposed. The procedure has good asymptotic properties and guarantees a specified mean-square estimation accuracy for any values of the nuisance parameters. This procedure, however, can be quite complex to implement in the case of many unknown parameters, since it comprises two stages and requires a system construction from a random number of LS estimates.

This paper considers the problem of estimating a periodic signal in a regression model with the autoregressive noise whose parameters are unknown. We propose a sequential sampling scheme with a special stopping time which ensures the estimation of unknown signal parameters with a prescribed mean square precision. The asymptotic formula for the mean duration of the procedure has been obtained.

1 Problem Formulation. Construction of the Sequential Procedure

Consider the problem of estimating the parameters μ_1, μ_2, β_{j1}, β_{j2}, $j = 1, ..., r$ of a signal

$$S_1 = \mu_1 + (-1)^n \mu_2 + \sum_{j=1}^{r} \beta_{j1} \cos \omega_j n + \beta_{j2} \sin \omega_j n \tag{1}$$

by observations of the process

$$x_n = S_n + \xi_n, \tag{2}$$

where ξ_n is the stable autoregressive process of order p obeying the equation

$$\xi_n = \lambda_1 \xi_{n-1} + ... + \lambda_p \xi_{n-p} + \epsilon_n. \tag{3}$$

Here $\{\epsilon_n\}$ is the sequence of independent identically distributed random variables, $E\epsilon_n = 0$, $E\epsilon_n^2 = \sigma^2$; $\lambda_1, ..., \lambda_p$ are the unknown parameters such that all roots of the characteristic polynomial

$$P(z) = z^p - \lambda z^{p-1} - ... - \lambda_p, \tag{4}$$

lie inside the unit circle. ω_j are known parameters such that $0 < \omega_j < \pi$, $\omega_i \neq \omega_j$, $i \neq j$.

It is well known [1] that any periodic signal with integer period T can be approximated by the trigonometric polynomial (1). In this case $r = \left[\frac{T-1}{2}\right]$, $\omega_j = \frac{2\pi j}{T}$, $[a]$ is the integer part of a number a.

In view of (1) and (3), the observed process (2) satisfies the equation

$$x_n = m_1 + (-1)^n m_2 + \textstyle\sum_{j=1}^{r} (\gamma_{j1} \cos \omega_j n + \gamma_{j2} \sin \omega_j n) + \sum_{k=1}^{p} \lambda_k x_{n-k} + \epsilon_n, n \geq p+1,$$

$$m_1 = \mu_1 \left(1 - \sum_{l=1}^{p} \lambda_l\right), m_2 = \mu_2 \left(1 - \sum_{l=1}^{p} (-1)^l \lambda_l\right),$$

$$\gamma_{j1} = \beta_{j1} (1 - \textstyle\sum_{l=1}^{p} \lambda_l \cos \omega_j l) + \beta_{j2} \sum_{l=1}^{p} \lambda_l \sin \omega_j l,$$

$$\gamma_{j2} = -\beta_{j1} \left(1 - \sum_{l=1}^{p} \lambda_l \sin \omega_j l\right) + \beta_{j2} \left(1 - \sum_{l=1}^{p} \lambda_l \cos \omega_j l\right).$$

Introducing the notation

$$Y_n = \begin{pmatrix} \Phi_n \\ X_{n-1} \end{pmatrix}, X_n = \begin{pmatrix} x_n \\ : \\ x_{n-p+1} \end{pmatrix}, \Phi_n = \begin{pmatrix} \phi_1(n) \\ : \\ \phi_{2r+2}(n) \end{pmatrix},$$

$$\phi_1(n) = 1, \phi_2(n) = (-1)^n, \phi_k(n) = \cos\omega_{k-2}n, \text{if } 3 \le k \le r+2,$$

$$\phi_k(n) = \sin\omega_{k-r-2}, \text{if } r+3 \le k \le 2r+2,$$

one can rewrite this equation as

$$X_n = \alpha' Y_n + \epsilon_n, n \ge p+1, \alpha \in \Lambda, \tag{5}$$

where $\alpha = (m_1, m_2, \gamma_{11}, \gamma_{12}, ..., \gamma_{r1}, \gamma_{r1})'$ is the column-vector of unknown parameters of size $(2r+2) \times 1$; Λ is some set of parameters meeting the requirements imposed on the parameters $\lambda_1, ..., \lambda_p$ in (3)-(4).

Least squares estimator (LSE) of the vector α of unknown parameters based on the observations of processes $X_k, Y_k, k = 1, ..., n$ in (5) is given by the formula

$$\alpha(n) = M_n^{-1} \sum_{k=1}^{n} Y_k x_K. \tag{6}$$

Here $M_n = \sum_{k=1}^n Y_k Y_k'$ is the sample Fisher information matrix of size $l \times l$,
$l = 2r + 2 + p$. We will assume that the minimum eigenvalue $\lambda_1(M_n)$ of matrix M_n satisfies the conditions $\lambda_1(M_n) \to \infty$ if $n \to \infty$ P-a.s.

In view of (5), it will be noted that, in general, the inverse matrix M_n^{-1} in (6) is random. Due to this fact, analysis of the LSE becomes problematic. To overcome these difficulties we propose a sequential LSE based on a special stopping rule. By making use of the upper bound for the bias norm of LSE obtained in [6], one gets

$$\|\alpha(n) - \alpha\|^2 \le \|M_n^{-2}\| \cdot \|m_n\|^2, \text{where } m_n = \textstyle\sum_{k=1}^n Y_k \epsilon_k.$$

We define the stopping rule as

$$\tau = \tau(h) = \inf\left\{ n \ge 1 : \left\|M_n^{-2}\right\|^{\frac{1}{2}} \le \frac{1}{h} \right\}, h > 0. \tag{7}$$

Sequential LSE $\alpha^*(h)$ for parameter α is given by the formula

$$\alpha^*(h) = \tilde{M}_{\tau(h)}^{-1} \sum_{k=p+1}^{\tau(h)} \beta_k Y_k x_k, \text{where } \tilde{M}_{\tau(h)}^{-1} = \sum_{k=p+1}^{\tau(h)} \beta_k Y_k Y_k', \beta_k = \begin{cases} 1 & \text{if } k < \tau(h), \\ \nu(h) & \text{if } k = \tau(h). \end{cases} \tag{8}$$

Here $\nu(h)$ is the factor, $0 < \nu(h) \le 1$, which is found from the equation

$$\left\| \left(\textstyle\sum_{k=1}^{\tau(h)-1} Y_k Y_k' + \nu(h) Y_{\tau(h)} Y_{\tau(h)}' \right)^{-2} \right\|^{\frac{1}{2}} = \tfrac{1}{h}.$$

2 Theoretical Properties of the Sequential Estimation Procedure

The properties of the sequential design (7)-(8) depend on the threshold value h. Further we assume that the vector of unknown parameters α in (5) belongs to some known compact subset of the set of the admissible values of the parameter vector. The properties of the analysis of the sequence plan (7)-(8), relating to the mean duration of procedure, and its accuracy are given in the following theorems.

Theorem 1. Let $(\epsilon_n)_{n\geq 1}$ be a sequence of independent identically distributed random varuables, $E\epsilon_n = 0$, $E\epsilon_n^2 = \sigma^2$, $E\epsilon_n^8 < \infty$, Λ be a set of admissible values of the parameter vector α. Then for any compact set $K \subset \Lambda$

$$\lim_{n\to\infty} \sup_{\alpha\in K} \left| E_\alpha \tfrac{\tau(h)}{h} - \|F^{-2}\|^{\frac{1}{2}} \right| = 0.$$

Here F is the limit matrix of the form

$$F = \left\| \begin{matrix} M_0 & M_1 \\ M_1 & F_0 + DM_0D \end{matrix} \right\|, \text{where } M_0 = diag\left(1, 1; \tfrac{1}{2}, ..., \tfrac{1}{2}\right),$$

$$V_k = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & (-1)^k & 0 & 0 \\ 0 & 0 & V_1(k) & V_2(k) \\ 0 & 0 & -V_2(k) & V_1(k) \end{pmatrix},$$

$$V_1(k) = diag\left(\cos\omega_1 k, ..., \cos\omega_r k\right), V_2(k) = diag\left(\sin\omega_1 k, ..., \sin\omega_r k\right),$$

$$A = \begin{pmatrix} \lambda_1 & ... & \lambda_p \\ I_{p-1} & ... & 0 \end{pmatrix}, \Gamma = \begin{pmatrix} m_1 & m_2 & \gamma_{11} & \gamma_{12} & \cdots & \gamma_{r1} & \gamma_{r2} \\ 0 & ... & ... & ... & ... & ... & 0 \\ 0 & ... & ... & ... & ... & ... & 0 \end{pmatrix},$$

$$D = \textstyle\sum_{k\geq 0} A^k \Gamma V(k), M_1 = M_0 V'(1) D', F_0 = \lim_{n\to\infty} \frac{1}{N} \sum_{n=p+1}^{N} \zeta_n \zeta_n' \text{ P} - \text{a.s.},$$

$$\zeta_n = A\zeta_{n-1} + \eta_n, \zeta_p = 0, \eta_n = (\epsilon_n, 0, ..., 0)'.$$

Theorem 2. For any compact set $K \subset \Lambda$ the mean-square accuracy of sequential plan (7)-(8) satisfies the inequality

$$\sup_{\alpha\in K} E_\alpha \left(\|\alpha_*(h) - \alpha\|^2\right) \leq \tfrac{b_k}{h} \left(1 + o(1)\right),$$

where $b_k = \sup_{\alpha\in K} \phi(\alpha), \phi(\alpha) = Q(\alpha) \|F^{-2}\|^{1/2}, o(1) \to 0$ as $h \to \infty$, the function $Q(\alpha)$ is defined by the equation $Q(\alpha) = trF_0 + 2r + 2 + (2r+2)^2 \left(\|D\|^2 + \frac{c}{1-q}\right)$.

Remark. According to Theorem 2 the mean square accuracy may be controlled by choosing a threshold h accounting for the number b_K can be calculated in advance. In this case the mean duration of the procedure grows linearly as h increases.

Conclusions

The paper proposes a sequential procedure for estimating the parameters of a trigonometric signal in a regression model with an autoregressive noise. A special stopping time based on the observed Fisher information is used to control the mean-square accuracy of the estimates of unknown parameters.

The results may be applied in the problems of control and identification of dynamic systems.

References

[1] Anderson T.W. (2011). *The statistical analysis of time series*. John Wiley & Sons, New York.

[2] Ibragimov I.A., Khasminsky R.Z. (1981). *Asymptotic estimation theory*. Springer-Verlag, New York.

[3] Liptser R., Shiryaev A.N. (2001). *Statistics of Random Processes*. Springer-Verlag, Berlin Heidelberg.

[4] Konev V., Pergamenshchikov S. (1997). On guaranteed estimation of the mean of an autoregressive process. *Ann. Statist.* Vol. **25**, pp. 2127-2163.

[5] Konev V.V., Pergamenshchikov S.M. (1997). Guaranteed estimation of a periodic signal distorted by an autoregressive noise with unknown parameters. *Probl. Peredachi Inf.* Vol. **33**, pp. 26-44. *(in Russian)*.

[6] Galtchouk L., Konev V. (2005). On sequential least squares estimates of autoregressive parameters. *Sequential Analysis*. Vol. **24**, pp. 335-364.

Oil Prices, Labor Force, and other Factors in the Forecasting Model of Russia's Economy Growth

Marina Lifshits
ISESP RAS, Moscow, Russia
e-mail: lifmarina@yandex.ru

Abstract

Adaptable for predictive calculations, econometrical models for Russia's economy growth (1999-2014) are offered in this article. The models take into consideration oil prices, demographic changes, world economy growth, foreign direct investment, and economy policy, and account for 90% of the variation of economic growth from 3Q 1999 to 1Q 2015. Forecasts of economic growth until 2024 are built, and this also includes predicative estimations for explanatory factors.

Keywords: Russia, GDP grow, GDP per capita grow, oil prices, labor force, foreign direct investment, economy policy.

Introduction

The article offers an econometrical model of economic growth in Russia (3Q 1999 – 1Q 2015) adaptable for predictive calculations. The model includes indicators of oil prices, demographic changes, world economy growth, foreign direct investment, and economy policy.

The impact of oil prices on economic growth in Russia is beyond doubt. This effect is described in detail in [6], and the authors believe that the model of economic growth should include both price changes and the absolute values of oil prices. Depending on changes in price are oil and gas revenues, and on the level of oil prices – the dynamics of investment in fixed assets. In most econometric models of economic growth in Russia, in [4; 9] in particular, change in the oil price is an important factor.

The dynamics of the labor force in the Russian Federation as a factor of economic growth has not yet received adequate attention in econometric modeling. The present work fills in this gap. It is possible that, with time, the demographic factor will become increasingly important.

Economic policy is reflected in the model by a factor of time, owing to a certain method of constructing this variable.

As an indicator of economic growth the percentage change in the volume of the Russian economy compared with the same quarter of the previous year is used.

Also, a forecast for economic growth in Russia until 2024 is built in the paper. To this end, projections for the independent variables are made too.

Data of Rosstat and World Bank are used.

1 The model, variables, interpretation of results

The proposed model of economic growth in Russia (3Q1999–1Q2015) has the form:

$$\Delta gdp\% = a_0 + a_1 * brent + a_2 * brent(t-1) + a_3 * \Delta brent\% + a_4 * gdpWgr + \\ a_5 * forinv + a_6 * \Delta empl\% + a_7 * toempl(t-4) + a_8 * trend + \varepsilon, \quad (1)$$

where $\Delta gdp\%$ – percentage change in the volume of the Russian economy compared to the same quarter of the previous year (SQPY);

$brent$ – the price of Brent crude oil in the quarter, dollars per barrel;
$brent(t-1)$ – the price of Brent crude oil in the previous quarter;
$\Delta brent\%$ – percentage change in the Brent price in comparison with SQPY;
$gdpWgr$ – global economic growth this year, %;
$forinv$ – foreign direct investment in the Russia's economy, % of GDP in the year;
$\Delta empl\%$ – percentage change in employment in the Russian economy compared with the SQPY;
$toempl(t-4)$ – the ratio of economy volume to the employee number in the SQRY;
$trend$ – reflects the economy policy quality, there are two variants in the models, $ln(d;t)$ in the Model 2 and *dummy* in the Model 3 ($d;t$ is set to 1 from 3Q 1999 to 1Q 2004, then 1 is added in each subsequent quarter, *dummy* is set to 0 from 3Q 1999 to 1Q 2004, then 1 in subsequent quarters);
ε – the residuals of the equation;
a_0 is the constant, a_n – coefficients given in Table 1.

Table 1: Models of economic growth in Russia

	Model 1		Model 2		Model 3	
	Coefficients	t-statistic	Coefficients	t-statistic	Coefficients	t-statistic
Constant	9.37****	4.132	1.85	0.747	2.58	1.123
$brent$	–	–	0.0143	0.620	–	–
$brent(t-1)$	0.0170	1.161	0.0486**	2.308	0.0212*	1.740
$\Delta brent\%$	0.0405****	5.115	0.0368****	4.774	0.0428****	6.497
$gdpWgr$	0.835****	3.507	1.159****	5.157	1.194****	5.699
$forinv$	0.167	0.721	0.347*	1.694	0.631***	2.977
$\Delta empl\%$	1.308****	4.571	0.856***	3.130	1.105****	4.590
$toempl(t-4)$	–0.0883****	–3.766	–0.0263	–1.107	–0.0297	–1.317
$ln(d;t)$	–	–	–2.166****	–4.636	–	–
dummy	–	–	–	–	–4.50****	–5.134
	$N = 63$ $R^2 = 0.861$ $AdjR^2 = 0.846$ $D - W = 1.258$		$N = 63$ $R^2 = 0.903$ $AdjR^2 = 0.888$ $D - W = 1.032$		$N = 63$ $R^2 = 0.906$ $AdjR^2 = 0.894$ $D - W = 1.322$	

*, **, ***, **** – significance level 10%, 5%, 1%, and 0.1% respectively

The constructed models confirm that the most important factor in the growth of the Russian economy is oil price. Only one variable $\Delta brent\%$ can account for 48% of the variation $\Delta gdp\%$. Also the result obtained previously Kazakova and Sinelnikov-Mourylev that the economic growth in Russia is under the impact of not only the change in oil prices, but also their absolute level has been confirmed. It is not the current level of prices, but prices in the previous quarter that are of significance.

Also justified has been the statement of Kudrin and Gurvich [5] that for Russia's economy not only oil prices, but the global economic dynamics as well are important.

When the relatively short time series are brought into focus, not infrequently, in order to get a more accurate assessment, various mathematical tools are applied, such as cointegration models. In this case, we chose a different path: a regression equation by the method of least squares including variables, reflecting the change in the Russian economic policy which began with the events related to the arrest of Khodorkovsky (25 October 2003). Perhaps even more important was not the arrest itself but the dismissal of Prime Minister Mikhail Kasyanov (Feb. 24, 2004), who earlier had repeatedly and publicly stated that the "Yukos case" harmed the investment attractiveness of Russia. Indeed, this history has demonstrated the world the insecurity of property rights in Russia. It is exactly from that time that such negative phenomena as corruption and the excessive role of government in the economy have begun to grow. The latter, Kudrin and Gurvich [5] described as follows: “The major impediment to growth is marked weakness of the market environment, explained primarily with dominance of state-owned and quasi-state companies.”

If we subtract the impact of changes in oil prices from the index of economic growth, we will see that the results of the economic policy pursued since 2004 have been growing less satisfactory (Fig. 1); there is a clear downward trend. A similar trend, but mirrored, has been laid out in the explanatory variable $ln(d;t)$, which combines the features of a dummy variable and the factor of time (Fig. 2).

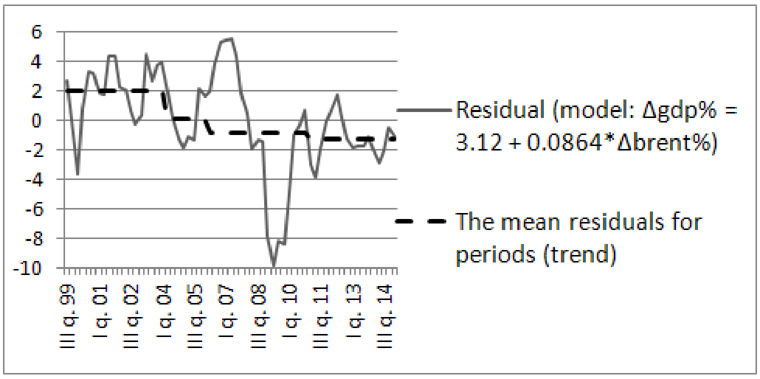


Figure 1: The residuals of the equation $\Delta gdp\% = 3.12 + 0.0864 * \Delta brent\%$ ($R^2 = 0.478$)

In fact, this variable reflecting the dominant trend has become a kind of framework that allowed the impact of other factors to reveal more precisely in the model 2.

The variable *trend* is missing in Model 1 and presented in a different form in Model 3.

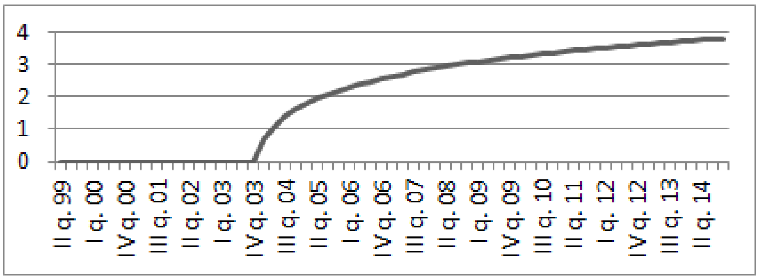


Figure 2: The values of the variable $ln(d;t)$

The values of Durbin-Watson statistics indicate the presence of autocorrelation in all three models. If you align the residuals of the Model 1 by using the moving average (Fig. 3), you will see two major waves of a descending line which can be explained by two political shocks that Russia's economy experienced in the period under review, "Yukos case" and War with Georgia. It is also possible that this is the result of some factors (foreign direct investment and oil prices) are undervalued in Model 1. The deterioration of Russia's image entails a reduction in the investment attractiveness of the country for international investors. That happened, in particular, in 2005. Variable $\Delta empl\%$ needs qualification. When unemployment is high, it is not the number of employees that has an impact on the economy but it is the economic growth that affects the change of employment. However, at low unemployment, decline in the labor force will be an additional factor limiting the opportunities for economic growth. According to Model 2, reducing the number of employees by 1% will be accompanied by a reduction of economic growth to 0.856 percentage points. In theory, reducing the number of employees could be more significantly offset by the increase in labor productivity. However, this requires investment. Therefore, a favorable investment climate is indispensable.

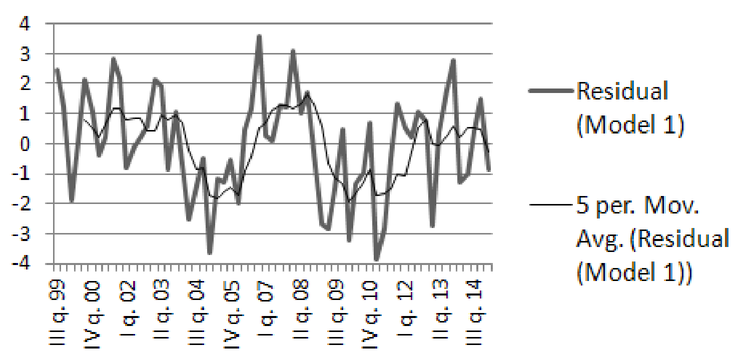


Figure 3: The residuals of Model 1

Variable $toempl(t-4)$ reflects the achieved level of economic development. Like in many other models of economic growth [2; 8; 10], the impact of this variable is negative. As has been said above, there is a close correlation between some variables in the period under review. It is possible that over time, when the random correlations weaken, the overall picture may be slightly different.

2 Projections for the explanatory variables and economic growth in Russia until 2024

Oil prices in winter 2014/2015 came as a surprise to many experts who had previously expected a slow decline, but not collapse. It is difficult to predict how fast and high prices could rise again. It is only evident that the previous $ 110 per barrel in the next 10 years will not be met, because the situation on the world market has changed dramatically. Now, in principle, there are no players who can raise and maintain prices at a high level as the price due to the development of new technologies will never be oligopoly. And the risk premium, which was part of the price in the 2007–2014 years, had disappeared forever, as the level of world production will no longer depend on the situation in the politically unstable regions. Thus, in the next few years the market will be looking tentatively for the equilibrium of the market price. To simplify, we can assume that in the period from March 2008 to September 2014 the average price ($ 96.7 per barrel Brent) was as high as consumers agreed to, and the future equilibrium price can be as low as it will be acceptable (breakeven) for a sufficient number of competing manufacturers. And it will depend solely on production technologies. Monograph [3] discusses in detail the pricing of oil, and also gives highly contradictory assessments of international experts as to what may be the equilibrium price in the foreseeable future. In paper [1] the bottom bracket payback for the extraction of shale oil in the US is estimated to be $35 in 2006 prices. The following options of economic growth forecasting in Russia are calculated under the assumption that the price of Brent will at first grow about the same path as in 2009, and then stop at the level of either $96 or $83 or $70 (Fig. 4).

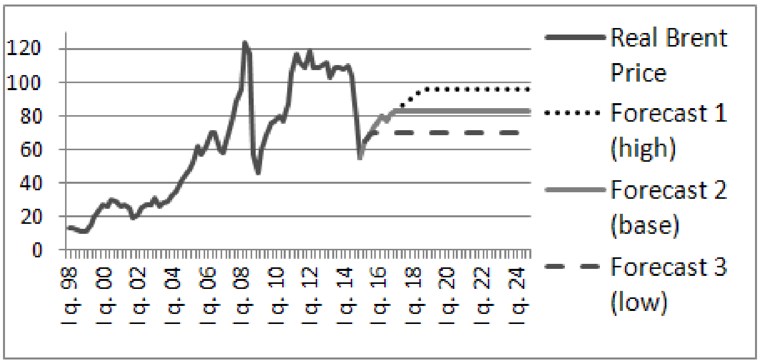


Figure 4: Real prices of Brent and 3 variants of the forecast of the writer, $ per barrel

World economic growth in all versions of our forecast is taken as 3.5% in 2015, then 3.75% per year (it was the average world economic growth from 1960 to 2014).

Reducing the number of **labor resources** in Russia is inevitable. This issue is discussed in detail in the writer's paper [7]. We can assume that the labor force is proportional to the size of the age group 18–64, as economic activity outside this age group is negligible. When calculating the forecast number of the age group 18–64 we accept the following conditions:

- the number of inhabitants of the Crimea is not taken into account;
- for the base case (forecast A) we take the coefficients of advancing age to be the same as the average in 2011–2013, l. e. a decrease of net migration will be fully offset by a decline in mortality;
- at lower forecast (B), age-specific death rates are the same as in 2010, and international migration will be zero.

As shown in Figure 5, the number of the age group 18–64 in both forecast variants will decline for years 2015–2024 significantly: by 9.2% for the base variant and 12.2% for the lowest.

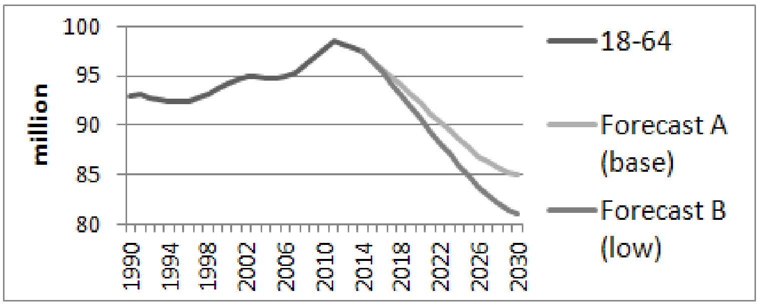


Figure 5: The population of the age group 18–64 at the beginning of the year. To 2014 – Rosstat data, from 2015 – designed by the writer

However, the number of employees does not depend solely on the size of the labor force, but also on the level of economic activity and unemployment. The level of economic activity of women, among other factors, depends on the level of fertility. We must bear in mind here that women are granted leave to care for a child up to the age of three. The number of **births** in the quarter *qborn* will be estimated through a simplified formula (2), the coefficients of which are found empirically for the period 1Q1999–1Q2015:

$$\Delta qborn2034_{(t;t-4)} = 0.259 + 0.0910 * \Delta qgdp.pc_{(t-3;t-7)}, \tag{2}$$

where $\Delta qborn2034_{(t;t-4)}$ – change in the number of births per 1,000 women aged 20-34 in comparison with SQPY;

$qgdp.pc_t$ – the ratio of GDP in the quarter t to the population (thous. rubles in 2008 prices);

$\Delta qgdp.pc_{(t-3;t-7)}$ – the increase in the value of $qgdp.pc_t$ in the quarter in which a decision about the birth of a child is made, in comparison with SQPY.

Note that in the 1999–2013 years, from 79.3 to 83.3% of all children were born to women in the age group 20–34.

For the level of **foreign direct investment** we make the following assumptions:

- for the high variant (a), we assume that $forinv = 0$ in 2015, and 3.37% in the following years (this was the meaning of this variable in 2013);
- for the base case (b) we assume that $forinv = 0$ in all years; Such a situation may arise if the Minsk Agreement will not be implemented.

The results of calculations for the 36 variants of the forecast are shown in Figure 6.

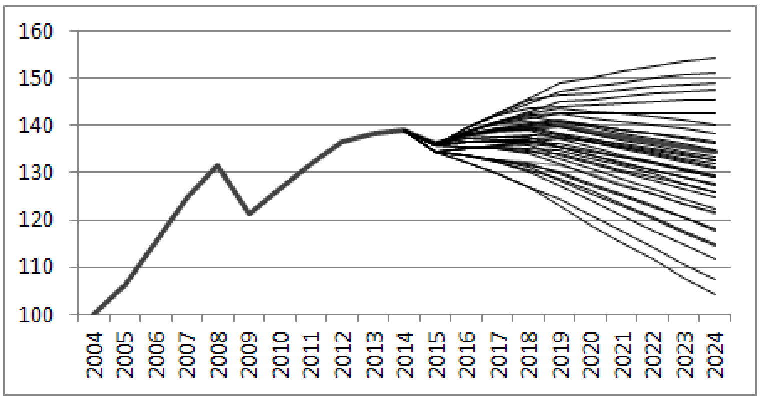


Figure 6: 36 variants of the Russian economy forecast, GDP (2004) = 100%

Conclusions

The paper suggests models for a plausible forecast of economic and demographic development of Russia in mutual relationship. The models of economy growth contain a variable that reflects the overall negative impact of the policy pursued by the central government since 2004. If the policy does not change the prospects for the Russian economy are as following – in the most favorable and, unlikely, scenario (1Aa, Model 3), the population will decline by 1.0 million people for 10 years and the GDP will grow by 10.9%; in the worst scenario (3Bb, Model 2), the population will decrease by 6.2 million people, and the GDP will decrease by 25%. To improve the model we must bind not only fertility, but also migration and mortality, to the level of economic development. In addition, the variable reflecting the level of demographic load needs improvement.

References

[1] Aguilera R.F., Eggert R.G., Lagos C.C. (2009). Depletion and the Future Availability of Petroleum Resources. *Energy Journal*. Vol. **30**, pp. 141-174.

[2] Barro R.J., Sala-i-Martin X. (2003). *Economic Growth, Second Edition*. The MIT Press, London.

[3] Bushuev V.V., Konoplanik A.A., Mirkin Y.M. (2013). *Tseny na neft': analiz, tendentsii, prognoz. [Oil Prices: Analysis, Trends, Forecast]*. ID "Energiya", Moscow.

[4] Ito K. (2008). Oil Price and Macroeconomy in Russia. *Economics Bulletin*. Vol. **17**, No. **17**, pp. 1-9.

[5] Kudrin A., Gurvich E. (2014). A New Growth Model for the Russian Economy. [in Rus]. *Voprosy ekonomiki*. Vol. **12**, pp. 4-36.

[6] Kazakova M., Sinelnikov-Mourylev S. (2009). The Global Market for Energy Sources and the Pace of Economic Growth in Russia. [in Rus]. *Ekonomicheskaya politika*. Vol. **5**, pp. 7-37.

[7] Lifshits M. (2012). Global regularities and Russia's prospects: Could natural loss of manpower be completely replenished by international migration in 2012–2028? *Demographic development: challenges of globalization (The seventh Valenteevskiye Chteniya): International Conference: Moscow, Russia, 15-17 November 2012: Proceedings*. MAKS Press, Moscow.

[8] Lifshits M. (2013). The influence of migration and natural reproduction of labor force upon economic growth in the countries of the world. [in Rus]. *Applied Econometrics*. Vol. **3**, pp. 32-51.

[9] Rautava J. (2013). Oil Prices, Excess Uncertainty and Trend Growth: A Forecasting Model for Russia's Economy. *Focus on European economic integration*. Q. **4**, pp. 77-87.

[10] Sachs J., Warner A. (1997). Fundamental sources of long-run growth. *American Economic Review*. Vol. **87 (2)**, pp. 184-187.

Creating Composite Measures for assessment conditions of fiscal potential mobilization

Loseva A. V.
Novosibirsk State University, Novosibirsk, Russia
e-mail: lav78@yandex.ru

Abstract

The study employs an algorithm for creating a composite measure which allows to rank regions according to their fiscal characteristics. Using the method proposed fiscal territorial disparities in Siberian Federal District has been analyzed.

Keywords: fiscal potentials, revenue potentials, regional fiscal independency, composite measure, ranking.

Introduction

To reach optimal decisions about governing budget process at sub-national (regional) level requires tools for measure and comparison regional fiscal capacity. We propose an algorithm for creating a composite measure which allows to rank regions according to their fiscal characteristics. In this case regional fiscal capacity is considered in the context of conditions of revenue potential mobilization. It takes a set of indicators which is different from previous studies in the literature. We propose 29 indicators divided into three units according different aspects of considered issue. Various alternative methods of normalization and aggregating creating a composite measure were used.

The results of analysis have allowed to find the following: to emphasize the regions-leaders and the regions-outsiders in Siberian Federal District; to show how much do the conditions of revenue potential mobilization vary across regions; to stress regional advantages and disadvantages. The source data came from the official statistics for 2011 - 2012 yr.

1 Review

Studies with inter-regional comparisons often include a process of aggregating information, the convolution of a set of selected data in one or some integral indicators and obtaining multidimensional summative measure.

There is a wide range of methodological approaches to composite measures divided on two distinct branches:

- a formal approach, when designing an integrated indicator is obtained by the methods of factor analysis, principal components etc.;

- an approach based on semantic analysis of the problem, when integrated indicators are the result of various transformations of the selected indicators and weights in additive or multiplicative convolution forms are set by experts or based on intuitive or qualitative reasoning of the researcher. The second approach “contains a certain degree of subjectivity and can often lead to different results when analyzing the same problem” [4]. However, because of its simplicity, compared to the first one, it is rather popular in cases of decisions-makers analysis. Some researchers supplement and adjust the methods of second approach by procedures of statistical analysis [1].

The International Statistics to date has accumulated a significant experience in constructing integral measures which compare and rank country performance in areas such as health, quality of life, industrial competitiveness, corruption, sustainable development, globalization, innovation and etc. Composite indicators are developed and used by the international organizations, such as the HDI - Human Development Index (UN), TAI - Technology Achievement Index (OECD). Their feature is the ease of calculation and the use of indicators that are available and comparable for a wide set of countries.

In Russia integral estimates for inter-regional comparisons are widely used in the methods of distribution of funds within the framework of intergovernmental relations, as well as in the work of individual researchers on the different problems of the regional socio-economic development. A general algorithm for building a composite indicator is similar in all the considered methods and include the following steps:

- data selection;
- normalization;
- integration of normalized data.

A number of normalization and integration methods exist. The objective is to identify the most suitable procedures to apply. For example, different normalisation methods will produce different results for the composite indicator. Therefore, overall robustness tests are recommended to carry out to assess their impact on the outcomes [3].

2 Data normalization

The indicators in a data set often have different measurement units. Therefore normalization is required in most cases of data aggregation. The most widespread methods are the followings.

1. Distance to a reference measures the relative position of a given individual indicator (x_i) vis-a-vis a reference point. The reference region could be the average or highest region of the group:

$$x_n = \frac{x_i}{\bar{x}} \tag{1}$$

or

$$x_n = \frac{x_i}{x_{max}} \tag{2}$$

Thus, if the excess value of the indicator over the other is seen in a negative light, normalization is performed in reverse order:

$$x_n = \frac{\bar{x}}{x_i} \tag{3}$$

or

$$x_n = \frac{x_{max}}{x_i} \tag{4}$$

2. Standardisation (or z-scores) converts indicators to a common scale with a mean of zero and standard deviation (σ) of one:

$$x_n = \frac{x_i - \bar{x}}{\sigma} \tag{5}$$

3. Min-Max normalisation "could widen the range of indicators lying within a small interval, increasing the effect on the composite indicator more than the z-score transformation" [3]:

$$x_n = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{6}$$

Normalisation of indicators using the average, maximum and minimum values prevalent in the practice in analysis for public administration decision-makers.

3 Weighting and aggregation

Existing techniques may use an additive or a multiplicative approach of aggregation.

There is a widespread use for this purpose different formulas of average. For example, when building Sectoral innovation index [2].used a simple arithmetic average. In the methods comprising the step of assigning indicators weights coefficients of significance with its attendant procedures (for example, the assessment of differences of expert opinion), applies the weighted arithmetic mean, weighted averages of standardized indicator values. When constructing the Gender-related Development Index integration occurs in the form of average weighted harmonic, where values are weighted for the proportion of men and women in the total population of the country. Researchers that do not use the opinions of experts and give the equal weights often use geometric average formula.

Reviewing the experience of constricting composite measures for the analysis of territorial disparities we offer a simple approach, to-can be used to get detailed operational information for estimating, comparing and monitoring the level and features of regional fiscal capacity.

4 The indicators

The general objective of the selected indicators is to reflect conditions of revenue potential mobilization in a region (table 1).

Table 1: Selected indicators for inter-regional comparison of fiscal capacity and conditions of revenue potential mobilization

Unit	Group	Indicators
I Revenue	1. Level of fiscal capacity, level of local budgets revenues	X_1-X_3
sufficiency for	2. Region disparities in fiscal capacity	X_4-X_5
public goods	3. Proportion of tax potential to public expenditure	X_6
and services	4. Contribution of regional taxes	X_7-X_8
	5. Fiscal discipline and tax evasion	X_9-X_{12}
II Tax potential	6. Corporate financial capacity	X_{13}-X_{15}
exhaustiveness	7. Households financial capacity	X_{16}-X_{17}
	8. Performance of tax administration	X_{18}-X_{20}
III Tax burden	9. Tax burden level	X_{21}-X_{22}
and investment	10. Investment activity	X_{23}-X_{29}

5 The territorial disparities between the regions of the Siberian Federal district

Within each of the three units, for each of the individual indicators, key features of distribution were calculated: maximum and minimum characteristic values, average, median, the variation coefficient, asymmetry factor. The result shew territorial disparities appear to be generally larger than initially expected - a number of parameters differ significantly.

The high level of variation appears with the indicators that are indirectly associated with the development of the regional economy and the financial status of the taxpayer. Extremely strong differences in the amount of profit per organization may be due to high variation in the amounts of tax evasion. A marked degree of regional differences evident in the indicator of the size of the shortfall in income that may indicate significant regional differences tax policy selected regions of the Russian Federation. It is noticeable that the Siberian Federal district unprofitable different from the national average indicators of overdue payables of enterprises and the share of population with incomes below the subsistence minimum.

A moderate homogeneity of the regions appears in total tax burden - a share of taxes and fees in gross regional product. High level of variation in other, more particular indicators of the tax burden (25, 26, 28, 29) suggest that differences in the tax environment in selected regions. The high variation in the degree of tax revenues disparities by activity indicates significant difference Siberian regions on the tax potential economic structure.

6 Methods of aggregation

For the creating an integral indicator three approaches was used: options A, B and C to investigate their comparability. Under option A the normalization of the indicators was carried out by comparing them with the average. In this case we used a geometric aggregation (also called deprivational index):

$$x_j = \sqrt[k]{\prod_{i=1}^{k} x_{ni}} \tag{7}$$

where X_j is an intermediate integral indicator on a unit j, k is the number of indicators that characterize a unit j.

The option B uses standardization (z-score) method to normalize the values of the indicators.

Intermediate aggregation in this case is proposed by summing the positive and negative deviations and averaging the values obtained:

$$X_j = \pm \sum_{i=1}^{n} \frac{x_{ni}}{k}$$

Option C provides the data normalization by the Min-Max method. Intermediate aggregation was carried out by using the arithmetic average.

Methods of normalization and intermediate aggregation denote the way of the further data aggregation. Method B is not possible to assign different weights to the integral estimates X_j, and the final aggregated indicator (I_B) will be defined as the distance:

$$I_B = \sum_{j=1}^{3} X_j$$

When using the normalized values in form of distance to the average or Min-Max values (options A and C) final aggregated indicator can be calculated as a summation of weighted intermediate aggregation values (the most widespread method of linear aggregation):

$$I_{A,C} = X_1 * 0.2 + x_2 * 0.5 + X_3 * 0.3$$

when the weights reflect the ability of regional authorities to influence the situation in short-time term.

7 The results of intermediate and final aggregations

Table 2 shows the intermediate aggregated indicators, generated for each unit of individual indicators by three options (A, B, C).

Table 2: Aggregated value of different aspects of conditions of revenue potential mobilization, 2011

	Aggregated indicators for Unit I Revenue sufficiency for public goods and services X_1			Aggregated indicators for Unit II Tax potential exhaustiveness X_2			Aggregated indicators for Unit III Tax burden and investment X_3		
	A	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*
Republic of Altay	0.957	0.098	0.515	0.855	0.110	1.502	1.452	0.322	0.972
Republic of Buryatia	0.075	-0.018	0.490	1.287	0.218	1.542	1.339	0.408	1.006
Republic of Tyva	0.917	-0.446	0.368	1.638	-0.505	1.192	2.002	0.666	1.070
Republic of Khakasia	1.154	0.115	0.534	0.937	0.069	1.398	1.400	0.581	1.059
Altay Territory	1.011	-0.009	0.494	0.982	0.037	1.475	1.321	0.221	0.952
Zabaikalsk Territory	1.174	0.326	0.590	1.100	-0.068	1.447	1.307	-0.245	0.840
Krasnoyarsk Territory	0.874	0.056	0.518	1.199	0.362	1.559	1.389	0.324	1.005
Irkutsk Region	1.016	0.073	0.528	1.256	0.507	1.592	1.281	-0.037	0.888
Kemerovo Region	1.049	0.280	0.580	1.208	0.111	1.475	1.305	-0.251	0.851
Novosibirsk Region	0.957	0.100	0.534	0.972	0.112	1.430	1.286	-0.032	0.891
Omsk Region	0.953	-0.012	0.498	1.093	0.098	1.488	1.290	-0.626	0.713
Tomsk Region	0.863	-0.556	0.345	1.366	0.606	1.632	1.250	-1.236	0.599

Regions were ranked according the levels of their aggregated indicators (Table 3).

The results of intermediate aggregation and ranking allow to estimate the position of the Siberian regions within each condition group both in relation to each other and in dynamics. According to the criteria of “Revenue sufficiency for public goods and services” first places were taken by Zabaikalsk Territory, Kemerovo Region and Republic of Khakasia. Their advantage - a very high ratio of local budgets incomes of the total revenues. The Siberian regions that took the last places are Tomsk, Omsk regions and Republic of Tyva. The latest one is characterized by an extremely low degree of budgetary provision and, in general, unsatisfactory evaluations almost all aspects of tax and budget process. The Omsk and Tomsk regions, in spite of its significant tax potential, are at a disadvantage: they have a high degree of centralization of regional tax revenues and a high degree of concentration of tax capacity in certain types of economic activity.

The leading regions in terms of “Tax potential exhaustiveness” are Republic of Buryatia, Tomsk and Irkutsk regions. They took first places thanks to the level of tax discipline and performance of tax administration as well as the relatively low level of corporate overdue debt. Absolute outsider in this case is Novosibirsk region, in addition it is the “leader” according the size of tax evasion and the minimum level of

Table 3: Intermediate ranking of regions (2011)

	Ranks for Unit I Revenue sufficiency for public goods and services				Ranks for Unit II Tax potential exhaustiveness				Ranks for Unit III Tax burden and investment			
	A	B	C	R_m	A	B	C	R_m	A	B	C	R_m
Republic of Altay	7	5	7	6	11	11	5	9.2	2	5	5	4.4
Republic of Buryatia	3	10	10	8.6	2	1	4	2.1	5	3	3	3.4
Republic of Tyva	10	11	11	10.8	12	8	12	10	1	1	1	1
Republic of Khakasia	2	3	3	2.8	10	5	11	7.8	3	2	2	2.2
Altay Territory	6	8	9	7.9	8	6	7	6.7	6	6	6	6
Zabaikalsk Territory	1	1	1	1	6	7	9	7.4	7	9	10	8.9
Krasnoyarsk Territory	11	7	6	7.5	5	2	3	2.9	4	4	4	4
Irkutsk Region	5	6	5	5.5	3	3	2	2.7	11	8	8	8.6
Kemerovo Region	4	2	2	2.4	4	10	8	8.2	8	10	9	9.3
Novosibirsk Region	8	4	4	4.8	9	9	10	9.3	10	7	7	7.6
Omsk Region	9	9	8	8.7	7	12	6	9.2	9	11	11	10.6
Tomsk Region	12	12	12	12	1	4	1	2.5	12	12	12	12

average profit per organization in the region.

Tables 2 and 3 show a highest level of Tax burden in Kemerovo, Omsk and Tomsk regions. It is a kind of payment for their leading position in terms of total amounts of tax revenues which, in large part, are not for their own regional budget, but for further distribution at the Federal level.

Further the intermediate aggregated indicators were weighted and combined into a final composite variable scores I_B and $I_{A,C}$ which were converted in ranks .

The distribution of Siberian regions, respectively the composite measure of the conditions for the mobilization of tax potential obtained by averaging ranks is presented in the Table 4.

Conclusions

The results of methods suggest that some aspects of the conditions for the mobilization of revenue capacity appear differently in the Siberian regions. We find two types of regions: 1) the regions with stable ranking position in time (the Republic of Tyva, Republic of Khakassia, Omsk, Tomsk region), 2) the regions which radically change their position (Irkutsk and Kemerovo Regions). It is concluded that to date in considered context the Krasnoyarsk territory is the undisputed leader of the Siberian Federal district. The Republic of Tyva that can really be designate as subsidized region by its nature. Other Siberian regions have a number of advantages and disadvantages that should be analyzed separately for each subject of the Russian Federation, to identify the best ways to improve the situation and achieve the highest possible level of fiscal independence.

The method proposed provides an opportunity to improve the analysis of territo-

Table 4: Rating of Siberian regions, respectively the level of conditions for the mobilization of tax potential (2011 – 2012 yr.)

	R_m rank Average between options A, B, C		General rating	
	2011	2012	2011	2012
Republic of Buryatia	2.0	4.0	1	2
Krasnoyarsk Territory	2.7	5.0	2	4
Irkutsk Region	3.3	10.3	3	12
Republic of Khakasia	4.7	4.7	4	3
Kemerovo Region	6.3	1.3	5	1
Republic of Altay	7.0	8.3	6	9
Altay Territory	7.3	5.3	7	5
Zabaikalsk Territory	7.7	6.0	8	6
Tomsk Region	8.0	10.0	9	11
Novosibirsk Region	9.0	7.3	10	8
Omsk Region	9.7	9.7	11	10
Republic of Tyva	10.3	6.0	12	7

rial fiscal disparities. The results of these studies can be directly applied in subnational governments activity.

Acknowledgements

This research has been supported by the Strategic development program of the Novosibirsk State Technical University (project 2.3.1), 2014 yr.

References

[1] Chudilin G.I., Geniatulina K.V. (2005). Variability of multidimensional assessments of the economic development of municipalities. *Voprosy Statistiki.* Vol. **12**, pp. 38-43 (In Russian).

[2] Makarova I.A., Flood I.A. (2008).Statistical evaluation of innovative development. *Voprosy Statistiki.* Vol. **2**, pp. 15-29.(In Russian).

[3] OECD. (2008). *Handbook on constructing composite Indicators: methodology and user guide. ISBN 978-92-64-04345-9.* OECD Publications, Paris.

[4] Smagin B.I., Neujmin S.K. (2005). Nature and technique of definition of region development indicators. *Voprosy Statistiki.* Vol. **12**, pp. 19-23.(In Russian).

System Analysis Principles for Forecasting Financial Processes

O. A. Kozhukhivska[1], P. I. Bidyuk[2] and A. D. Kozhukhivskyi[1]

[1] *Cherkassy state technological university, Cherkassy, Ukraine*

[2] *Institute of Applied System Analysis of National technical university of Ukraine «Kyiv Polytechnic Institute»*

e-mail: olga-kozhuhovska@mail.ru, pbidyuke@gmail.com, andrejdk@mail.ru

Abstract

A computer based decision support system is proposed the basic tasks of which are adaptive model constructing and forecasting of financial and economic processes. The system is developed with the use of system analysis principles, i.e. the possibility for taking into consideration of some stochastic and information uncertainties, forming alternatives for models and forecasts, and tracking of the computing procedures correctness during all stages of data processing. A modular architecture is implemented that provides a possibility for the further enhancement and modification of the system functional possibilities with new forecasting and parameter estimation techniques. A high quality of final result is achieved thanks to appropriate tracking of the computing procedures at all stages of data processing: preliminary data processing, model constructing, and forecasts estimation. The tracking is performed with appropriate set of statistical quality parameters. Examples are given for modeling and forecasting of nonlinear and nonstationary financial and economic processes. The examples show that the system developed has good perspectives for the practical use. It is supposed that the system will find its applications as an extra tool for decision making when developing the strategies for enterprises of various types.

Keywords: model, forecasting, financial and economic processes, system analysis principles.

Introduction

The forecasting problems are to be solved practically in all areas of human activities. However, the problems of mathematical modeling, estimation and forecasting process dynamics are particularly urgent for micro- and macroeconomics, banking sphere, insurance, investment companies, industrial enterprises that are functioning in conditions of tough competition, and many others kind of activities. There are many ideologically different approaches to mathematical description of processes dynamics and their volatility that are based on known statistical and recently developed techniques of intellectual data analysis.

Volatility forecasts are used widely as a measure of various kinds of financial risks in the frames of Value-at-Risk (VaR) and other methodologies. The market and some other types of risks are estimated with different modifications of VaR methodology that provides a possibility to reach practically acceptable quality of risk estimates [1, 2]. For forecasting financial processes and enterprises bankruptcy the nonlinear classification type models have found wide application, for example, logistic regression as well as support vector machine (SVM), fuzzy logic, neural networks and neuro-fuzzy techniques, Bayesian networks, decision trees and combinations of the approaches mentioned [3 - 5].

Selection and application of a specific technique for process description and forecasts estimation depends on application area, availability of statistical data, qualification of personnel, who work on the financial analysis problems, and availability of appropriate applied software. Better results for estimation of financial processes forecasts is usually reached with application of ideologically different techniques combined in the frames of one computer based system. Such approach to solving the problems of quality forecasts estimation can be implemented in the frames of modern decision support systems (DSS). DSS is a powerful instrument for supporting decision making as far as it combines a set of appropriately selected data processing procedures aiming to reach final result of high quality – objective high quality alternatives for a decision maker. Development of a DSS is based on modern theories of system analysis, information processing systems, estimation theory, mathematical and statistical modeling and forecasting, decision making theory as well as many other results of theory and practice of processing data and expert estimates [6, 7].

The paper considers the problem of DSS constructing for solving the problems of modeling and forecasting processes evolution in time with the possibility for application of alternative data processing techniques, modeling and estimation of parameters and states for the processes under study.

1 Problem formulation

The purpose of the study is as follows: 1) analysis and development of requirements to the modern decision support systems; 2) development of the system architecture; 4) selection of mathematical modeling and forecasting techniques for financial and economic processes; 3) illustration of the system application to solving the problem of financial and economic processes forecasting with statistical data.

2 General requirements to modern DSS

Such systems should satisfy the following general requirements: 1) - contain highly developed bases of data, mathematical models, quality criteria and rules, as well as necessary computational procedures; 2) - their interface should be user friendly, convenient and simple for use, as well as adaptive for the users of various levels (e.g., engineering and managerial staff); 3) - the hierarchy of a system functioning should correspond to the hierarchic process of human decision making; 4) - the system should possess an ability for learning in the process of its functioning, i.e. accumulate appropriate knowledge regarding possibilities of solving the problems of definite (selected) class; 5) - the organization and techniques for computing procedures should provide for appropriate rate of computing that corresponds to the human requirements with regard to the rate of alternatives generation and reaching the final result; 6) - computing (precision) quality should satisfy preliminary established requirements; 7) - intermediate and final results of computations should be controlled with appropriate sets of analytic quality criteria, what will enhance significantly quality and reliability of the final result; 8) - DSS should generate all necessary for a user forms and types of intermediate and final results representations with taking into consideration the users of various levels; 9) - the system should contain the means for exchange with data and knowledge with other information processing systems via local or global computer nets; 10) - DSS should be easily expandable with new functions.

Satisfaction of all the requirements mentioned above provides a possibility for effective practical application of a system developed and enhancing general behavioristic effect of the DSS as a whole for a specific company or an enterprise.

3 General and special purpose mathematical tools for DSS

All mathematical methods that are hired for development and implementation of DSS could be divided into two following groups: - general purpose methods that provide for implementation of system functions, and special purpose methods that are necessary for solving specific problems regarding data processing, model constructing, alternatives generating, selecting the best alternative for implementation and forecasting of the implementation consequences.

The group of the general purpose methods includes the following methods: - data and knowledge collecting and editing procedures; - preliminary data processing techniques such as digital filtering, normalization, imputation of missing values, detecting special effects (regime switching, seasonal effects, trends etc); - the methods for accumulating information regarding previous applications of DSS to problem solving for the retrospective use; - computer graphics techniques; - techniques for syntactic analysis to be used in command interpreter; - methods for organizing communications with other information processing systems via local and global nets; - logical rules to control the system functioning. The set of the methods mentioned could be modified or expanded depending on specific application.

Selection of the application defined mathematical methods for a DSS depends on the specific system application area, possible problem statements regarding data processing, model building, processes forecasting, and alternatives generation. However, it is possible to state that in most cases of DSS development it is necessary to use the following mathematical methods: - methods and methodologies for mathematical (statistical and probabilistic) modeling using statistical/experimental data; - forecasting techniques on the basis of the models constructed with possibilities for combining the forecasts computed with different techniques; - operations research optimization techniques and dynamic optimization (optimal control) methods; - the methods for forecasting/foresight of decision implementation consequences; - the sets of special analytic criteria to control the processes of computations performed at each stage of data processing and alternative generation aiming to reach high quality of final results.

All the methods and methodologies mentioned are described well in special modern literature. For example, time series modeling and forecasting are presented in many references, more particularly in [8, 9]. The task for a DSS developer is in appropriate selection of model classes, modeling and optimization techniques, quality criteria as well as relevant methodologies for organizing computational processes.

4 General and special purpose mathematical tools for DSS

Decision making process includes rather sophisticated procedures that could be partially or completely iterative, i.e. executed repeatedly when the alternative found is not satisfactory for a decision making person (DMP). DSS can return automatically (or on DMP initiative) to the previous stages of data and knowledge analysis.

The whole process of making and implementing decision could be considered as consisting of the stages given below.

1. A thorough analysis of the decision problem using all available sources of information, collection of data and knowledge relevant to the problem. At this stage it is also important to consider and use former solutions to the problem if such are available. The information regarding former solutions of similar problem can be helpful for correcting problem statement, to select appropriate techniques for data analysis, to speed up alternative generation, and to decline the alternatives that turned out to be ineffective in the past.

2. Selection of a class (classes) of mathematical models for the problem description, and analysis of a possibility for the use of available (previously developed) models. The models could belong to different classes as far as they can be formulated in continuous or discrete time, be linear or nonlinear, they can be developed according to the structural or functional approach etc. In some cases it is necessary to construct complex simulative model that would include a set of simpler models of various classes.

3. Development of new models for the problem (process, object, system) under study what includes structure and parameter estimation for candidate models using available data (and possibly expert estimates) and knowledge of various types. The alternative structures of candidate models provide a possibility for selecting the best one of them for generating alternative decisions (forecasts, control actions, risk estimates etc) on their bases.

4. Analysis of the candidate models constructed and selecting of the best one of them with application of a set of statistical quality criteria and expert opinion (estimation). At this stage again more than one model can be selected for

the further use as far as the best model (for a particular application) can be found only after application of the candidates for solving particular problem, i.e. after alternatives generating and estimating possible consequences of their implementation.

5. Application of the model (models) selected to solving forecasting and/or control problem (when necessary). If the forecasts or controls computed are not satisfactory we should return back to stage one or stage three, and repeat the process of model constructing. At this stage another set of statistical quality criteria should be applied to the analysis of forecasts or controls determined.

6. Generating of a set of alternatives with the use of the model (models) constructed and various admissible initial conditions and constraints on variables. In a case of controls generating the alternatives could be built with different optimality criteria, utility functions or other criteria.

7. Analysis of the alternatives generated with the experts of an enterprise or a company, and final selection of the best one for practical implementation. In a case when no alternative is acceptable we should return back to the model constructing or alternative generating stages. New knowledge or data can be required for the next iteration of computing the decision support.

8. Planning of actions and estimation of financial, material and human resources that are necessary for implementation of the alternative selected. Determining of a time horizon (horizon of control) necessary for implementing the decision made.

9. Implementation of the decision made: current monitoring of availability and spending the necessary resources, estimation of necessary time frames, registering and quality estimation of intermediate and final results.

10. Application of possible analytic and expert quality criteria to estimation of final results.

11. Analysis of the final results achieved by the company experts, and final elucidation of advantages and disadvantages of the alternative implemented; analysis of the decision making and implementing process, and forming forecasts (foresights) for the future.

12. Writing the final report on the tasks performed.

5 Architecture of DSS for forecasting of financial and economic processes

DSS architecture is a generalized large-scale representation of system elements with links between them. Architecture gives a notion for the general purpose of system constructing and its basic functions. DSS functionality is controlled by user commands, correctness of which is monitored by the command interpreter which constitutes a part of user interface. The user commands are implemented by the main operation module that coordinates functioning of all system elements. The specific commands and actions can be the following: expanding and modification of bases available in the system; initiation and starting of data and knowledge processing procedures; model constructing, forecasts estimation and alternative generating; viewing intermediate and final results of computing; retrospective analysis of previous results of decision making; comparing of current results with previous.

The system interface is considered as its basic element from the point of view of its presentation to the user. This is justified by the fact that interface construction influences substantially convenience as well as rate and effectiveness of user interaction with the system [10]. The principles of interface constructing and its implementation create a separate special task that is not considered here.

6 Coping with uncertainties

The problem of identifying and taking into consideration various uncertainties is a very important one and is practically always present in decision making procedures. Here we do not consider the cases when uncertainties are taken into consideration with expert estimates. Though expert estimates are not excluded from the process of alternative generation. For example, expert estimate can be used for selecting special types of mathematical models that for some reason cannot be chosen automatically due to sophisticated structure or necessity to apply special estimation procedures. Expert judgment can also be useful for final selection of the best alternative from the set of generated decisions.

Statistical data uncertainties such as skipped measurements, extreme values and high level jumps of unknown origin could be processed with appropriately selected statistical procedures. There exist a number of data imputation procedures that help to make collected data complete. For example, very often skipped measurements for time series can be generated with appropriately selected distributions. Processing of jumps and extreme values helps with adjusting data stationarity and to estimate correctly probability distribution.

Application of Kalman filter requires knowledge of covariance for state disturbances and measurement errors. As far as these parameters are often unknown it is useful to apply appropriate adaptive estimation algorithms that provide acceptable estimates for the statistical parameters. An experience of practical application of the filter shows that it better to use separate procedure for covariance estimation to avoid divergence of filtering algorithm.

Fuzzy logic can be hired for coping with the level uncertainties for variables when we consider linguistic variables instead of numerical ones. There are possibilities for transforming fuzzy values into numerical and vice versa. Thus, there is no problem for processing fuzzy and exact variables in the frames of one computing procedure.

Probabilistic types of uncertainties regarding whether or not some event will happen can be taken into consideration with various probabilistic models. Among them are analysis of distributions, Bayesian networks and other possibilities. From the computational point of view it is easier to process discrete variables as far as they accept a final number of values. In this case probabilities are assigned to outcomes using a probability mass function (PMF). Mass function tells us what “weight” (or a mass) should be assigned to each outcome. The sum for all the masses is 1,0. In a case of dealing with continuous variables, that may accept infinite number of values, we use probability density function (PDF). An integral over the density function should be equal to 1,0. When more than one random variable is considered we have to use joint distribution functions.

Generally speaking the modern instrumentation for coping with uncertainties is very powerful and it should be used in the frames of decision support systems for enhancing their possibilities with respect to reaching the best models and forecasts, and the best possible decisions.

7 Data, model and forecasts quality criteria

To achieve reliable high quality final result of forecasting at each stage of computational hierarchy separate sets of statistical quality criteria have been used. Data quality control is performed with the following criteria:

- database analysis for missing values using developed logical rules, and imputation of missed values with appropriate techniques;
- analysis of data for availability of outliers with special statistical tests, and processing of outliers to reduce their negative influence on statistical properties of data;
- normalizing of data in a case of necessity;
- application of low-order digital filters (usually that’s low-pass filters) for separation of observations from measurement noise;
- application of principal component method to achieve desirable level of orthogonalization between the variables selected;
- computing of extra indicators for the use in regression models.

It is also useful to test how informative is the data collected. Very formal indicator for data informativeness is sample variance. It is considered formally that the higher is the variance the richer is data with information. Another criterion is based on computing derivatives with a polynomial that describes data in the form a time series. For examples, such polynomial may describe rather complex process trend as follows:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + c_1 k + c_2 k^2 + ... + c_m k^m + \varepsilon(k), \qquad (1)$$

where $y(k)$ is the main variable; a_i, c_i are model parameters; $k = 0, 1, 2, ...$ is discrete time which is linked to real continuous time t via data sampling period T_s as follows: $t = kT_s$; $\varepsilon(k)$ is a random process that integrates influence of external disturbances to the process being modeled as well as model structure and parameters errors. Autoregressive part of the model (1) describes the deviations that are imposed on a trend, and the trend itself is described with the m-th order polynomial. In this case maximum number of derivatives can be m, though in practice actual number of derivatives is defined by the largest number i of the parameter c_i, that is statistically significant.

To select the best models constructed the following statistical criteria are used: determination coefficient (R^2); Durbin-Watson statistic (DW); Fisher F-statistic; Akaike information criterion (AIC), and residual sum of squares (SSE). The forecasts quality is estimated with hiring the following criteria: mean squared error (MSE); mean absolute percentage error ($MAPE$); and Theil inequality coefficient (U). To perform automatic model selection the following combined criterion is proposed:

$$V_N(\theta,\, D_N) = e^{|1-R^2|} + \ln(1+\frac{SSE}{N}) + e^{|2-DW|} + \ln(1+MSE) + \ln(MAPE) + e^U, \quad (2)$$

where θ is a vector of model parameters; $D_N -$ is a dataset of power N. The power of the criterion was tested experimentally and proved with a wide set of models and statistical data. Thus, the three sets of quality criteria are used to insure high quality of final result.

8 Coping with uncertainties

When considering mathematical models it is important to use a unified notion of model structure which we define as follows:

$$S = \{r,\ p,\ m,\ n,\ d,\ z,\ l\},$$

where is model dimensionality (number of equations); is model order (maximum order of differential or difference equation in a model); is a number of independent variables in the right hand side; is a nonlinearity and its type; is a lag or output reaction delay time; is external disturbance and its type; are possible restrictions for the variables.

The process of constructing a model in the form of BN can represented with the following steps: 1) - a thorough analysis of the process (object) under study aiming to detecting of its special functioning features and identification of parent and daughter variables; 2) - search and analysis of existing process models and determining the possibility of their usage in DSS; 3) - determining degree of relations between the process variables using special tests and expert estimates; 4) - reduction of the process dimensionality whenever this is possible; 5) - scaling and discretization of the data available when necessary; 6) - determining semantic restrictions on the future model; 7) - estimation of candidate model (directed acyclic graphs) structures using appropriate optimization procedures and score functions; 8) - candidate models analysis and selection of the best one using model quality criteria (including values of score functions); 9) - application of the model(s) constructed to solve the problem stated; 10) - computing inference with the model(s) constructed with regards to the variables selected, quality analysis of the result. In our case the final result of the model application is computing of client default probability with the conditions established by other model variables. According to alternative problem statement BM can be constructed for estimation of operational or other type of financial risks.

Support vector machine. Support vector machine (SVM) belongs to the group of techniques that determines classes with the limits for spaces. It also can be used for constructing SVM based regression models to solve forecasting problem. The support vectors are created with the vectors of data that lay on these limits. The classification result is successful if the space between the limits is empty. Usually SVM is hired to solve the problems of linear classification and regression analysis. The basic idea of SVM is in transformation of input vectors to the space of higher dimension with subsequent search of separating hyperplane with maximum distance in this space. Two parallel hyperplanes are built on both sides of separating hyperplane, and the separating hyperplane is the one that maximizes the distance between the two extra parallel hyperplanes. The algorithm is based on maximization of distance between the parallel hyperplanes what minimizes mean classification error.

Conclusions

The methodology for constructing of decision support system for mathematical modeling of economic and financial processes that is based on the system analysis principles: hierarchical system structure, taking into consideration of probabilistic and statistical uncertainties, generating of decision alternatives, and tracking of computational processes for all the stages of data processing.

The system proposed has a modular architecture that provides a possibility for the easy extension of its functional possibilities with new model parameter estimation methods, forecasting techniques, and alternative generating. High quality of the final result is achieved thanks to appropriate tracking of the computational processes for all data processing stages: preliminary data processing, model structure and parameter estimation, computing of short- and middle-term forecasts, as well as thanks to convenient for a user intermediate and final results representation. The system is based on different (ideologically different) techniques of modeling and forecasting what creates a good base for combination of various approaches to achieve the best results. The examples of the system application show that it can be used successfully for solving practical forecasting problems.

The DSS can be used for decision making support in various areas of human activities including strategy development for industrial enterprises, transportation and investment companies etc. Further extension of the system functions is planned with new forecasting techniques based on probabilistic technologies.

Acknowledgements

The work is supported by the state budget scientific research project at the National Technical University of Ukraine “Kyiv polytechnic institute”: “Development of information technology for modeling and estimation of financial risks (on the basis of Bayesian approach)" (the state registration number: 0113U000650). The authors are also grateful to the staff of local insurance company Brit Mark for their valuable advices regarding practical issues touching upon reinsurance strategies.

References

[1] McNeil A.J. Quantitative Risk Management / A.J. McNeil, R. Frey, P. Embrechts. – Princeton (New Jersey): Princeton University Press, 2005. – 538 p.

[2] International Convergence of Capital Measurement and Capital Standards. A Revised Framework. Comprehensive Version. – Basel Committee on Banking Supervision, Bank for International Settlements. – Basel, 2006. – 158 p.

[3] Mays E. (Ed.) Handbook of Credit Scoring / E. (Ed.) Mays. – Chicago: Glen lake Publishing Company, Ltd., 2001. – 460 p.

[4] Neil M. Using Bayesian networks to model expected and unexpected operational losses / M. Neil, N.E. Fenton, M. Tailor // Risk Analysis, 2005. – P. 34-57.

[5] Zgurovsky M.Z. Method of constructing Bayesian networks based on scoring functions / M.Z. Zgurovsky, P.I. Bidyuk, O.M. Terentyev // Cybernetics and System Analysis, 2008.– Vol. 44.– No. 2.– P. 219-224.

[6] Polovcev O.V. A System Approach to Modeling, Forecasting, and Management of Financial and Economic Processes / O.V. Polovcev, P.I. Bidyuk, L.O. Korshevnyuk. – Donetsk: Oriental Publishing House, 2009. – 286 p.

[7] Hollsapple C.W. Decision support systems / C.W. Hollsapple, A.B. Winston. – Saint Paul: West Publishing Company, 1996. – 860 p.

[8] Tsay R.S. Analysis of financial time series / R.S. Tsay. – Hoboken: Wiley & Sons, Inc., 2010. – 715 p.

[9] Bidyuk P.I Methods of Forecasting / P.I. Bidyuk, O.S. Menyailenko, O.V. Polovcev. – Lugansk: Alma Mater, 2008. – 608 p.

[10] Burstein F. Handbook of Decision Support Systems / F. Burstein, C.W. Holsapple. – Berlin: Springer – Verlag, 2008. – 908 p.

The Tourist Flows Formation Features of Domestic Tourism in the Russian Regions

Julia Vladykina and Andrey Faddeenkov
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: `j_vladikina@ngs.ru, faddeenkov@corp.nstu.ru`

Abstract

Features of formation of tourist traffic in the territorial aspect are considered. The econometric model of the average length of stay in hotels is built. The main factors affecting the duration of the stay of tourists in the regions are identified and analyzed.

Keywords: tourist attraction, econometric modeling, panel data model, reduction model, dummy variables.

Introduction

The relevance of domestic tourism, assimilating financial resources on its own territory and develop local infrastructure, today more than ever evident. Complex indicators of activity for the emerging domestic tourism destinations Russia have not made yet. This is due to the fact that the calculation of tourist traffic in general, and on the domestic market alone is not accurate enough and is focused mainly on income from international outbound tourism. Development of local travel agencies and, especially, tour operator networks, is at an early stage of formation. Therefore, the main indicator of the interest in the tourist center in Russia today can only be the average length of stay in hotels.

This figure is significant in European tourist flow measurement systems [1-4], because it is considered as the most visible and identifiable. This indicator is linked to the policy airlines' excursion fare "(agreement with hotel accommodation facilities to increase overnight stays), and they are trying to increase the rate of event programs and various of other accents in the sights of tourist centers. This indicator is shown in the brochures as an indicator of the interest of tourists in the Tourist Center, and as an indicator of investment attractiveness of the region.

1 Panel data model for the average length of stay

In the constructing of the model, average length of stay data were used from the site of the Federal Service of State Statistics (http://cbsd.gks.ru/). Baseline data are available for analysis, have a panel structure. In this regard the panel data model with fixed effects was chosen as the main model [5, 6].

In the role of effects of objects were the effects of regions objects (76 regions of Russia involved in the analysis). Time moments match with years from 2002 to 2013. Input factor reflecting the development of the region road infrastructure was also included in the model.

The equation of the model is as follows:

$$y_{it} = \mu + R_i + A_t + \theta\rho_{it} + \varepsilon_{it}, \tag{1}$$

where y_{it} - the average length of stay (number of nights/number of placed persons) in a objects of collective accommodation; i - number of the region ($i = 0, \ldots, 75$; number $i = 0$ corresponds to the Altai Region); t - number of the year ($t = 0, \ldots, 11$ number $t = 0$ corresponds to 2002 year); μ - unknown general mean; R_i - the effect of the i-th region; A_t - the effect of the year with the number t; ρ_{it} - the density of paved public roads per 1,000 sq.km. territory "i" in year "t"; θ - unknown parameter; ε_{it} - random error with respect to which the assumptions are valid: in all cases errors are independent and identically distributed with zero mean and variance σ^2_ε: $\varepsilon_{it} \sim (0, \sigma^2_\varepsilon)$, $i = 0, \ldots, 75$, $t = 0, \ldots, 11$.

The fact that the corresponding matrix of independent variables has linearly dependent columns is a feature of the model (1). The consequence of this is the impossibility of uniquely estimation of the model parameters. The solution is possible if to make a reduction of the model [7]. Reduction procedure is to move to a new dummy variables reflecting the effects of the difference of each of the quality factors with some chosen level. For regional factors such baseline was selected Altai region, for the time factor - in 2002. After the reduction of the model (1) takes the form

$$y_{it} = \mu + r_i + a_t + \theta\rho_{it} + \varepsilon_{it}, \tag{2}$$

where $r_i = (R_i - R_0)$ - dummy variable reflecting the difference of i-region's effect with the effect of the Altai Region ($i = 1, \ldots, 75$; $a_t = (A_t - A_0)$ - analogous a dummy variable for the year t ($t = 1, \ldots, 11$).

2 The results of econometric analysis

Statistical estimation of parameters and statistical hypothesis testing in this study were performed using the statistical package R [8]. After the calculations were as follows.

The coefficient of determination, reflecting the quality of the constructed model was equal to 0.9001. Statistics F-test during the test on the significance of the model is 85.33. Hypotheses about the relevance of all input factors is not rejected for the probability of error of 0.001 or less. The regional factor has the greatest explanatory power. It accounts for 74.11% of the explained variance. This is followed by the time factor (7.82%) and road infrastructure factor (0.17%).

A detailed analysis of the estimates of the model parameters (2) allowed a number of conclusions. Kabardino-Balkaria, Stavropol and Krasnodar region stand out from all the regions. They are the "old" tourist centers of the Soviet era and is actively are reforming of tourist and recreational infrastructure in order to resume the former tourist traffic. Kurgan, Tambov, Bryansk region and the Jewish Autonomous Region, are the regions of border trade with Kazakhstan, China and Belarus. Length of stay is significantly higher than the average in Russia (see Figure 1).

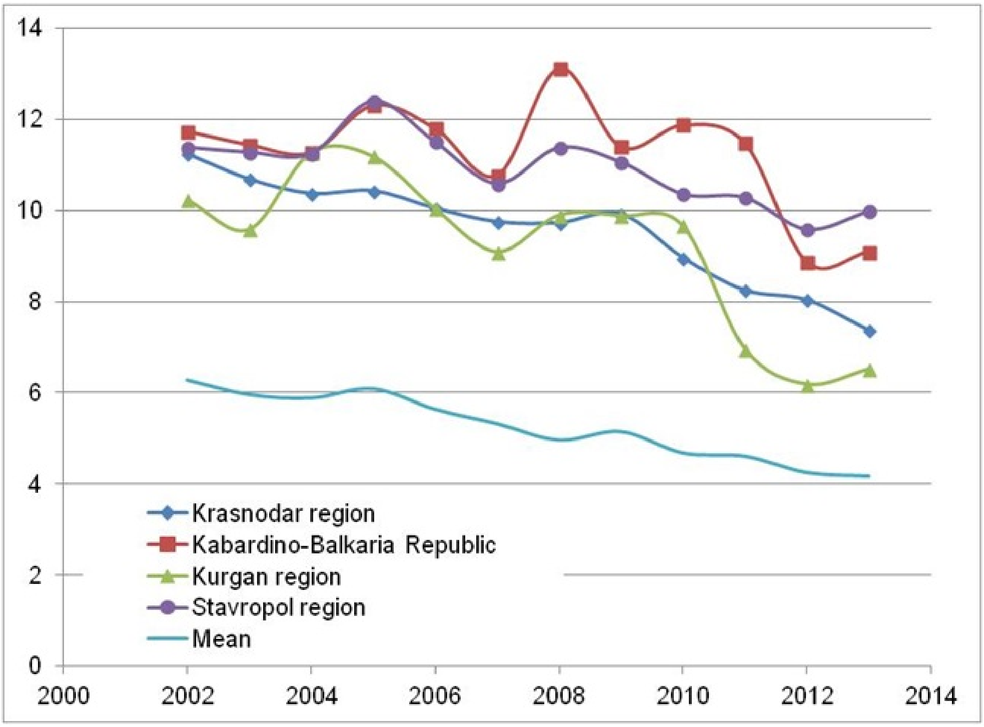


Figure 1: Dynamics of the average length of stay in the areas of resort and therapeutic types

Over time, the length of stay in the tourist centers in the whole of Russia is gradually reduced from an average of 6 to 4 days. The rate of decline is increasing, and has a definite linear form (see Figure 2).

Development of road infrastructure in comparison to other factors has the least impact, but that impact can not be ignored. Increasing of the road density works upon reduce of the average length of stay. Transport accessibility increases the proportion of tourists "Weekend" with a short stay. Increased mobility of tourists.

If we consider the average length of stay of tourists in the regions of Russia, the situation is as follows. Maximum length of stay of citizens there in the North Caucasus Federal District, and is approximately 9 days. This is due to the presence of the Stavropol Territory (10 days on average), and Kabardino-Balkaria (9 days), with are the maximum "points" in the spectrum of weeks long stay in Russia (see Figure 3).

The same can be said about the Southern Federal District with an average for the Krasnodar Territory about 8 days. Thus both of these regions in terms of reducing the overall length of stay in accommodation facilities of the Russian Federation can be considered the most affluent (average decline of about 30 percent compared to 45% in Russia for ten years from 2002 to 2013).

All other federal districts have approximately the same average length of stay (from 3.5 to 6 days), showing an equal potential for different segment groups: business objectives, recreational, cultural and historical, a longer stay is the result only of

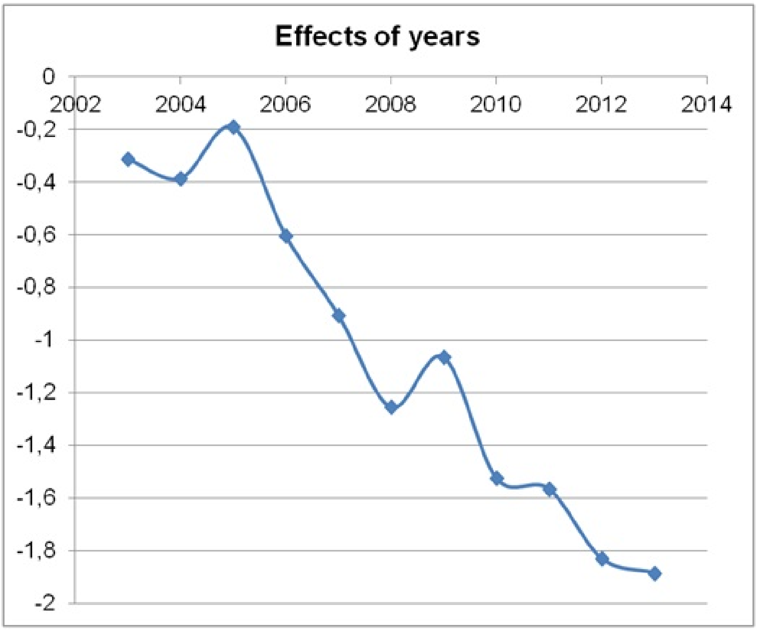


Figure 2: The effects of time factor by years

medical and beach tourism. We should also note, mark the situation in the Volga Federal District, where there is the sharpest decline in 2002-2013gg. (42%) due to a decrease in length of stay in the Perm region and the republic of Udmurtia with 9.5-10 days in 2002, respectively, to 4.9 and 5.2 days in 2013.

Conclusions

Reducing of the average length of stay in the Russian regions by 45% in 10 years, due to the increase of the total mobility of the population, reducing the time of holidays and the transition to a European standard short rest twice a year, or a weekend getaway. The tourism as enjoying of the cultural, historical and recreational value of the territory, in this case acts as an additional product to the main part of the business - tourism, the Tourist Center without commercial and business attractions, or an established brand hyped event, can expect to attract on tourists only in case of possession of exceptional natural and recreational or medical balneological resources.

Thus, the initial hypothesis of the interdependency road and stay of tourists, as the possibility of the development of domestic automobile tourism is not completely accurate. First accommodation facilities considered in those places where there are roads, and secondly a significant "regional bias" shows the benefits the policy of promotion and branding of regions, as one of the most effective tools for attracting tourists and residents of destination in its own recreation.

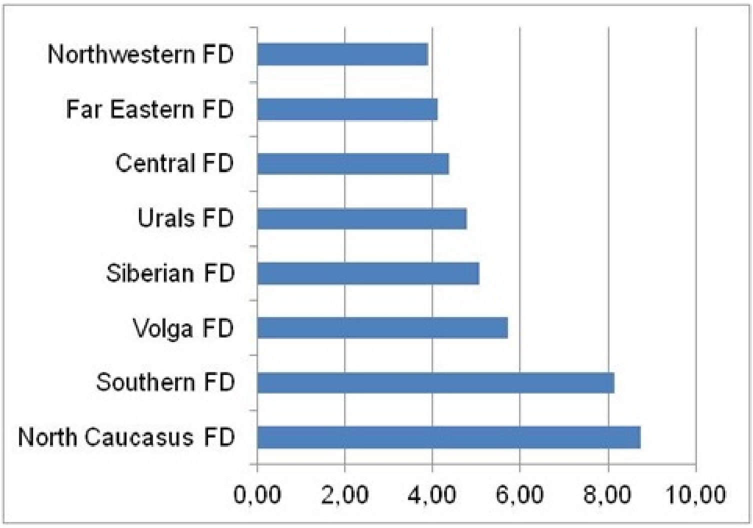


Figure 3: Average length of stay in the Russian regions (average for the 2002-2013 biennium)

References

[1] Azar V.I., Tumanov S.Y. (1998). *The economy of the tourist market*. Moscow, 239 p.

[2] Alexandrova A.Y. (2014). Tourist traps Regional Development *Initiative of the XXI century*. Vol. **2**, pp. 52-57.

[3] Kuznetsov Y.V. (2002). *Actual problems of tourism development in Russia at the present stage and tasks of the National Academy of Tourism*. St. Petersburg, 167 p.

[4] Vladykina J.O. (2014). Application the concept of marketing strategy in formation tourism and recreational clusters of the Novosibirsk region *Scientific journal "Vestnik NSUEM"*. Vol. **2**, pp. 304-312.

[5] Faddeenkov A.V. (2007). The analysis algorithms of linear regression models on panel data *Science bulletin of NSTU*. Vol. **3(28)**, pp. 65-78.

[6] Hsiao C. (2003). *Analysis of panel data*. Cambridge University Press, New York, 364 p.

[7] Searle S.R. (1971). *Linear models*. Wiley, New York, 532 p.

[8] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Bases of Stochastic Similarity of Difficult Systems

V.A. Shpenst[1], D.A. Pervukhin[1], D.A. Garanin[2]
[1] *The national mineral resource of the University "Mining", St. Petersburg, Russia*
[2] *St.-Petersburg Polytechnic University, St. Petersburg, Russia*
e-mail: kss1959@mail.ru

Abstract

In article is offered direction of the development to theories of the stochastic resemblance for efficient decision of the problems of the estimation and checking quality complex technical systems in process of their creation and usages. It Is Offered, is theoretically motivated and experimental is confirmed criterion of the stochastic resemblance in the manner of relations function distribution parameter (the features) of the systems. The most Further development to theories of the stochastic resemblance will allow to advance in study nonlimiting distribution, which little are described in classical theory of chances and mathematical statistics.

Keywords: product, system, control, quality, test, stochastic similarity, invariant similarity, modelling, the similarity theorem, criterion.

Introduction

Modern development of technics is characterised by sharp complication of the problems solved at manufacturing of products, high requirements to their reliability, deadlines of creation and introduction in operation, aspiration to reduce an expense for product working out at satisfaction of the set conditions. Existing methods of an estimation and quality assurance of a product by results of tests appear often inefficient in the conditions of the determined experiment or the available diverse, statistical information limited on volume on results of physical modelling, prototyping and tests of a small number of samples. One of solutions of a problem of maintenance of quality and reliability at a design stage and manufacturing of products is application of the theory of similarity and modelling.

The doctrine about similarity and modelling has started to be created more four hundred years ago. Leonardo da Vinci, Michelangelo, Galileem became attempts to prove methods of modelling and to apply them in various areas: to architecture, the mechanic, geometry, astronomy. However the first scientific formulations of a condition of similarity have been received by I.Newton in its work "the Mathematical beginnings of natural philosophy" in which it considers movement of material bodies and establishes laws of their similarity. It had been opened ways of application and modelling of mechanical systems and their criteria.

Academician M.V.Kirpichev [1] in the works has shown that the similarity theory is the experiment and modelling theory. She specifies, how it is necessary to put experience, to process the skilled data, and also to generalise and extend the received results to other objects.

Now the most urgent problem is development of the theory of similarity with reference to research problems of the big, difficult and non-uniform systems and objects.

Despite variety of the works devoted to methods of reduction of volumes and duration of tests of products for an estimation and control of their reliability, there are no the effective methods which are based on the uniform approach to the decision of problems on the basis of association of the diverse information by results of experimental working off, available in the limited volume.

As a whole, the similarity theory shows that any functional dependence between physical parametres of investigated object can be presented in the form of dependences between the criteria of similarity made of physical parametres. Thus, distinguish full, incomplete (partial, local, functional), approached and other kinds of similarity used in corresponding ways of modelling [3].

The special place occupies the stochastic similarity which description is given as in “the theory of stochastic similarity”, and in “the stochastic theory of similarity”. The essence of both statements is reduced to the description of similarity of stochastic objects by methods of probability theory and the mathematical statistics. Using geometrical interpretation, we will notice that speech in this case can go about similarity of set of the polygons which parties have the casual sizes.

In the generalised concept similarity of the phenomena is defined as proportionality each other all sizes characterising the phenomena, and proportionality factors (a similarity constant) keep constant value in all points of system for the certain name of sizes, but are various for sizes of the different name. Similarity of the phenomena it is possible to express also and so-called invariant similarity (idem) that means "same".

It is necessary to distinguish concepts “a similarity constant” and “invariant similarity”. The constant keeps constant value in all points of system, but it will be other when one steam of the similar phenomena another is replaced. Similarity constants are not any. Communication of the sizes entering into constants of similarity, is defined by law of the physical phenomenon and expressed in the form of the equations. Presence of such equation establishing dependence between sizes, imposes certain restrictions and on similarity constants.

Here communication of concepts “a similarity constant” and “ invariant similarity” stochastic the similar phenomena with regress of the random variables characterising these phenomena is traced. Really, let two phenomena characterised by two systems (vectors) of casual parametres (sizes)$(X_1, X_2, ..., X_i, ..., X_n)$ and $(Y_1, Y_2, ..., Y_j, ..., Y_n)$ identical dimension are observed n. Normalized the correlation matrix looks like [8]

$$\|r_{ij}\| = \left\| \begin{array}{ccc} r_{11} & ... & r_{1n} \\ ... & ... & ... \\ r_{n1} & ... & r_{nn} \end{array} \right\| \tag{1}$$

Where r_{ij} - correlation factor between X_i and Y_j.

Then a diagonal matrix $\|r_{ij}\|$ with elements $r_{ij} = 0$ if $i \neq j$, contains factors of correlation, the best approach of linear regress x on y with regress factors $\beta_{ij} = r_{ij}\frac{\sigma_x}{\sigma_y}$ where σ_x and σ_y - mean square deviations of random variables x and y accordingly [10].

Let’s consider substantive provisions of the theory of similarity.

The first theorem of similarity is formulated as follows [3]: at the similar phenomena criteria of similarity are numerically identical (necessary conditions of similarity).

The similar phenomena are characterised by a number of certain properties.

1. The sizes defining the phenomena in all points of system in which processes of the given phenomenon proceed, concern in homologic points the sizes with the same name from group of the similar phenomena, as constant numbers. Each size is answered with the number, various for each pair of the phenomena. Thus it is necessary to mean that the studied phenomena proceed and in geometrically similar systems.

2. The sizes characterising the considered phenomenon, are independent from each other, and between them there are certain communications. If these communications can be expressed in the form of mathematical dependences the last letter are identical to the similar phenomena.

The essence **of the second theorem** of similarity (π - theorems) consists in carrying over of the data of individual experience on all phenomena similar to it, with the help the criteria equations. That is, functional dependence between sizes characterising process can be presented in the form of dependence between the criteria of similarity made of them.

The third theorem of similarity establishes signs on which it is possible to learn, whether two phenomena each other are similar. The theorem recognises that the equations which connect among themselves sizes of the first phenomenon are known, and these equations of communication answer also to living conditions of unlimited number similar to the first phenomenon, i.e. existence of group of the similar phenomena is possible.

Besides, following additional positions to the basic theorems of similarity [3] are known.

1. Difficult systems are similar, if subsystems corresponding to them are similar and the criteria of similarity made of sizes, not entering in any of subsystems are equal.

2. Similarity conditions, fair for systems with constant parametres, it is possible to extend and to systems with variable parametres under condition of coincidence of relative characteristics of variable parametres.

3. Similarity conditions, fair for isotropic (homogeneous) systems, can be extended to anisotropic (non-uniform) systems if anisotropy in compared systems is rather identical. *That is, it actually a condition of association of non-uniform stochastic objects (selection).*

Conditions of similarity of the phenomena.

1. Geometrical similarity of systems and alphabetic similarity of the equations of communication (a necessary condition).

2. Similarity of conditions of unambiguity of the phenomenon allocating it from group of others (a necessary condition). If these conditions same, as well as at the first phenomenon, only numerical values of the sizes entering into them, at the second phenomenon others these sizes name *monovalently.* Accordingly, unambiguity condi-

tions name conditions of monovalency of the phenomenon. The choice of constants of similarity in similar systems is not any, for there are the causing equalities demanding that the *indicators of similarity* received from the equations of communication, equaled to unit. From here:

3. A necessary condition of similarity is equality to unit of indicators of the similarity made of constants of sizes, entering into an unambiguity condition. This requirement is answered with equality of criteria which carry the name defining for their invariancy is included into the conditions defining similarity of the phenomena.

Thus, the **third theorem of similarity** consists that those phenomena which occur in geometrically similar systems are similar, submit to the same equations of communication at which monovalents are in numerically constant relation and the criteria made of them are equal.

Hence, the phenomena are similar, if their expressed in relative units and monovalency criteria are identical.

The presented principles of similarity in stochastic sense are based that the parametres entering into criteria of similarity, are random variables, and criteria of similarity - functions of these random variables [3].

Then similarity stochasticcertain physical systems should be based on equality of functions of distribution of the parametres (sizes) characterising these systems.

The decision of a problem of an estimation of stochastic similarity of two systems characterised in casual parametres, is spent by check of a statistical hypothesis about equality of functions of distributions of these parametres. For this purpose it is necessary to know the law of distribution of the chosen criterion.

Thus, in stochastic statement by a similarity condition equality of functions of distribution of random variables (parametres) is. If to compare two classes of objects for which as a result of tests estimations of functions of distribution of their parametres the condition of the approached stochastic similarity according to the first can be written down similarity theorems in the form of affinity of selective criteria of similarity on probability [4] are received. According to the third theorem of similarity a sufficient condition of similarity of two systems is equality of any two corresponding criteria of similarity of these systems made of their key parametres and initial (boundary) conditions. Defining criteria are made of sizes independent among themselves which enter into unambiguity conditions (geometrical parities, physical parametres, regional conditions, initial and boundary).

Having chosen as criteria of stochastic similarity of function of distribution of some parametres, it is possible to write down a condition (criterion, the indicator) stochastic similarity in the form of the relation

$$Q = \frac{F_1(x_1, x_2, ..., x_n)}{F_2(y_1, y_2, ..., y_n)} \tag{2}$$

Where $F_1(\cdot)$ and $F_2(\cdot)$ – functions of distribution of parametres (characteristics) of objects by number n.

Let's note communications of criterion (2) with existing concepts.

1. Expression (2) represents *in regular intervals the* most powerful criterion of the relation of credibility [9].

2. At $Q = 1$ expression (2) degenerates in identity which is used as likelihood integrated transformation of random variables from different general totalities. For example, in Figure 1 procedure of transformation of a random variable x_1 with distribution function $F_1(x_1)$ in a random variable x_2 with distribution function is represented $F_2(x_2)$.

3. At $F_2(\cdot) = 1$ expression (2) degenerates in the stochastic superindicator [5] which is used for identification of laws of distribution for samples mainly small volume.

It is obvious that the size Q in expression (2) is casual. Then, having entered into consideration function of distribution of this relation (criterion of stochastic similarity), it is possible to estimate criterion, using concept of the significance value αused for check of statistical hypotheses.

According to a lemma [9] random variable $z = F(x)$ where $F(x)$ - distribution function, is in regular intervals distributed in an interval [0; 1]. Then, according to expression (2) there are two independent random variables z_1 andz_2, in regular intervals distributed in an interval [0; 1]. It is necessary to define the law of distribution private $q = \frac{z_1}{z_2}$, under a condition $z_1 \leq z_2$. (3)

Let's notice that the condition $z_1 \leq z_2$ is postulated from reasons of reception of foreseeable area of distribution of a random variable q. Further advantage of introduction of this condition will be shown. The density of joint distribution of the ordered random variables z_1 also z_2 looks as follows [6]

$$f(z_1, z_2) = 2! f_{z_1}(z_1) f_{z_2}(z_2) = 2, \tag{3}$$

Where $f_{z_1}(z_1) = 1$ and $f_{z_2}(z_2) = 1$ – density of distribution of independent random variables z_1 and z_2.

In a general view the density of distribution of the relation q looks like [7]

$$g(q) = -\int\limits_{-\infty}^{0} z_1 f(z_1,qz_1)dz_1 + \int\limits_{0}^{\infty} z_1 f(z_1, qz_1)dz_1.$$

Having substituted in this expression the formula (3), and, having rejected the first integral as $z_1 \in [0; 1]$, we will receive

$$g(q) = \int\limits_{0}^{1} z_1 f(z_1, qz_1)dz_1 = \int\limits_{0}^{1} 2z_1 dz_1 = 2 \cdot \frac{z^2}{2}\Bigg|_{0}^{1} = 1\,. \tag{4}$$

That is, the relation from two independent random variables in regular intervals distributed in an interval [0; 1], there is a random variable in regular intervals distributed in an interval [0; 1].

For acknowledgement told in Figure 1 the typical histogram of distribution of the random variable q, constructed by results of computing experiment is resulted.

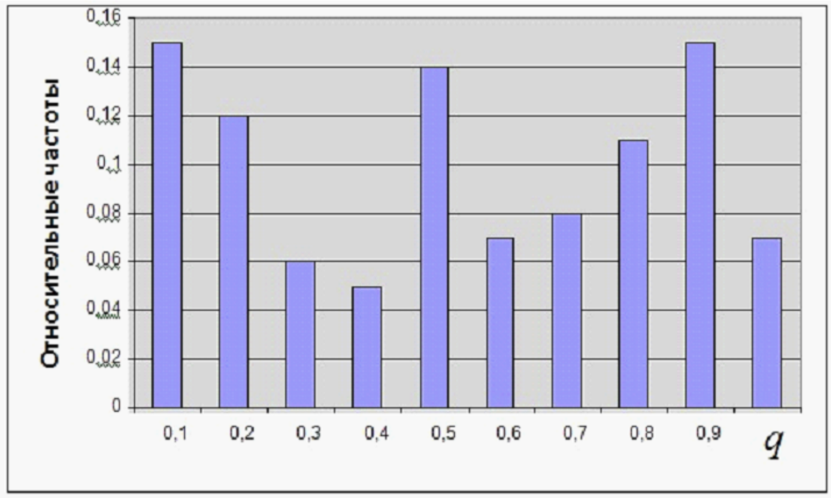


Figure 1

It is possible to show that the statistics of *criterion Kolmogorov* for hypothesis check about uniform distribution of a random variable q is equal

$$\lambda = D\sqrt{n} = |\max(\Delta)|\,\sqrt{n} = 0,07\sqrt{100\ =}0,7.$$

Probability $P(\lambda) = 0,711$ that much more level $\alpha = 0,1$. Therefore the bases to reject a hypothesis about uniform distribution of a random variable q in an interval [0; 1] is not available [8].

For comparison in **Figure 2** the histogram of empirical distribution of the simple relation of two independent in regular intervals distributed in an interval [0 is resulted; 1] random variables.

Apparently from the figures, the given distribution of the general with the uniform has no anything. Besides, at it long enough "tail" that would complicate decision-making because of weak discernability of threshold values of criterion on "tail".

Thus, results of computing experiment (imitating modelling) testify to justice of expression (4) for the description of density of distribution of criterion (2).

Now for hypothesis check about stochastic similarity of two objects characterised by two samples of supervision, it is enough to compare settlement value of criterion (2) with theoretical, having integrated density (4) on the area limited to a significance value α. It is obvious that for the uniform law it does not represent any difficulties.

Generalising told, it is necessary to notice that the theory of stochastic similarity is some generalisation of probability theory and the mathematical statistics on the similarity theory. It is intermediate between full uncertainty and probability theory with mathematical statistics (which operate with laws of distribution and their parametres) and allows to do a conclusion about similarity of stochastic objects in certain conditions.

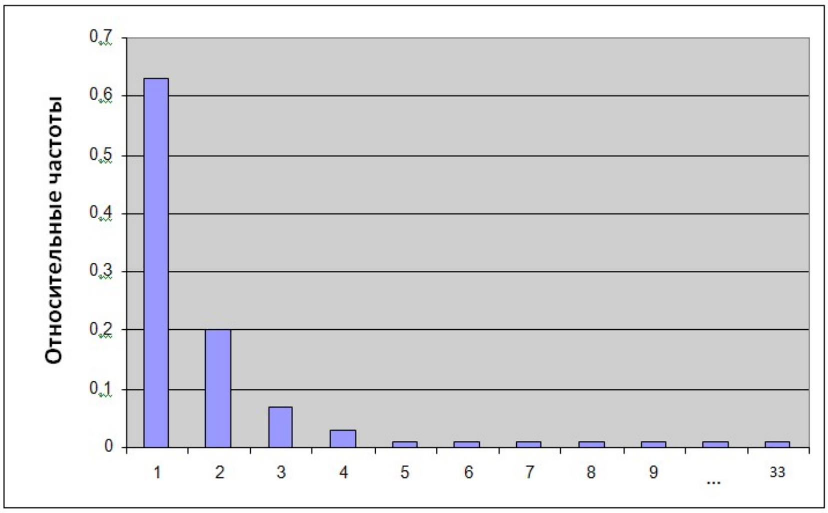


Figure 2

Its use gives the chance carrying out of certain manipulations with stochastic similar objects. The further development of the theory of stochastic similarity will allow to promote in research of nonlimiting distributions which are a little described in classical probability theory and the mathematical statistics. Especially that their part which will give the chance to manipulate samples of small volume and censor samples, taking from them an information maximum.

References

[1] Kirpichev M.V. Teorija of similarity. M: AH the USSR, 1953. - 213 p.

[2] N.Chhaidze. Methods of similarity and mathematical modelling in research of difficult systems. Tbilisi: the Publishing house "Technical university", 2009. - 193 p.

[3] Reliability and efficiency in the technician: the Directory: Vol. 4.: similarity Methods in reliability/under V.A.Melnikov, N.A.Severtseva. - M: Mechanical engineering, 1987. - 280 p.

[4] Guhman A.A. introduction in the similarity theory. M: the Higher school, 1973. - 296 p.

[5] Martyshchenko L.A., Volovik A.V., Klavdiev A.A., etc. methods of rationing of reliability of difficult systems of the weapon. - L: MO, 1992. 330 p.

[6] N.Johnson, F.Lion. Statistics and experiment planning in the technician and a science. Data processing methods. - M: the Word, 1980. - 610 p.

[7] Venttsel E.S., Ovcharov L.A.applied of a problem of probability theory. - M: Radio and communication, 1983. - 416 p.

[8] Ventceli E.S.probability theory. - M: the State publishing house of the physical and mathematical literature, 1962. - 560 p.

[9] Gnedenko B. V, Beljaev J.K., A.D.mathematical's Nightingales methods in reliability theory. - M: the Science, 1965. - 523 p.

[10] Cornet G, Korn T.Spravochnik on the mathematician for science officers and engineers. - M: the Science, 1984. - 832 p.

Potential Distribution of Probabilities and Stochastic Similarity in Comparison Problems

O.V. Afanasyeva, I.N. Kivaev, I.A. Klavdiev
The national mineral resource of the University "Mining", St. Petersburg, Russia
e-mail: ovaf@rambler.ru, kivin83@yandex.ru, kss1959@mail.ru

Abstract

Existing methods of an estimation of quality (efficiency) of samples in the determined statement appear unacceptable in uncertain conditions of their use. Working out of such samples is based on performance of requirements on reaching some quality. The choice of this or that variant is based on procedures of estimation which product indicators and criterion substantiation (or criteria) includes comparisons. The greatest complexity is represented thus by a choice of a way of data of private indicators to generalised (criteria) on which the decision is made. In article the concept of a method of potential distribution of probabilities and stochastic similarity in comparison problems is resulted.

Keywords: potential distribution of probabilities, stochastic similarity, product indicators substantiation, criteria of comparison.

Introduction

Modern development of technics is characterised by complication of the problems solved by various samples which owing to integration of the diverse and distributed elements more and more can be considered as difficult systems. Working out of such samples, as a rule, is based on performance of requirements to destination (achievement of some quality), which satisfaction, generally, in probably various ways. The choice of this or that variant is based on the estimation which procedure includes product indicators and criterion substantiation (or criteria) comparisons. The greatest complexity is represented thus by a choice of a way of data of private indicators to generalised (criteria) on which the decision is made.

Existing methods of an estimation of quality (efficiency) of samples in the determined statement appear unacceptable in uncertain conditions of their use that causes necessity of working out of ways of the decision of a problem for the stochastic statement which urgency is dictated by the requirement of timely reaction of market conditions in the conditions of a rigid competition and uncertainty of demand and application of samples to destination.

The problem of comparison of samples consists in a general view in:
formation of product indicators, characterising the most important properties;
choice of criterion of comparison (the generalised indicator);
estimation of compared samples.

Results of the decision of a problem are used for acceptance of this or that decision depending on a research objective.

1. Potential distribution of probabilities

The information situation at which use potential distribution of probabilities, is characterised by that are known only given about corresponding characteristics of analogues is exemplary [1]. In this case it is represented expedient to put forward a hypothesis about linear convolution of some private dimensionless indicators. For definition of weight factors of such convolution there is enough of various methods. All of them are based on this or that model of behaviour of environment which, as a rule, is postulated in the informal image. Meanwhile objectivity the models constructed with use of a principle of a maximum of uncertainty possess. One of possible approaches of an estimation of the specified weight factors who is based on this principle, the method of potential distribution of probabilities is. The maintenance of a considered situation thus can be presented the following scheme.

Is available n samples of products who on the shape and appointment can be considered as analogues of some prototype. Set of characteristics (product indicators) defining its degree of quality (efficiency) is put each of these samples-analogues in conformity. Let such characteristics will be m. We will designate through x_{ij} $(i = 1, n)$ $j = \overline{1, m}$ – private indicators of compared objects. The initial data thus it is convenient to have in the form of a matrix

$$X = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots \\ x_{1m} & x_{2m} & \dots & x_{nm} \end{bmatrix}.$$

The weight j – of that characteristic in distribution of means to achievement of demanded level of an indicator of quality of samples generally is unknown. It is required to compare available samples-analogues taking into account objectively existing uncertainty.

The principle of potential distribution postulates for comparison use of criterion of Bajesa as the complex indicator reflecting quality of this or that sample. It has the following appearance

$$b_i = \sum_{j=1}^{m} p_j \, r_{ij}, \tag{1}$$

Where r_{ij} – dimensionless indicators, and $r_{ij} = x_{ij}/x_{Bj}$ if the increasex_{ij} leads to growth b and $r_{ij} = x_{Bj}/x_{ij}$ if the increasex_{ij} leads to reduction b;

x_{Bj} – the characteristic of the base sample in which quality any of the presented samples is considered.

Then weight factors p_j, $(j = \overline{1, m})$, reflecting certain model of behaviour of environment, are by maximisation of entropy of Shennona

$$H = -\sum_{j=1}^{m} p_j \ln p_j \to \max; \tag{2}$$

At restrictions

$$\sum_{j=1}^{m} p_j = 1 \quad ; \quad \prod_{j=1}^{m} \overline{r}_j^{\,p_j} = const.$$

It is possible to show that expression for estimations of weight factors has the following appearance

$$p_j = \left(\sum_{i=1}^{n} r_{ij} \right)^{-1} \left[\sum_{j=1}^{m} \left(\sum_{i=1}^{n} r_{ij} \right)^{-1} \right]^{-1}. \tag{3}$$

Restrictions in (2) postulate a condition of normalization and a constancy of the compound. Physically the last means that the relative increment of weight j – – of that characteristic to proportionally relative increment of an indicator concerning level of the same characteristic on all set of considered samples, and proportionality factor depends on the reached level.

The lack of a subjective choice of the base sample is inherent in the given method that, however, does not influence quality standard by a principle "is better-is worse".

2. Stochastic similarity

Now the most urgent problem is development of the theory of similarity with reference to research problems of the big, difficult and non-uniform systems and objects. As a whole, the similarity theory shows that any functional dependence between physical parametres of investigated object can be presented in the form of dependences between the criteria of similarity made of physical parametres [3].

The special place occupies the stochastic similarity which objects of research in considered statement are samples of supervision of the random variables, supposing any physical interpretation.

Known principles of similarity in stochastic sense are based that the parametres entering into criteria of similarity, are random variables, and criteria of similarity - functions of these random variables [2]. *Then similarity stochasticcertain physical systems should be based on equality of functions of distribution of the parametres (sizes) characterising these systems.*

The decision of a problem of an estimation of stochastic similarity of the systems characterised in casual parametres, is spent by check of a statistical hypothesis about equality of functions of distributions of these parametres [4].

It is necessary to notice that at considered above potential distribution of probabilities and stochastic similarity much in common. Really, transition to dimensionless indicators r_{ij} in expression (1) is that other, as the similarity relation. Besides, an essence of the second theorem of similarity (π – theorems) is carrying over of the data of individual experience on all phenomena similar to it, with the help criteria the equations. That is, functional dependence between sizes characterising object can be presented in the form of dependence between the criteria of similarity made of them [2].

Conclusion

It is possible to show that at comparison by a method of potential distribution of probabilities of two objects which private characteristics are proportional with factor c, their complex indicators (1) will differ also with this factor. It means that stochastic similarity, being a special case of potential distribution of probabilities, establishes a parity between objects in the conditions of proportional (with the factor equal to a constant of similarity) of change of their characteristics.

Thus, stochastic similarity, being one of methods of research of difficult systems, allows to solve problems of their comparison by a direct estimation of a parity of similar indicators.

References

[1] Ivchenko B. P, Martyshchenko L.A., Tabuhov M. E. Management in economic and social systems. The system analysis. Decision-making in the conditions of uncertainty. - SPb.: "Nordmed-Izdat", 2001. - 248 p.

[2] Reliability and efficiency in the technician: the Directory: T. 4.: similarity Methods in reliability/under V.A.Melnikova, N.A.Severtseva. - M: Mechanical engineering, 1987. - 280 p.

[3] Venttsel E.S., Ovcharov L.A.applied of a problem of probability theory. - M: Radio and communication, 1983. - 416 p.

Nonparametric Identification of Stochastic Models of Difficult Systems

A.A. Klavdiev[1], A. V. Volovik[2], S.V. Efimenko[1]

[1] *The national mineral resource of the University "Mining", St. Petersburg, Russia*

[2] *OAO"Klimov", St. Petersburg, Russia*

e-mail: kss1959@mail.ru, volovik_aleksandr@mail.ru, falcon.sergey@yandex.ru

Abstract

In article the effective nonparametric method of identification of models of refusals of radio-electronic equipment of means of automation of management is proved by dynamic systems. The essence of the offered method consists that on the small samples presented in the form of a variation number, practically always it is possible to find such transformation in which result the statistics which is not dependent on parametres of distribution of general totality will turn out. Function of distribution of such statistics it is represented expedient to define as a result of statistical modelling if its analytical construction is complicated.

Keywords: model of refusals, identification, a nonparametric method, dynamic system, the superindicator.

Known experience of working out and application of separate mathematical receptions for identification of models of refusals of radio-electronic equipment of means of automation of management by dynamic systems [1] is by this time stored. However till now there was no uniform methodical basis of reception of discrimination functions. The problem becomes complicated that decisions are correct only under condition of the account of set of characteristics between which stochastic communication is possible. It is necessary not only to reveal and estimate this communication, but also to consider in the course of the decision of problems of synthesis and the analysis of electronic systems.

Owing to that character of signals does not give in to the strict regular description, it makes sense to use for formalisation by probability theory elements. Having entered into consideration casual character of investigated parametres x_i, it is possible to apply known receptions of their formal description. It is obvious that the fullest information on casual process the density of distribution of its co-ordinates $f(x_1, x_2, ..., x_n)$. However bears complexity of use of multidimensional distributions consists in known difficulties of their identification and limitation of forms of the obvious description. So, perhaps, unique generalisation of multidimensional distribution in an explicit form is the multidimensional density of the normal law

$$f(x_1, x_2, ..., x_n) = Ce^{-Q(x_1-a_1, x_2-a_2, ..., x_n-a_n)},$$

Where

$Q(x_1, x_2, ...x_n) = \sum_{k,l=1}^{n} q_{kl} x_k x_l$ – Positively certain square-law form;

$a_1, a_2, ..., a_n$ – Population means of random variables $x_1, x_2, ..., x_n$;

Factors C also $q_{kl} = q_{lk}$ are expressed through dispersions $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ and correlation factors r_{kl} between x_k and x_l.

Normal distribution plays a fundamental role in the theory of probability and the mathematical statistics since substantive provisions of these sections of mathematics are based on the assumption of normal distribution of general totality. However the information situation in which conditions the decision is made, often does not allow to postulate unequivocally an assumption about normal distribution.

Complexity and heterogeneity of conditions of realisation of electronic systems causes necessity of research, for which basis it is expedient to put specific methods of the statistical analysis. Experience of carrying out of similar researches in various areas of a science and technics has allowed to reveal a number of positions which can effectively be applied to an estimation of dynamics of processes of functioning of electronic systems. However feature of fast ageing of the information and its limited volume causes application of the methods using invariant statisticians of the theory of stochastic indication [2].

The essence of the given approach consists that on the small samples presented in the form of a variation number, practically always it is possible to find such transformation in which result the statistics which is not dependent on parametres of distribution of general totality will turn out. Function of distribution of such statistics it is represented expedient to define as a result of statistical modelling if its analytical construction is complicated. That is the given approach belongs to the class of nonparametric methods of check of hypotheses about a kind of the law of distribution.

Basis of construction of the transformation leading to formation of the invariant statistics, the variation number $x_1^{(m)} \leq x_2^{(m)} \leq ... \leq x_m^{(m)}$ made of sample of independent random variables serves $x_1, x_2, ..., x_m$. The density $f(x_1, x_2, ..., x_m)$ of joint distribution of members of a variation number is defined by expression

$$f(x_1, x_2, ..., x_m) = m! \prod_{i=1}^{m} f_i(x_i) \quad ,$$

Where $f_i(x_i)$ – density of distribution of a random variable x_i;

m – Quantity of supervision in sample.

To get rid of parametres of distribution of general totality it is possible, having subjected to members of a variation number to intermediate transformation. So, for example, for sample of random variables x_i in volume m=2 of general totality with exponential distribution law such transformation looks like

$$\aleph = \frac{x_1}{x_2}, \quad x_1 \leq x_2.$$

Really, having applied N.V. Smirnova's return transformation to random variables x_1 and x_2, we will receive expression

$$\aleph = \frac{\ln(1 - \alpha_1)}{\ln(1 - \alpha_2)},$$

Which does not depend on parametres of exponential distribution, but depends only on the random variables $\alpha_1 \leq \alpha_2$ in regular intervals distributed with joint density of probability

$$f_\alpha(\alpha_1, \alpha_2) = 2! \quad .$$

It is similarly possible to show that for sample in volume m=3 of general totality with the uniform law of distribution intermediate transformation looks like

$$\aleph = \frac{x_2^{(3)} - x_1^{(3)}}{x_3^{(3)} - x_1^{(3)}} = \frac{\alpha_2 - \alpha_1}{\alpha_3 - \alpha_1},$$

Where$\alpha_1 \leq \alpha_2 \leq \alpha_3$ – the ordered random variables in regular intervals distributed in the range of [0,1].

For sample of the same volume of general totality with the normal law of distribution:

$$\aleph = \frac{x_2^{(3)} - x_1^{(3)}}{x_3^{(3)} - x_1^{(3)}} = \frac{\eta_2 - \eta_1}{\eta_3 - \eta_1},$$

where $\eta_1 \leq \eta_2 \leq \eta_3$ – the ordered random variables distributed under the standard normal law.

The increase in supervision in sample allows to build set of intermediate transformations under the similar scheme. Such set is characterised by integrated function of joint distribution $G(\aleph_i, i = \overline{1, m-r})$, where r – number of parametres of distribution of general totality. Owing to an ambiguous arrangement of critical zones for $\aleph_i, i = \overline{1, m-r}$ at a preset value G it is represented expedient to apply a method of stochastic indication according to which $G(\aleph_i, i = \overline{1, m-r})$ acts in a role of the superindicator [1,2].

Stochastic superindicator S represents probability of the event which outcome depends on a parity of two or several random variables. In our case the superindicator acts in a role of nonparametric criterion of the consent. Legitimacy of its use is based on the following statement.

The statement. Let it is required to check up a hypothesis$H_0' : G(S) \equiv G_1(S_1)$, where $G_1(x)$ – function of hypothetical distribution of a random variable x. We will enter into consideration random variables $S = G(x)$ and $S_1 = G_1(x)$. Then, if equality $G(x) = G_1(x)$ expression $H_0' : F(S) \equiv F_1(S_1)$ where $F(S)$ and $F_1(S_1)$ – functions of distribution of superindicators is fair is carried out S and S_1. Hence, hypothesis check H_0 is equivalent to hypothesis check H_0'.

Process of formation of superindicator S, its function of distribution $F(S)$ and the nonparametric criterion of the consent based on it for some organic laws of distribution of general totality, and also research of its capacity, are in detail stated in [2].

Let's notice that superindicators for various laws of distributions can be in a similar way generated. However to receive final analytical dependences not always it is obviously possible. In such cases the problem can be solved numerical methods.

Besides, in the presence of some aprioristic data about a class of distribution of general totality, for example, when parametres of the prospective law are known, to check up a hypothesis it is possible, having spent transformations of available random variables in uniform, normal or exponential. Then enough to use the corresponding superindicator for identification of the transformed random variables.

Really [3,4], sample $x_1^{(n)}...x_n^{(n)}$ of random variables of general totality with the normal law of distribution $N\ (m,)$, where m – a population mean let is given; n – rms a deviation.

If sequence of random variables $x_1^{(n)}...x_n^{(n)}$ to present in the form of a variation number

$$x_1^{(n)} < ... < x_k^{(n)} < ... < x_n^{(n)},$$

Where $x_k^{(n)} = \frac{\sum\limits_{i=1}^{n-1} x_i^{(n)}}{k}$ and $k = n$-2,

Then density of joint distribution of the relation

$$\aleph = \frac{x_k^{(n)} - x_1^{(n)}}{x_n^{(n)} - x_1^{(n)}}, \quad \aleph \in [0,1], \quad n = 3, ...\infty$$

Looks like

$$f(\aleph) = \frac{\left(\frac{k+1}{k} - \aleph\right)}{arctg\sqrt{k}\sqrt{\aleph^2 - 2\aleph + \frac{k+1}{k}} \cdot \left(\aleph^2 - 2\aleph + \frac{k+1}{k}\right)}.$$

The proof.

It is necessary to enter random variables into consideration:

$$y_1 = \frac{\eta_k^{(n)}}{\sqrt{k}} - \eta_1^{(n)}; \quad y_2 = \frac{\frac{\eta_k^{(n)}}{\sqrt{k}} - \eta_1^{(n)}}{\eta_n^{(n)}}; \quad \aleph = \frac{\frac{\eta_k^{(n)}}{\sqrt{k}} - \eta_1^{(n)}}{\eta_n^{(n)} - \eta_1^{(n)}}.$$

Then

$$\eta_1^{(n)} = y_1 \left(\frac{1}{y_2} - \frac{1}{\aleph}\right), \quad \eta_k^{(n)} = y_1\sqrt{k}\left(1 + \frac{1}{y_2} - \frac{1}{\aleph}\right), \quad \eta_n^{(n)} = \frac{y_1}{y_2}.$$

Hence, Jakobian

$$\frac{\partial\left(\eta_1^{(n)},\eta_k^{(n)},\eta_n^{(n)}\right)}{\partial(y_1,y_2,\aleph)} = \begin{vmatrix} \frac{1}{y_2} - \frac{1}{\aleph} & -\frac{y_1}{y_2^2} & \frac{y_1}{\aleph^2} \\ \left(1 + \frac{1}{y_2} - \frac{1}{\aleph}\right)\sqrt{k} & -\frac{y_1}{y_2^2}\sqrt{k} & \frac{y_1}{\aleph^2}\sqrt{k} \\ \frac{1}{y_2} & -\frac{y_1}{y_2^2} & 0 \end{vmatrix} = -\sqrt{k}\frac{y_1^2}{y_2^2\aleph^2}.$$

Means

$$f\left(y_1, y_2, \aleph\right) = \frac{1}{C\eta} f\left(\eta_1^{(n)}, \eta_k^{(n)}, \eta_n^{(n)}\right) \cdot \begin{vmatrix} \frac{1}{y_2} - \frac{1}{\aleph} & -\frac{y_1}{y_2^2} & \frac{y_1}{\aleph^2} \\ \left(1 + \frac{1}{y_2} - \frac{1}{\aleph}\right)\sqrt{k} & -\frac{y_1}{y_2^2}\sqrt{k} & \frac{y_1}{\aleph^2}\sqrt{k} \\ \frac{1}{y_2} & -\frac{y_1}{y_2^2} & 0 \end{vmatrix} =$$

$$= \frac{3!\, e^{-\frac{y_2}{2}d}\, y_1^2\sqrt{k}}{(2\pi)^{3/2}\, y_2^2\aleph^2 C\eta}.$$

Therefore $f\left(y_2, \aleph\right) = \frac{3}{\pi C\eta}\frac{\sqrt{k}}{y_2^2\aleph^2 d\sqrt{d}}$,

Where $d = \frac{y_2^2\left(k\aleph^2-2k\aleph+k+1\right)+y_2\left(k\aleph^2-2k\aleph+k+1\right)+k\aleph^2+2\aleph^2}{y_2^2\aleph^2}$.

Further designations are entered: $a = k\aleph^2 - 2k\aleph + k + 1$;

$$b = 2k\aleph^2 - 2k\aleph - 2\aleph;$$

$$c = 2\aleph^2 + k\aleph^2;$$

$b_1 = b/a$ and $c_1 = c/a$.

Then $f\left(y_2, \aleph\right) = \frac{3}{\pi C\eta}\frac{\aleph y^2\sqrt{k}}{a\sqrt{a}(y_2^2+b_1 y_2+C_1)\sqrt{y_2^2+b_1 y_2+C_1}}$.

Hence, $f(\aleph) = \frac{3\sqrt{k}\aleph}{\pi a\sqrt{a}C\eta}\left\{y_1 - \frac{b_1}{2}y_2\right\}$ and$f(\aleph) = \frac{3\sqrt{k}\aleph\cdot 4\left(2k\aleph^2-2k\aleph-2\aleph\right)}{\pi C\eta\sqrt{a}(b^2-4ac)}$,

Where $b^2 - 4ac = 4\aleph^2\left\{-2\aleph^2 k + 2k\aleph - (k+1)\right\}$.

Means $f\left(\aleph\right) = \frac{3}{\pi C\eta}\frac{\left(\frac{k+1}{k}-\aleph\right)}{\pi C\eta\sqrt{\aleph^2-2\aleph+\frac{k+1}{k}}\left(\aleph^2-2\aleph+\frac{k+1}{k}\right)}$.

Considering, as $F\left(\aleph\right) = \frac{6}{\pi C\eta}arctg\left(\frac{\aleph}{\sqrt{\aleph^2-2\aleph+\frac{k+1}{k}}}\right)C\eta = \frac{6}{\pi}arctg\sqrt{k}$, we receive a definitive kind of expression for$f\left(\aleph\right) = \frac{\frac{k+1}{k}-\aleph}{arctg\sqrt{k}\left(2\aleph^2-2\aleph+\frac{k+1}{k}\right)\sqrt{2\aleph^2-2\aleph+\frac{k+1}{k}}}$.

Hence $F\left(\aleph\right) = \frac{arctg\frac{\aleph}{\sqrt{\aleph^2-2\aleph+\frac{k+1}{k}}}}{arctg\sqrt{k}}$.

Procedure of check of a hypothesis about an accessory of sample of general normal set consists in the following.

On sample of random variables a $x_1^{(n)}...x_n^{(n)}$ variation number is under construction $x_1^{(n)} < ... < x_k^{(n)} < ... < x_n^{(n)}$,

Where$x_k^{(n)} = \frac{\sum_{i=1}^{n-1} x_i^{(n)}}{k}$ and k=n-2 also it is calculated $\aleph = \frac{x_k^{(n)}-x_1^{(n)}}{x_n^{(n)}-x_1^{(n)}}$.

Value of the superindicator pays off

$$S_P = F\left(\aleph\right) = \frac{arctg\frac{\aleph}{\sqrt{\aleph^2-2\aleph+\frac{k+1}{k}}}}{arctg\sqrt{k}}.$$

Critical value of superindicator S_{KP} =a, where and - a significance value is defined.

Are compared settlement S_P and critical S_{KP} values of the superindicator. Thus, if $S_P < S_{KPthe}$ hypothesis about the normal law of distribution of initial sample is rejected; if $S_P >$ S_{KP} there are no bases to reject the put forward hypothesis.

Expansion of possibilities of application offered for identification of models of refusals (including is likelihood-physical) method radio equipment is reached by carrying out preliminary operational (algebraic or integro-differential) transformations of the initial information.

References

[1] Afanasyeva O.V., Glozshteyn N.V. Raschchet reliability of complex systems // Proceedings of the IX International scientific-practical conference of young scientists and graduate students "analysis and forecasting management systems", St. Petersburg. SZTU. 2008. Pp. 210-213.

[2] Klavdiev A.A., Volovik A.V., V. A. An identification method processes of Markov under the limited information. Magazine "Military technology Standardization", No 2, Moscow 1993, 58-69 p.

[3] Klavdiev A.A., Soskin A.V., Soskin S.V. Method of nonparametric identification of functions of requirement and cost of equipment elevating means of transport terminals. Works of the Russian-Polish conference "the Analysis, forecasting and management in difficult systems" (APS-2006), SPb, SZTU, 2006, 283 p.

[4] The patent for useful model No. 65664 (Russian Federation), G06F15/36, the Device for identification selection / A.A.Klavdiyev, O V. Afanasyev, etc. It is published 10.08.2007, the bulletin No. 21.

Information and Statistical Justification of the Alternative Acceptance Control Products

A.V. Volovik[1], A.A. Klavdiev[2], S.V. Efimenko[2]
[1] *St.-Petersburg Polytechnic University, St. Petersburg, Russia*
[2] *The national mineral resource of the University "Mining", St. Petersburg, Russia*
e-mail: volovik_aleksandr@mail.ru, kss1959@mail.ru,
falcon.sergey@yandex.ru

Abstract

The Plan of the statistical reception checking - a set of the rules, on which check the party and come to a conclusion about her(its) acceptability or unacceptability. Consequently, problem consists in motivated choice of the variant of the plan of the checking, which takes into account the information about reliability from usage product. Firm operating the system of the provision to reliability product different purpose possible at presence to feedback, which allows to develop and correct the system of the selective checking in process production (the repair). The Purpose given work was a show to need of the account to serviceability product at organizations of the system of the reception checking the enterprise.

Keywords: sustainable functioning of the system, ensuring reliability, sampling plan, statistical acceptance control.

Steady functioning of system of maintenance of reliability of products of different function probably in the presence of the debugged feedback which by data about refusals and the malfunctions arisen for the industrial reasons and revealed in operation, allows to develop and correct system of selective control in the course of manufacture (repair).

The *system of selective control* represents a set *of plans of selective control* and corresponding schemes [1]. In turn, the plan *of statistical acceptance control* is a set corrected, on which supervise party and the decision on its acceptability or unacceptability make. And the scheme *of selective control* – the *procedure* establishing rules of switching from one plan for another, and returnings to the initial plan. Hence, the problem consists in a well-founded choice of a variant of the plan of control which considers data on reliability from operation of products.

As variant of the plan of control is called set of dependences between volume of sample and $n$party volume N, at various values of a rejection degree of quality q_m(level of discrepancies [2]), corresponding to a preset value of risk of the consumer [β3]. Thus, for a choice of the plan of control it is necessary to define [3]:

Value of risk of the consumer β;

Value of a rejection degree of quality q_m;

Rejection variant.

Value of risk of the consumer β is established by competent bodies or the agreement between the supplier and the consumer. For definiteness we will be set by less rigid requirement of the consumer to quality of controllable production $\beta = 0,1$.

Value of a rejection degree of quality q_m should be chosen, proceeding from a boundary degree of quality which represents as much as possible admissible share of defective products in party.

Thus, **value** of a rejection **degree of quality** q_m **should not exceed value of a boundary degree of quality.** Or on the contrary, on the value of a boundary degree of quality calculated by data from operation to establish value of rejection level equal to it q_m**.**

The choice of a variant of rejection demands separate consideration and in the given work is not resulted.

Let's consider an example of scoping of sample for control of a lot of products in the size $N = 165$ in which it has been while in service revealed $D = 25$ defective for the industrial reasons. For sample scoping we will calculate value of boundary level

$$q_m = \frac{D}{N} \cdot 100\% = \frac{25}{165} \cdot 100\% = 15\%. \tag{1}$$

It is obvious that $q_m > \beta \cdot 100\%$**.** Therefore selective control for party with such level of defective products it agree GOST 16493-70 [3] it is not provided.

According to later GOST P 50779.52-95 [2] account of operational reliability of products by working out (updating) of the plan of acceptance control is carried out by the task of an estimation of expected actual (entrance) level of discrepancies in the shown party (the next party from sequence of parties). Thus recognise that if by the time of creation of the plan the estimation of actual (entrance) level of discrepancies of the party which have arrived on control is known, use it. In the absence of such given by an expert method choose an interval of this estimation and the corresponding plan. Further this plan is necessary for specifying periodically on the basis of the subsequent estimations.

Under table 1 [2] at value of risk of the consumer $\beta = 0,1$ trust degree to the supplier is established T2. From table 4 at volume of party N from 151 to 280 we choose plan A.21 and the scheme for normal control A.81.

According to plan A.21 and scheme A.81 at level of discrepancies by data from operation (1) and to standard level $NQL = \beta \cdot 100\% = 10\%$ there is a situation, when $q_m > NQL$ in which any of admissible plans, including continuous control, does not provide high probability of acceptance.

Thus, according to operating GOSTs the party with such level of defects (discrepancies) should be exposed to continuous control.

By the way, for a considered case of inadmissibility of defective products in sample ($d = 0$) to the cores is GOST 16493-70 [3] which urgency is confirmed by its reprinting in 2011. In this document of the table are made with the minimum value of volume of sample $n = 20$. Believing that in standards it is impossible to provide all variants of control, we will solve a problem analytically in the following statement.

From the party in the size N containing according to operation D of defective products, the sample in volume n taken without returning is taken n. All products from this sample are exposed to control check about revealing of discrepancies (defects). Let in sample it has appeared d defective products. Then distribution of probabilities of number of occurrence of defective products in sample is described by hypergeometrical distribution with function [4]

$$P\{d(n) = d\} = \frac{\begin{pmatrix} D \\ d \end{pmatrix}\begin{pmatrix} N-D \\ n-d \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}, \tag{2}$$

Where $\begin{pmatrix} a \\ b \end{pmatrix} = \frac{a!}{b!(a-b)!}$.

For the plan of type of unitary sample in case of inadmissibility of defective products in it $d = 0$. Probability of display at least one defect

$$Q = 1 - P\{d(n) = 0\} \tag{3}$$

In sample of the decreasing volume, since some moment, decreases (Figure 1) see.

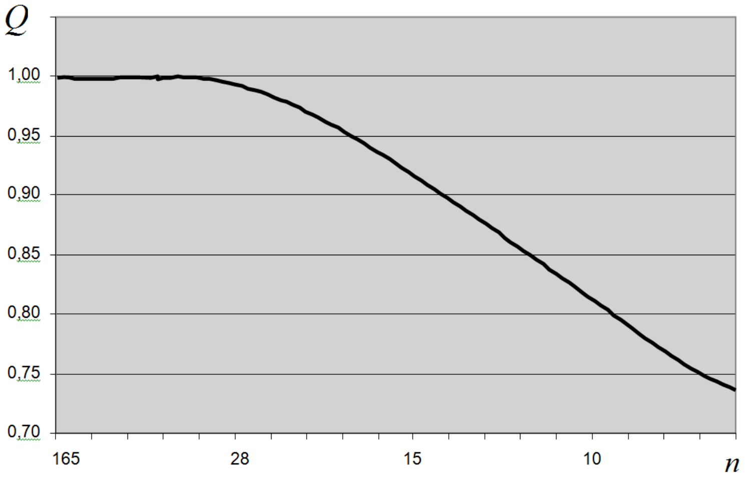


Figure 1: Probability to "catch" at least one defective product in sample. Physically it means that *the the smaller number of products from party is exposed to control, the it is less probability to reveal presence defective in it.*

Having set by level $\beta = P\{d(n) = 0\} = 1 - Q$ which in acceptance control is treated as risk of the consumer [2], it is possible to define sample volume n, at check of products from which the defective should not be. In particular, at $\beta = 0,1$ sample volume should be not less $n = 14$. It means that for party in the size $N = 165$ of products from which in operation it has been revealed $D = 25$ defective, control should subject casually chosen one product from everyone 12. Thus, if at least one of the products which are exposed to control, appears defective all party (a rejection variant in the given problem is rejected is not considered).

Let's consider the decision of the same problem with use of elements of the theory of stochastic similarity [5]. Factor of stochastic similarity of control sample of checked party

$K = \frac{Q_l^1}{Q_N^1}$,At, $Q_l^1 < Q_N^1$ (4)

Where indexes at Qdesignate a variant of the plan of control (to control one product casually chosen from each l products of sample) is exposed l. It is obvious that sample volume in this case ($n = \frac{N}{l}$ with a rounding off to the whole value).

The schedule of this function is presented in the Figure 2 which step character is caused by step-type behaviour of hypergeometrical distribution. The analysis of Figure 2 shows that with reduction of volume of sample (n it is defined by a variant of the plan of control "1 of l", i.e. with increase l at the fixed size of party N) the factor of stochastic similarity decreases. Physically it means that the samples generated thus with increase l reflect stochastic essence of controllable party ever less.

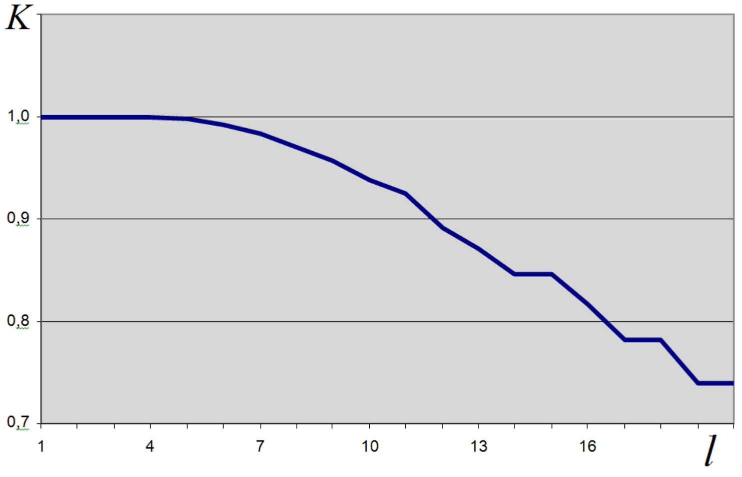


Figure 2

Having put that the size in K expression (4) is casual, for a substantiation of value of the size l characterising a variant of the plan of control, it is possible to use a consequence of postulation and hypothesis check about stochastic similarity considered samples. For this purpose it is necessary to know the law of distribution of a random variable Q. According to a lemma [the 6] random variable Q representing probability, is in regular intervals distributed in an interval [0; 1]. Then distribution of a random variable K grows out of the following theorem.

The *theorem. The* relation from two independent random variables in regular intervals distributed in an interval [0; 1], there is a random variable in regular intervals distributed in an interval [0; 1].

The *proof.* Let two independent random variables and x_1x_2, in regular intervals distributed in an interval [0 are given x_1x_2; 1]. The density of their joint distribution after streamlining ($x_1 \leq x_2$) looks as follows [8,9]

$$f(x_1,x_2)=2!f_{x_1}(x_1)f_{x_2}(x_2)=2, \tag{4}$$

Where and $f_{x_1}(x_1)=1$ – $f_{x_2}(x_2)=1$ density of distribution of independent random variables and x_1x_2.

The density of distribution of a random variable $y=\frac{x_1}{x_2}$ looks like [7]

$$g(y)=-\int\limits_{-\infty}^{0}x_1f(x_1,yx_1)dx_1+\int\limits_{0}^{\infty}x_1f(x_1,yx_1)dx.$$

Having substituted in this expression density of joint distribution (4), and, having rejected the first integral as $x_1\in[0;1]$, we will receive

$$g(y)=\int\limits_{0}^{1}x_1f(x_1,yx_1)dx_1=\int\limits_{0}^{1}2x_1dx_1=2\cdot\frac{x^2}{2}\Bigg|_{0}^{1}=1\,. \tag{5}$$

Expression (5) corresponds to density of uniform distribution of the relation y. The theorem is proved.

Now, proceeding from justice of the postulated hypothesis about stochastic similarity considered samples with probability $(1-\alpha$for equal risks of the supplier and the consumer $\alpha=\beta=0,1)$, it is possible to solve a return problem of definition of the size $l$characterising a variant of the plan of control. For this purpose it is enough to resolve the equation (4) rather graphic l (Figure 2 see) or numerical way. It is possible to show that for the set conditions $l=12$. It means that one product in a random way taken from each 12, making checked party should be exposed to control. In this case from party in the size $N=165$ sample in volume $n=14$ of products, in the absence of defects in which will be generated $n=14$, the party is considered accepted.

It is easy to be convinced that the received result coincides with the decision under formulas (2) and (3) under a condition in $Q_N^1=1$ expression (4). It is natural, since the probability of occurrence at least one defective product in party with $D=25$ defects is authentic. If necessary, when $Q_N^1\neq1$, it is possible to extend the offered approach and for an estimation of stochastic similarity samples, generated on different variants of control among themselves, instead of with party. For the same reason the form of curves in $Q(n)$ Figure 1 and in $K(l)$ Figure 2 coincides.

Let's consider influence of operational reliability on necessity of updating of a variant of the plan of control. For this purpose in Figure 3 dependences of factors of stochastic similarity samples, generated on various variants of control of the same party in volume $N=165$ in which operation the various number D of defective products has been revealed are presented $N=165D$.

The analysis of Figure 3 shows that **more reliable products are necessary for supervising more often** (curves are displaced towards reduction of number l

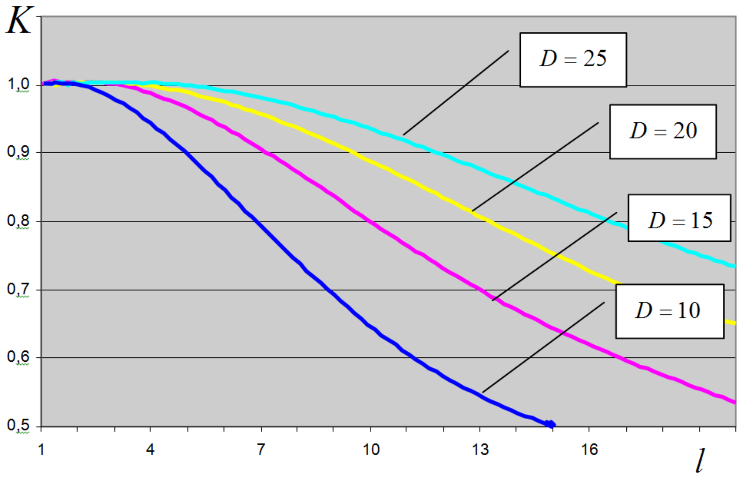


Figure 3: Influence of operational reliability on variant of the plan of control

of products of which one gets out for control). Rather the reverse, as unsophisticated experts in the statistican imagine, believing that in highly reliable products defect is shown seldom, and expenses for control can be saved.

Really, in Figure 4 the function $l(D)$ corresponding to variants of control at factor of similarity samples of controllable party is presented $l(D)K = 1-\alpha = 1-0,1 = 0,9$. Apparently is almost linear in the given range of change D dependence.

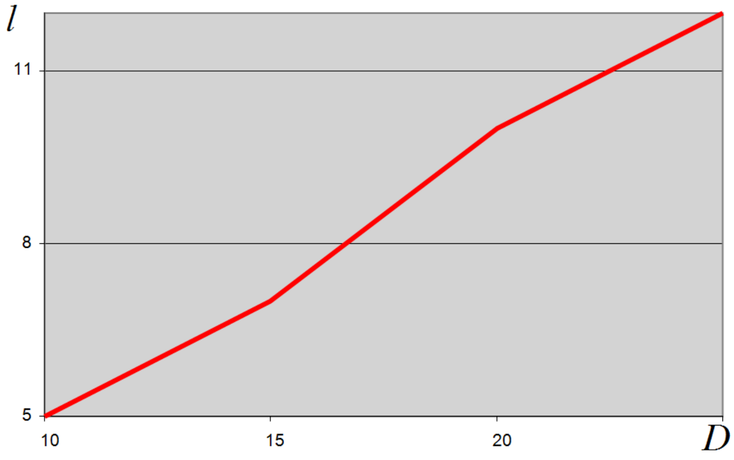


Figure 4: Dependence of a variant of control (size$l$) from number the defects D revealed in operation of a lot of products

For this function that is characteristic (as it is paradoxical sounds) that for maintenance of stochastic similarity of control sample of checked party (i.e. that sample really reflected the maintenance of defective products in party) at increase in level of discrepancies in q it (shares of defective products) volume of control sample it is possible to reduce (increase$l$) in comparison with earlier appointed. And on the contrary, if in operation reduction of number D of the revealed defects in party (decrease in level of discrepancies q) the volume of control sample is necessary for increasing is registered Dq. That is more reliable products demand special attention (because of a rarity of display of defects), therefore for authentic revealing of discrepancies it would be required..

From here there is the economic nuance, consisting that high-quality products costly not only to make, but also to supervise their quality. Therefore GOST considers this feature of acceptance control and in table 1 [3] the nearest value for level (1) is $q_m = 10\%$. On set ($\beta = 0,1$ there corresponds to a column And tables) in a line "122 and more" corresponding to the size of party $N = 165$, value of volume of sample $n = 25$ that corresponds to a variant of the plan of control “1 of 6”.

At decrease in level to ($q_m = 8\%$decrease in number of defective products in party) at the same size of party in $N = 165$ line “138 and more” value of volume of sample $n = 40$ that corresponds to a variant of the plan of control “1 of 4”.

That is, the improvement of quality of the manufacture which have led to decrease of number of defective products in party by results of operation, demands increase of control of products. Differently: it is necessary to confirm display of more rare events with the increased volume of the statistical data.

Contrary to this logic, in practice often arrive differently: in the absence of the information on defects in party (i.e. quality high) some heads, aspiring to benefit, reduce volume of control sample, i.e. carry out acceptance control less often. For example, change a control variant “1 product from 5” to a variant “1 product from 10” on the ground that in one let out party of the state order there were no defective products. Here so! On one party – at once twice to reduce sample volume. Be accepted, such decision would lead to that for considered party, agree tables 1 [3], reduction of volume of sample in 2 times would raise rejection level in q_m 2,5 times. It would mean that for the same party instead of registered in operation 25 defective products because of industrial deviations at a variant of acceptance control “1 product from 5” could be passed to 62 defective products at a control variant “1 product from 10”.

Not propaganda for performance standards GOST, and display of necessity of the account of operational reliability of products at the organisation of system of acceptance control of the enterprise was the purpose of the given work. Without understanding of essence of selective control and adding elements of the theory of stochastic similarity in it situations of acceptance of erroneous decisions are possible. And for responsible products such decisions are fraught with infringement of their safe operation.

References

[1] GOST P 8550-1-2007 Statistical methods. A management at a choice and application of systems of statistical acceptance control of discrete units of production in parties. A part 1. The general requirements.

[2] GOST P 50779.52-95 Statistical methods. Acceptance quality assurance to an alternative sign.

[3] GOST 16493-70 Quality of production. Statistical acceptance control to an alternative sign. A case of inadmissibility of defective products in sample.

[4] GOST P 50779.10-2000 (ISO 3534-1-93) Statistical methods. Probability and statistics bases. Terms and definitions.

[5] Reliability and efficiency in the technician: the Directory: T. 4.: similarity Methods in reliability/under V.A.Melnikova, N.A.Severtseva. – M: Mechanical engineering, 1987. – 280p.

[6] Gnedenko B. V, Beljaev J.K., A.D.mathematical's Nightingales methods in reliability theory. – M: the Science, 1965. – 523p.

[7] Venttsel E.S., Ovcharov L.A.applied of a problem of probability theory. – M: Radio and communication, 1983. – 416p.

[8] N.Johnson, F.Lion. Statistics and experiment planning in the technician and a science. Data processing methods. – M: the World, 1980. – 610p.

[9] Volovik A.V., Klavdiev A.A., Klavdiev I.A., etc. Estimation of stochastic similarity of objects with casual parametres of difficult technical systems. – the mountain information-analytical bulletin No 6, 2014. – M: Publishing house "Mountain book". – 2014. – 432p.

Extreme Statistics

A.A. Klavdiev[1], A.A. Sulima[2], S.V. Efimenko[1]

[1] *The national mineral resource of the University "Mining", St. Petersburg, Russia*

[2] *Mikhaylovsky artillery academy, St. Petersburg, Russia*

e-mail: kss1959@mail.ru, falcon.sergey@yandex.ru

Abstract

In article safety of operation of difficult technical systems is considered. By the present moment in the theory of reliability the subject, the purpose, methods and research problems were created, but in the theory of safety they are in a development stage. The theory of adoption of statistical decisions on small number of supervision still needs now scientific justification and development. Complexity of statement and the solution of problems of creation of the best estimates at this volume of statistical material is caused by that the required decision in strong degree depends on concrete type of distribution, the volume of selection and can't be object of rather general mathematical theory.

Keywords: difficult technical systems, theory of reliability, theory of adoption of statistical decisions, distribution type, selection volume.

Introduction

The safe operation of complex engineering systems largely depends on the reliability of their constituent elements. The development of the theory of security as a scientific on-Board is similar to the recognition and establishment of reliability as a science. And, if so far in the theory of reliability has formed the subject matter, purpose, methods and objectives of the research, it is safe they are in development. This state of teaching caused a relatively "young age" of this research area, the feature information and the need to develop and unsaturated extremal-tion methods for safety assessment.

Feature information security is characterized by the notion of an event, the occurrence of which is the danger. For example, in aviation, in the classification of failures on the implications of concepts are used:
aviation incident (accident, accident);
the incident (non-localized failures, fires, shut down the engine in flight, failures of components and systems that do not have duplication);
emergency landing, etc.

The investigations have to spend in the face of very limited information on the so-called "tail" of distributions. This implies two aspects of the problem of safety assessment:
for a quantitative assessment is necessary to develop methods that provide the possibility of such assessment on small number of observations with the required accuracy;
for interval estimation methods necessary, adequately describing the marginal region of distributions based on limited information.

Generally speaking, statistical information on accidents and incidents are very limited and heterogeneous, because of the events leading up to them, as a rule, solitary or rarely repeated. Therefore, the safety assessment is currently being carried

out on a qualitative level, because for the quantification mathematical apparatus information is not enough.
The theory of making statistical decisions on small number of observations, for many tasks which are typical leasimpresa posing the problems currently still requires scientific substantiation and development. The complexity of formulating and solving the challenges of building the best estimates in this statistical volume of material is due to the fact that the desired solution is often to a great extent depends on the specific type of distribution, sample size and cannot be the object of a fairly General mathematical theory.

It is considered that the beginning of the theory of small samples was initiated in the first decade of the twentieth century the publication of the work W. Gosset, in which he placed the t-distribution. At the time, Gosset worked as a statistician in Breweries. He experimented with the idea of a significant reduction in the number of samples taken of a very large number of barrels, stocks brewery, to selectively control the quality of porter. In the end, he has published the results of their study compared a sample of quality control using the t-distribution for small samples and traditional z-distribution (normal distribution) anonymously, under the pseudonym "Student" (Student - hence the name t-student distribution).

At the time the question was raised about how much should be sampled, so that it can be considered small. A definite answer to this question simply does not exist. However, the conventional boundary between small and large samples considered to be n=30. The reason for this is to some extent an arbitrary decision is the result of the comparison t-distribution with the normal distribution. A simple visual inspection of the tabular value of t allows us to see that this approximation is quite fast, starting with n=30 and above. Therefore, the sample volume of less than 30 observations are small.

However, the statistics of accidents and incidents operates in much smaller volumes. It is, literally, about a few cases. In these conditions requires neosynthesis methods based on extreme distributions. That is, pardon the pun, extreme conditions, dictate the use of extreme methods to estimate distributions of extreme values of random variables. All this, it seems, should be the subject of study of extreme statistics.

For awareness of the problem it is useful to consider the statement and the possible way of solving one of the classical problems in the theory of reliability, which has independent significance in evaluating the safety of objects. In the operation of aircraft of a particular type occur failures (failure), leading to accidents. For the entire period of observation, there are no more than 3...5 cases, but the impact is significant (i.e. affecting safety). It is necessary with a given confidence probability to estimate the boundary of the safe operation of the product.

As can be seen from the statement of the problem the extremity of the conditions is that the sample size does not allow to count on acceptable from the viewpoint of reliability, the solution of the classical method based on the marginal distributions. Most preferred in this case is information approach using the principle of maximum uncertainty (the principle of Jaynes), based on consideration of the Shannon entropy.

This approach is the least sensitive to the initial assumptions, and in General allows to take into account any number of available information [3].

The formalism of the principle of maximum uncertainty (maximum entropy) postulates that the least questionable representation of the probabilities will be such a representation that maximizes the uncertainty in the light of all given information. In this case, the entropy serves as a measure of uncertainty. The essential difference of the maximum principle of uncertainty is the possibility of obtaining a priori estimates of the distribution of information in situations for which there are various restrictions in the form of a probability measure, a separate torque characteristics, etc., in the form of equalities and inequalities. From a mathematical point of view, using the principle of maximum uncertainty, the task of such restrictions leads to the solution of classical and non-classical optimization problems (extremum problems). The basis for analysis was the empirical observation. Consider the empirical density distributions of the smallest (extreme) values in samples of different size n obtained by simulation modeling of a General population with exponential distribution law. Their flattened view is presented in figure 1.

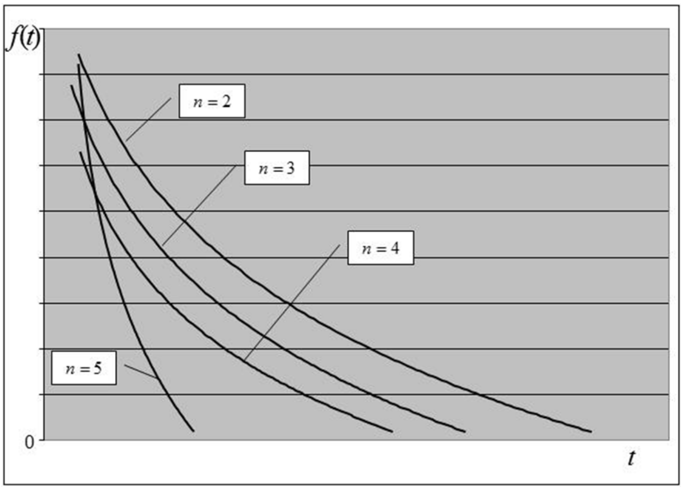


Figure 1: The Empirical density distribution of the lowest the values in the sample

It is easy to see that with increasing sample size, ceteris paribus, the smallest distribution of a random variable is shifted to the y-axis.

Theoretical justification of the issue are the following considerations. In General case, the distribution function of the smallest value in the sample size n has the form [1]

$$F_{t_{\min}}(t_{\min}) = 1 - [1 - F(t_{\min})]^n \tag{1}$$

and the density, respectively

$$f_{t_{\min}}(t_{\min}) = n\,[1 - F(t_{\min})]^{n-1} \cdot f(t_{\min}) \tag{2}$$

where $F(\cdot)$ and $f(\cdot)$ is the function and the density of the initial distribution. Then the density distribution of the smallest value in the sample from exponentiexponentially distributed General population can be written as follows

$$f_{t_{\min}}(t_{\min}) = n\lambda e^{-n\lambda t_{\min}} \tag{3}$$

where $\lambda = \frac{1}{T}$ – the parameter distribution;

T – the mathematical expectation of a random variable t.

Physically formula (3) means that the smallest value in the sample is shown with intensity proportional to the sample size.

Graphs of the theoretical density functions (3) are presented in figure 2.

Analysis of figures 1 and 2 shows the identity of the nature of the empirical and theoretical distributions of the lowest values in samples from a General population with exponential distribution law. Consider the quantile function of the distribution of the smallest value for the sample from an exponential population. To do this, in the expression (1) substitute the function of the exponential distribution and find the inverse function of quantiles

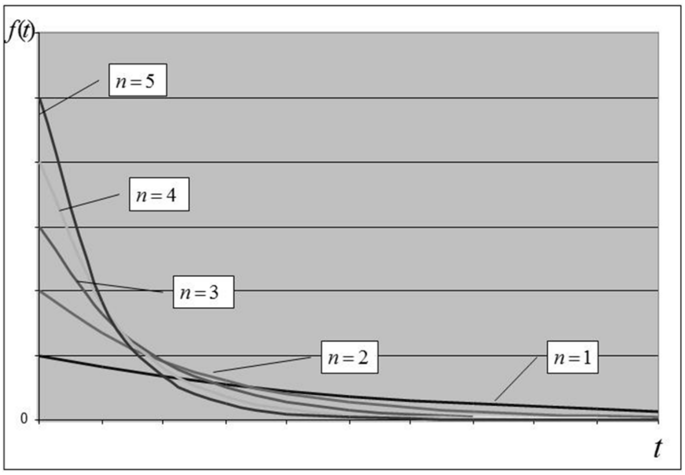


Figure 2: Theoretical density distribution of the lowest the values in the sample

$$T_{\min} = -\frac{T}{n}\ln(1-\alpha) \tag{4}$$

The graph of this function depending on n are depicted in figure 3.

The nature of the influence of the sample size together with the level of significance α is illustrated by figure 4.

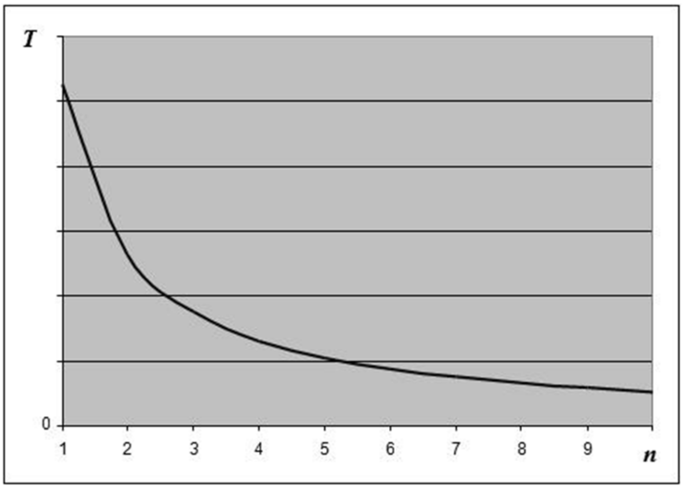


Figure 3: The Function of the quantiles of the distribution of the smallest value of the sample from exponential General population

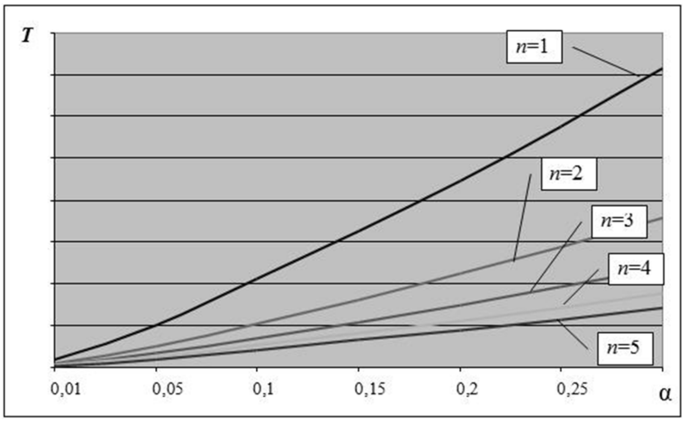


Figure 4: Dependence of the quantile function of the sample size

Analyzing figures 3 and 4, we can conclude that, as befits a rational function of the form
quantile function (4) has asymptomatic properties if $n > 10$.

When the sample sizes $n < 5$ not asymptomaticity dependence (4) becomes noticeable. Especially significant is the region of inflection of functions of quantiles $n \in [2, 5]$. As for the specific point $n = 2$, a particular point, it should be noted that it lies on the edge of the range $n \in [1, 2]$, where sufficient linearity of the function (4) gives reason to resort to approximations.

To do this, using the principle of maximum uncertainty (maximum entropy), one can show that the asymptotic representation of the function of quantiles in the case when the initial distribution is known only to the mathematical expectation T of a random variable has the form [3]

$$T_{\min} = \frac{nT}{n-1}\left[1 - (1-\alpha)^{\frac{n-1}{n}}\right] \tag{5}$$

Is of practical interest, the consideration of the asymptotic properties of the functions and the density function of a random variable $T_{\min}$ Thus, by substitution $F(T_{\min}) = \alpha$ in (5), we have distribution function

$$F(T_{\min}) = 1 - \left(1 - \frac{n-1}{nT}T_{\min}\right)^{\frac{n}{n-1}} \tag{6}$$

differentiating which, it is easy to deduce the density of the desired distribution

$$f(T_{\min}) = \frac{1}{T}\left(1 - \frac{n-1}{nT}T_{\min}\right)^{\frac{1}{n-1}} \tag{7}$$

Let us show, that functions (6) and (7) are normalized in interval

$$T_{\min} \in \left[0; \frac{n}{n-1}T\right]$$

In the particular case, when the minimum sample volume , density (7) is a linear dependence of

$$f(T_{\min}) = \frac{1}{T} - \frac{1}{2T^2}T_{\min}$$

and the distribution function is quadratic

$$F(T_{\min}) = 1 - \left(1 - \frac{1}{2T}T_{\min}\right)^2$$

In the limit (when) density (7) takes the following form

$$f(T_{\min}) = \frac{1}{T}$$

and distribution function

$$F(T_{\min}) = \frac{T_{\min}}{T} \tag{8}$$

what is asymptotically corresponds to the uniform distribution.

The distribution of the smallest value in the sample in General bilateral. However, in terms of solved problems we are interested in the left border, which characterizes the smallest (extreme) random variable. Therefore, we believe $F(T_{\min}) = \frac{\alpha}{2}$ Then for uniform distribution law (8) the expression is true

$$\hat{T}_H = \frac{\alpha}{2}T \tag{9}$$

which with sufficient accuracy for practice approximates the dependence (4) in the range

$$\alpha \in [0; 0,2]\,.$$

Theoretical values under conditions from (6) have the following form

$$T_H = 2T\left(1 - \sqrt{1 - \frac{\alpha}{2}}\right) \tag{10}$$

Figure 5 presents the theoretical and asymptotic functions of quantiles of the distribution of the smallest random variable from the sample exponentially distributed Noi General population. There is also shown the discrepancy of their $\Delta T_H = T_H - \hat{T}_H$, for clarity, multiplied by 10.

It is easy to see that the differenceΔT_H between the theoretical and asymptotic values for the minimum sample volume $n = 2$ in almost range used $\alpha \in [0; 0,2]$ is not more than 3%.

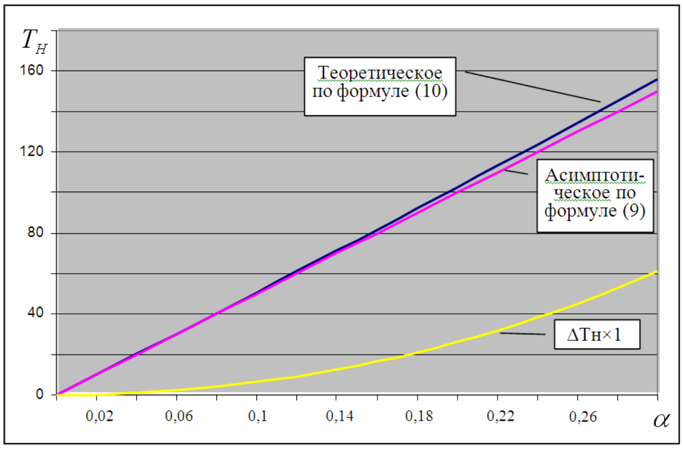


Figure 5: Theoretical Graphs and asymptotic functions quantile

Thus, the analysis did not reveal contradictions in the representation of the function of the quantiles of order statistics in the form of asymptotic dependence (9) theoretical positions and the results of the computational experiment. He confirmed the sufficient accuracy of the method for extreme case sample size $n = 2$.

The results of the study can serve as a basis for the practical use of the proposed approach when confirming the safety and reliability of highly reliable products in extreme cases, sampling, when the nature of the distribution is not known nothing but the mathematical expectation of a random variable.

References

[1] N. Johnson, F. Lyon. Statistics and experimental design in engineering and science. Methods of data processing. - M.: Mir, 1980.

[2] Korn G., Korn T., mathematical Handbook for scientists and engineers. - M.: Nauka, 1984. - P. 635.

[3] Ivchenko, B. P., Martyshchenko L. A., Tabakov M. E. Governance in the economic and social systems. System analysis. Decision making under uncertainty. - SPb.: “Normed-Izdat”, 2001. – P. 248

Numerical Probabilistic Approach for Data Nonparametric Analysis

Boris S. Dobronets and Olga A. Popova
Siberian Federal University, Krasnoyarsk, Russia
e-mail: BDobronets@yandex.ru, OlgaArc@yandex.ru

Abstract

The paper considers an approach allowing to build reliable estimates of the cumulative distribution function. The approach is based on smoothing an empirical cumulative distribution function using numerical probabilistic analysis. We propose the method to estimate the reliability indices of technical systems in conditions of limited information.

Keywords: Reliable estimates, empirical cumulative distribution function, numerical probabilistic analysis.

Introduction

In order to solve many practical problems need to know the reliable estimation of the distribution function, constructed under conditions of small amount of statistic data.

Monte Carlo method [12] is a powerful approach, but it has some serious shortcomings. These are difficulties in handling uncertain quantities having unknown dependency relationships or those with imprecise probabilities, that is, with not fully specified distributions.

Non-Monte Carlo methods have been developed since 1960s [2, 13]. A major non-Monte Carlo approach is interval analysis [11, 3].

In our work, we develop a technique that uses Numerical Probabilistic Analysis (NPA) to solve various problems with stochastic data uncertainty. The basis of NPA is numerical operations on probability density functions of the random values. These are operations "$+$", "$-$", "$\cdot$", "$/$", "$\uparrow$", "max", "min", as well as binary relations "$\leq$", "$\geq$" and some others. The numerical operations of the histogram arithmetic constitute the major component of NPA [8].

The aleatory uncertainty characterizes the inherent randomness in the behavior of the system under study. On the contrary, the epistemic uncertainty characterizes a lack of knowledge about a considered value. Generally, the epistemic uncertainty may be inadequate to the frequency interpretation, which is typical for classical probability and for uncertainty description in the traditional probability theory [7].

A probability box (or P-box) is a characterization of epistemic uncertainties that is often used in risk analysis or quantitative uncertainty modeling where numerical calculations must be performed [9, 10]. Second order histograms are alternative of P-boxes in the case of epistemic uncertainty [6].

1 Reliable Estimate Empirical Cumulative Distribution Function

In this section we look at ways to build reliable estimates of the cumulative distribution function (CDF). Let $(x_1, \ldots, x_n)$ be real random variables with the common cumulative distribution function $F(t)$. Then the empirical distribution function F_n is defined as

$$F_n(t) = \frac{m_t}{n}. \tag{1}$$

where m_t is the number of elements $x_i < t$.

Let $z_i = F(x_i), i = 1, \ldots, n$. Note that $z_i, i = 1, \ldots, n$ are uniformly distributed random variables. If $z_1 \leq z_2 \leq \ldots \leq z_n$ then expected value M$[z_i]$= $i/(n+1)$. Next, we use the points $(x_i, i/(n+1)$ for the construction of the approximation cumulative distribution function $F(t)$. For these purposes, we use splines s.

Let $a = x_0 < x_1 < x_2 < \ldots < x_n < b = x_{n+1}$ be mesh.

Suppose that interpolation conditions are

$$s(x_i) = i/(n+1), \quad i = 1, \ldots, n, s(a) = 0, s(b) = 1,$$

and let boundary conditions are

$$s'(a) = 0, s'(b) = 0.$$

Note that if instead of $i/(n+1)$ to use their exact values z_i, then for example cubic spline on a mesh $\{x_i\}$ with step $h = \max(x_{i+1} - x_i), i = 0, \ldots, n$ satisfies the estimate

$$||F^\nu - s^\nu|| \leq h^{4-\nu}||F^{(4)}||, \quad \nu = 0, 1, 2.$$

Thus, even with a small n, you can build a fairly accurate approximation for F. The task of building the spline is reduced to solving a system of linear algebraic equations with a tridiagonal matrix [1]

$$\lambda_j m_{j-1} + 2m_j + \mu_j m_{j+1} = d_j, \tag{2}$$

$$2m_0 + m_1 = 3(z_1 - z_0)/h_1 - h_1 z_0^2/2,$$

$$2m_N + m_{N-1} = 3(z_N - z_{N-1})/h_N + h_N z_N^2/2,$$

$$d_j = 3\lambda_j(z_j - z_{j-1})/h_j + 3\mu_j(z_{j+1} - z_j)/h_{j+1}, \quad j = 1, ..., N-1.$$

where $m_i = s'(x_i)$.

The matrices of these systems are deterministic and right-hand sides contain the random variables. Thus, by virtue of a deterministic matrix, the solution m_i $i = 0, \ldots, N$ can be represented as a linear combination of the elements of right side.

As a result, a cubic spline on the intervals $[x_{j-1}, x_j], j = 1, ..., N$ has the representation [1]:

$$s(x) = m_{j-1}(x_j - x)^2(x - x_{j-1})/h_j^2 - m_j(x - x_{j-1})^2(x_j - x)/h_j^2 +$$

$$+z_{j-1}(x_j - x)^2(2(x - x_{j-1}) + h_j)/h_j^3 + +z_j(x - x_{j-1})^2(2(x_j - x) + h_j)/h_j^3, \qquad (3)$$

Let p_z be joint density distribution of the vector $(z_1, z_2, \dots, z_n)$. Then replacing z of their joint probability density and using a numerical probability analysis we obtain estimates of the probability density for the components of the m_i and build a probabilistic extension of CDF. Probabilistic extension of CDF can be represented in the form of histogram P-box and then used for numerical analysis [9].

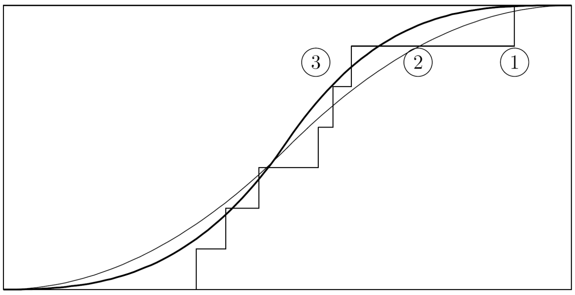


Figure 1: Smoothing the empirical CDF

Figure 1 shows an example of smoothing the empirical CDF of the sample dimension $n = 7$ distributed over a triangular law on the segment [0,2], with the vertex (1,1). Where line (1) is empirical CDF, line (2) is the exact CDF, line (3) is smoothed CDF.

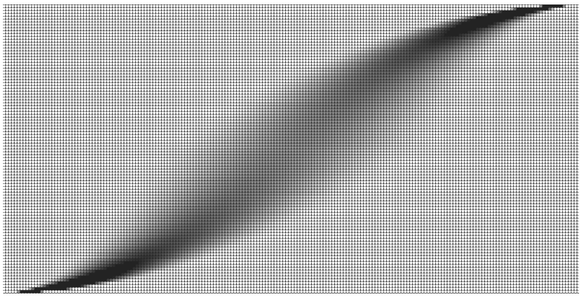


Figure 2: Probabilistic extension of CDF

Figure 2 shows the histogram P-box where the values of probability densities are shades in gray.

For some random functions we introduce the following concepts. Let the random function has the form

$$f(x) = \sum_{i=1}^{n} a_i g_i(x),$$

where a_i are random constants, $g_i \in C^m[a, b]$. Then the derivative of $f(x)$ defined in such a way:

$$\partial^k f(x) = \sum_{i=1}^{n} a_i g_i^{(k)}(x), \quad k = 0, ..., m.$$

So, differentiating probabilistic extension of CDF we get probabilistic extension of the probability density function.

2 Numerical Probabilistic Analysis

NPA is based on numerical operations on probability density functions of the random values and probabilistic extensions. Represent types of probability density function in details. Density function can be a discrete function, a histogram (piecewise constant function), and a piecewise polynomial function.

Using the arithmetic of probability density functions and probabilistic extensions, we can construct numerical methods that enable us solving systems of linear and nonlinear algebraic equations with stochastic parameter [4]. To facilitate more detailed description of the epistemic uncertainty, we introduce the concept of second order histograms, which are defined as piecewise histogram functions [5]. The second order histograms can be constructed using experience and intuition of experts.

Histograms. The histogram is called a random variable density which is represented piecewise constant function. Histogram P is defined grid $\{x_i|i = 0, \dots, n\}$, on each interval $[x_{i-1}, x_i]$, $i = 1, \dots, n$ histogram takes constant value of p_i.

Second order histogram. Next, we consider construction of a second order histogram in the case of epistemic uncertainty. Suppose that we have a series of histograms $\{Y_i, i = 1, 2, \dots, N\}$. Each Y_i assign a probability p_i: $\sum_i^N p_i = 1$. For simplicity, we assume that all the histograms Y_i are defined on the mesh $\{z_i, i = 0, 1, \dots, n\}$. On the interval $[z_{k-1}, z_k]$ histogram Y_i takes the value Y_{ik}. Thus, on each interval $[z_{k-1}, z_k]$, we have a random variable Y_k with values Y_{ik} and probability p_i. Using these values, we can on each interval $[z_{k-1}, z_k]$ restore histogram P_{zk}.

Operation on histograms. Let $p(x, y)$ be a joint probability density function of two random variables x and y. Let p_z be a histogram approximating the probability density of the arithmetic operations on two random variables $x * y$, where $* \in \{+, -, \cdot, /, \uparrow\}$. Then the probability to find the value z within the interval $[z_i, z_{i+1}]$ is determined by the formula

$$P(z_k < z < z_{k+1}) = \int_{\Omega_k} p(x, y) dxdy,$$

where $\Omega_k = \{(x, y)|z_k \leq x * y \leq z_{k+1}\}$.

This approach is generalized to a larger number of variables. Let us required to find a histogram p_z of sum

$$z = a_1x_1 + a_2x_2 + ... + a_nx_n$$

and let $p(x_1, x_2, ..., x_n)$ be density distribution the probability of a random vector $(x_1, x_2, ..., x_n)$. Then

$$P(z_i < z < z_{i+1}) = \int ... \int_{\Omega_i} p(x_1, x_2, ..., x_n)dx_1dx_2....dx_n,$$

where $\Omega_i = \{(x_1, x_2, ..., x_n)|z_i < a_1x_1 + a_2x_2 + ... + a_nx_n < z_{i+1}\}$.

Probabilistic extensions. One of the most important prblems that NPA deals with is to construct probability density functions of random variables. Let us start with the general case when $(x_1, \ldots, x_n)$ is a system of continuous random variables with joint probability density function $p(x_1, \ldots, x_n)$ and the random variable z is a function $f(x_1, \ldots, x_n)$

$$z = f(x_1, \ldots, x_n).$$

By *probabilistic extension* of the function f, we mean a probability density function of the random variable z.

Let us construct the histogram F approximating probability density function of the variable z. Suppose the histogram F is defined on a mesh $\{\, z_i \mid i = 0, \ldots, n \,\}$. The region is denoted as $\Omega_i = \{(x_1, \ldots, x_n) | z_i < f(x_1, \ldots, x_n) < z_{i+1}\}$. Then the value F_i of the histogram on the interval $[z_i, z_{i+1}]$ is defined as

$$F_i = \int_{\Omega_i} p(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n / (z_{i+1} - z_i). \tag{4}$$

By *histogram probabilistic extension* of the function f, we mean a histogram F constructed according to (4).

Let $f(x_1, \ldots, x_n)$ be a rational function. To construct a histogram of F, we replaced the arithmetic operation by the histogram operation, while the variables x_1, x_2, …, x_n are replaced by histogram of their possible values. It makes sense to call the resulting histogram of F as *natural histogram extension* (similar to “natural interval extension”).

3 Estimates of Failure Rates

The probability of failure-free operation $P(t)$ is likelihood that within the specified operating time there is no failure. Operating time is the duration or the amount of work

The failure rate is measure of failure per unit of time. The failure rate depends on a failure distribution, which is a cumulative distribution function that describes the probability of failure prior to time t,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \le \xi < t + \Delta t | t \le \xi)}{\Delta t} = \frac{f(t)}{P(t)} = \frac{f(t)}{1 - F(t)}.$$

Note that $f(t) = P'(t)$ and

$$\lambda(t) = -\frac{P'(t)}{P(t)}. \tag{5}$$

Consider a numerical simulation of failure rate. Let $(\xi_1, \xi_2, \ldots, \xi_n)$ be statistics failures obtained empirically. Then

$$-\ln(z_i) = \int_0^{\xi_i} \lambda(\xi) d\xi,$$

where $z_i = P(\xi_i)$. For to find $\lambda(t)$ we use the method of least squares. Let $\varphi_1, \varphi_2, \dots, \varphi_m$ be linearly independent functions and $\lambda(t)$ introduce as

$$\lambda(t) = \sum_{i=1}^{m} a_i \varphi_i(t).$$

To find the unknown coefficients $a_1, a_2, \dots, a_m$ consider the functional

$$\Phi(a_1, \dots, a_m) = \sum_{i=1}^{n} (z_i - \sum_{j=1}^{m} a_j \varphi_j(\xi_i))^2 \to \min .$$

The problem is reduced to solving a system of linear algebraic equations

$$A\vec{a} = b,$$

where $A = (a_{ij})$ is the Gram matrix, $\vec{a} = a_1, a_2, \dots, a_m$, $b = (b_i)$, $a_{ij} = (\varphi_i, \varphi_j)$, $b_i = (z, \varphi_i)$ and

$$(x, y) = \sum_{i=1}^{n} x(\xi_i) y(\xi_i).$$

Using instead of $z_1, z_2, \dots, z_n$ their joint probability density function $p(z_1, z_2, \dots, z_n)$, we can construct probabilistic extensions of $\lambda(t)$.

Model example. Given a resampling developments failures $\{0.0155, 0.0389, 0.2855, 0.5318, 0.7412, 1.0118, 1.1267, 1.2327, 1.8594.\}$. Suppose $\lambda(t)$ has the form

$$P(t) = \exp(-\int_0^t a_1 + a_2 x^2 dx),$$

and $a_1 = 1$, $a_2 = 0.3$.

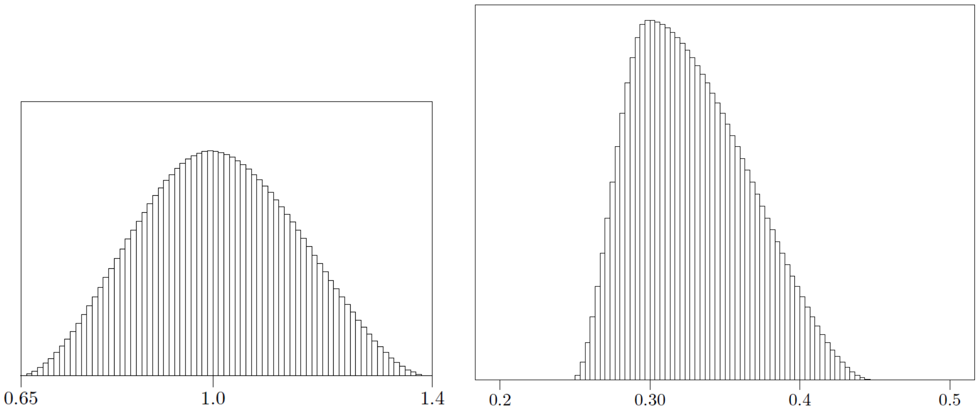


Figure 3: Histogram evaluation values a_1 and a_2

Using least squares method and assuming $z_i = i/(n+1)$, $\varphi_1 = 1$, $\varphi_2 = x^2$, get the following assessment $a_0 = 0.95$, $a_2 = 3.23$. Numerical probabilistic analysis allows to construct a probabilistic extension of $\lambda(t)$. Histogram estimation of probability density functions for a_1, a_2 shown in Figures 3.

Using probabilistic extension of λ can calculate the histogram evaluation values $P(t)$ at any time. Figure 4 shows the histogram evaluation $P(t)$ at the time $t = 1$.

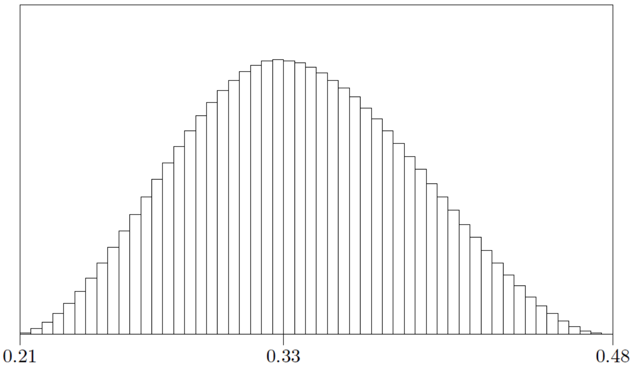


Figure 4: Histogram evaluation values $P(1)$

Conclusion

This method of estimating the failure rate of equipment critical applications, enables us to construct reliable estimates of parameters of reliability of complex technical systems in a small sample. The reliability of estimates are constructed using probabilistic extensions CDF. The proposed approach can be used to assess the various risks in complex technical systems in conditions of limited information.

References

[1] Ahlberg J.H., Nilson E.N., Walsh J.L. (1967) *The theory of splines and their applications*. Academic Press, New York.

[2] Gerasimov V.A., Dobronets B.S., and Shustrov M.Yu. (1991). Numerical operations of histogram arithmetic and their applications. *Automation and Remote Control*, Vol. **52(2)**, pp. 208–212.

[3] Dobronets B.S.(2004) *Interval Mathematics*. Krasnoyarsk: KSU. (Russian) (Russian)

[4] Dobronets B.S., Popova O.A. (2011). Numerical Operations on Random Variables and their Application, *Journal of Siberian Federal University. Mathematics & Physics* Vol. **4(2)**, pp. 229–239. (Russian)

[5] Dobronets B.S., Popova O.A. (2012). Elements of numerical probability analysis. *SibSAU Vestnik*, Vol. **42(2)**. pp. 19–23. (Russian)

[6] Dobronets B.S., Popova O.A. (2012). Numerical probabilistic analysis for the study of systems with uncertainty. *Journal of Control and Computer Science*,Vol. **21(4)** pp.39–46. (Russian)

[7] Dobronets B.S., Popova O.A. (2014). Numerical Probabilistic Analysis under Aleatory and Epistemic Uncertainty *Reliable Computing*. Vol.**19**. pp. 274–289.

[8] Dobronets B.S., Popova O.A. (2014) *The numerical probabilistic analysis of uncertainty data*. Sib. Feder. University, Krasnoyarsk. 167 pages. (Russian)

[9] Dobronets B.S., Popova O.A. (2014) Representation and processing of uncertainty based on histogram distribution functions and p-boxes. *informatization and communication* No **2**. pp. 23–26. (Russian)

[10] Ferson S., Kreinovich V., Ginzburg L., Myers D.S., and Sentz K. (2003). Constructing probability boxes and dempster–shafer structures. *Technical Report SAND2002–4015*, Sandia National Laboratoris.

[11] Moore R.E.. (1984). Risk analysis without monte carlo methods. *Freiburger Intervall- Berichte*, Vol. **84(1)**. pp. 1–48.

[12] Sobol I.M. (1994). *A Primer for the Monte Carlo method.* CRC Press, Boca Raton.

[13] Williamson R., Downs T. (1990). Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* Vol. **4**.

Piecewise Linear Processes on the Poisson Flow

Olga V. Sereseva and Vasily A. Ogorodnikov
Institute of Computational Mathematics and Mathematical Geophysics,
Novosibirsk State University, Novosibirsk, Russia
e-mail: `seresseva@mail.ru, ova@osmf.sscc.ru`

Abstract

The results of a study of two types of piecewise linear process on the Poisson flow are considered in the paper. The process of the first type is a non-stationary process, whose values in Poisson reference points are the sum of independent uniformly distributed random variables with a consistently increasing number of terms. In the second type of process values in Poisson reference points are independent random variables with the given distribution function. For the process of the first type are obtained exact expressions for the mean and variance as a function of time. For the process of the second type are shown that it is asymptotically stationary process of a one-dimensional distribution. On concrete examples numerically are shown that the second process fairly quickly reaches a steady level. The correlation structure of the processes is investigated numerically.

Keywords: piecewise linear process, Poisson flow, correlation functions, inflection point.

Introduction

In [1] was considered several approaches to modeling of ruled non-Gaussian processes with respect to the modeling of price series. In [2] was considered a ruled random process on the positive semiaxes $t > 0$ taking the following values in the interval (S_k, S_{k+1})

$$Y(t) = (Y_{k+1} - Y_k)\frac{t - S_k}{S_{k+1} - S_k} + Y_k = (Y_{k+1} - Y_k)\,Q(t) + Y_k, S_k \le t < S_{k+1},\ k = 0, 1, \ldots \tag{1}$$

Here $S_0 = 0$, $S_k = \sum\limits_{i=1}^{k} X_i$, X_i are independent positive random variables with density $f(x) = \lambda \exp(-\lambda x)$, $Y_k = \sum\limits_{i=0}^{k} \alpha_i$, α_i – are mutually independent random variables not dependent on X_i and uniformly distributed in the interval $[a\,,\,b]$, $a < b$

In [3] function (1) was considered in the form

$$Y(t) = (Y_{\nu(t)} - Y_{\nu(t)-1})\frac{t - S_{\nu(t)-1}}{S_{\nu(t)} - S_{\nu(t)-1}} + Y_{\nu(t)-1} = \alpha_{\nu(t)}\,Q_{\nu(t)-1}(t) + \sum_{i=0}^{\nu(t)-1} \alpha_i, \tag{2}$$

$$S_{\nu(t)-1} \le t < S_{\nu(t)},$$

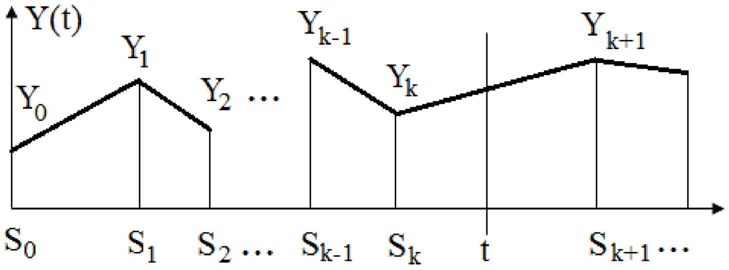


Figure 1

where integer random variable

$$\nu(t) = Min\{n \geq 1 : S_n \geq t\} \in [0, \infty), \;\; t > 0$$

is the number of a random interval covering the point t. The probability for $n = \nu(t)$ equals

$$\Pr\{\nu(t) = n\} = \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}, \;\; 0 < n < \infty.$$

First and second moments $\nu(t)$ equals

$$E[\nu(t)] = 1 + \lambda t, \;\; E[\nu^2(t)] = 1 + 3\lambda t + \lambda^2 t^2. \qquad (3)$$

Also in [3] formula for mathematical expectation $E[Y(t)]$ was obtained for the process $Y(t)$

$$E[Y(t)] = \frac{b+a}{4}\left(3 + \lambda t\left((2 + \lambda t)\Gamma(0, \lambda t) + 2\right) - (1 + \lambda t)e^{-\lambda t}\right). \qquad (4)$$

This paper is a continuation of research of non-stationary processes of the type (1) of [2], [3]. Using the representation (2) is obtained an exact expression for the variance of the process. We also consider a modification of the process (1), according to which the variables Y_k are independent random variables with the one-dimensional distribution $F(x)$. It is shown that this process is asymptotically stationary on the mean and variance. The corresponding expressions for mean and variance are obtained. Examples are given for the mean values and variances as a function of time for cases where Y_k is a uniform distribution, or exponential distribution. The correlation structure of the processes is investigated numerically.

1 Dispersion of piecewise-linear process with additive random components at the reference points

Dispersion $V[Y(t)]$ of the process $Y(t)$ is determined by the expression:

$$V[Y(t)] = E[Y^2(t)] - (E[Y(t)])^2$$

$$E[Y^2(t)] = E[\alpha^2_{\nu(t)}]\, E[\, Q^2_{\nu(t)-1}(t)] + 2E[\alpha_{\nu(t)}]\, E[\, Q_{\nu(t)-1}(t) \sum_{i=0}^{\nu(t)-1} \alpha_i] \, + E[(\sum_{i=0}^{\nu(t)-1} \alpha_i)^2]$$

for the case when Y_k is (2). Taking into account (1), the expressions for $E[\, Q_{\nu(t)-1}(t)]$ and $V[\, Q_{\nu(t)-1}(t)]$ [3]:

$$E[\, Q(t)] = \frac{1}{2}(1 + \lambda t(2 + \lambda t)\Gamma(0, \lambda t) - (1 + \lambda t)e^{-\lambda t}),$$

$$\begin{aligned} V[Q(t)] \; = \; & \tfrac{1}{12}(1 - \lambda t\,(2(6 + \lambda t(9 + 2\lambda t)) + 3\lambda t(2 + \lambda t)^2\Gamma(0, \lambda t))\Gamma(0, \lambda t) + \\ & +(2(1 + 7\lambda t + 2\lambda^2 t^2) + 6\lambda t(2 + 3\lambda t + \lambda^2 t^2)\Gamma(0, \lambda t) - 3e^{-\lambda t}(1 + \lambda t)^2)e^{-\lambda t}) \end{aligned}$$

and using the expressions

$$E[\alpha_n] = \frac{b + a}{2}, \quad E[\alpha^2_n] = \frac{a^2 + ab + b^2}{3},$$

$$E[\, Q^2_{\nu(t)-1}(t)] = \frac{1}{3}e^{-\lambda t}\,(-1 + e^{\lambda t} + 2\lambda t + \lambda^2 t^2 - \; e^{\lambda t}\lambda^2 t^2(3 + \lambda t)\Gamma(0, \lambda t)),$$

$$E[(\sum_{i=0}^{\nu(t)-1} \alpha_i)^2] \, = (1 + \lambda t)\, \frac{a^2 + ab + b^2}{3} + (2\lambda t + \lambda^2 t^2)\frac{(a + b)^2}{4},$$

$$\begin{aligned} & E[\, Q^2_{\nu(t)-1}(t)]\,[\, \textstyle\sum_{i=0}^{\nu(t)-1} \alpha_i] \; = \\ & = \tfrac{a+b}{2}\left(\lambda^2 t \int\limits_t^\infty \tfrac{e^{-\lambda y_3}}{y_3} dy_3 + \lambda^2 \int\limits_0^t \int\limits_t^\infty \tfrac{t - y_2}{y_3 - y_2} e^{-\lambda y_3} e^{\lambda y_2}(2 + \lambda y_2) dy_3 dy_2\right) = \\ & = \tfrac{a+b}{2}\lambda t\Gamma(0, \lambda t) + \tfrac{a+b}{12}e^{-\lambda t}\left(-2 - 5\lambda t - (\lambda t)^2 + e^{\lambda t}\left(2 + 3\lambda t - (\lambda t)^2(6 + \lambda t)\mathrm{Ei}[\text{-}\lambda t]\right)\right) \end{aligned}$$

After some transformations we get expression for $V[Y(t)]$

$$\begin{aligned} V[Y(t)] = & \tfrac{a^2+ab+b^2}{3}\tfrac{1}{3}e^{-\lambda t}\,(-1 + e^{\lambda t} + 2\lambda t + \lambda^2 t^2 - \; e^{\lambda t}\lambda^2 t^2(3 + \lambda t)\Gamma(0, \lambda t)) \; + \; 2\tfrac{b+a}{2}\cdot \\ & \cdot\left(\tfrac{a+b}{2}\lambda t\Gamma(0, \lambda t) + \tfrac{a+b}{12}e^{-\lambda t}\left(-2 - 5\lambda t - (\lambda t)^2 + e^{\lambda t}\left(2 + 3\lambda t - (\lambda t)^2(6 + \lambda t)\mathrm{Ei}[\text{-}\lambda t]\right)\right)\right) + \\ & +(1 + \lambda t)\,\tfrac{a^2+ab+b^2}{3} + (2\lambda t + \lambda^2 t^2)\tfrac{(a+b)^2}{4} - \\ & - \left(\tfrac{b+a}{4}\left(3 + \lambda t\,((2 + \lambda t)\Gamma(0, \lambda t) + 2) - (1 + \lambda t)e^{-\lambda t}\right)\right)^2 . \end{aligned} \tag{5}$$

If $-b = a \quad E[Y(t)] = 0$ and

$$V[Y(t)] = \tfrac{a^2+ab+b^2}{3}\left(1 + \lambda t + \tfrac{1}{3}e^{-\lambda t}\left(e^{\lambda t}\left(1 - \; e^{\lambda t}\lambda^2 t^2(3 + \lambda t)\Gamma(0, \lambda t)\right) + 2\lambda t + \lambda^2 t^2 - 1\right)\right)$$

For large $t \quad D[Y(t)] = \frac{a^2+ab+b^2}{9}(4 + 3\lambda t)$.

The graph of the function $E[Y(t)]$ is presented in Fig. 2 for the various parameter values a, b, $\lambda$and in Fig. 3, Fig. 4. graphs of the function $\sqrt{D[Y(t)]}$ at different time intervals. Fig. 4 also shows the value of $\sqrt{D[Y(t)]}$ calculated by the model samples. Choosing the parameters a, b and λ the model can be adjusted to the actual data. In this paper, for the calculation of the characteristics of the model samples number of trajectories are equal $n = 1000000$.

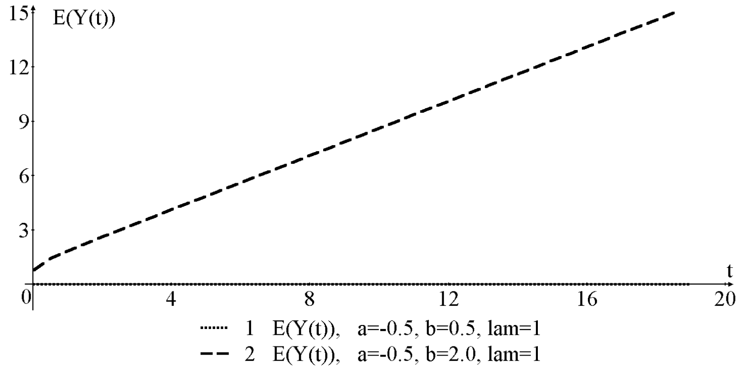


Figure 2: Graph of the function $E[Y(t)]$ calculated by the formula (4) for the cases: $-b = a = 0.5$ – 1, $a = 0.5$, $b = 2.0$ – 2 on the time interval (0,19).

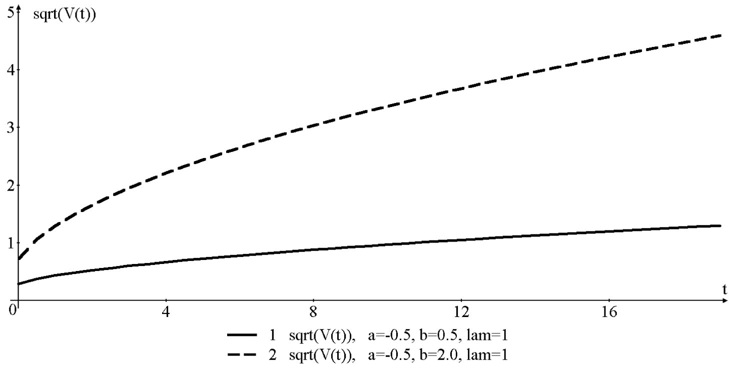


Figure 3: Graph of the function $\sqrt{D[Y(t)]}$ calculated by the formula (5) for the cases: $-b = a = 0.5$ – 1, $a = 0.5$, $b = 2.0$ – 2 on the time interval (0,19).

2 The piecewise-linear process with independent and identically distributed Poisson variables in reference points

Consider a ruled random process $Y(t)$ form (1) (or (2)), which Y_k are independent random variables with an arbitrary one-dimensional distribution $F(x)$. Let us obtain an expression for the mean value of the process $Y(t)$:

$$E[Y(t)] = E[\,(X_{\nu(t)} - X_{\nu(t)-1})\,Q_{\nu(t)-1}(t) \;+ X_{\nu(t)-1}] =$$
$$= E[\,X_{\nu(t)}\,]E[Q(t)] \;- E[\,X_{\nu(t)-1}\,]E[Q(t)] \;+ E[\,X_{\nu(t)-1}\,] = E[\,X_{\nu(t)-1}\,].$$

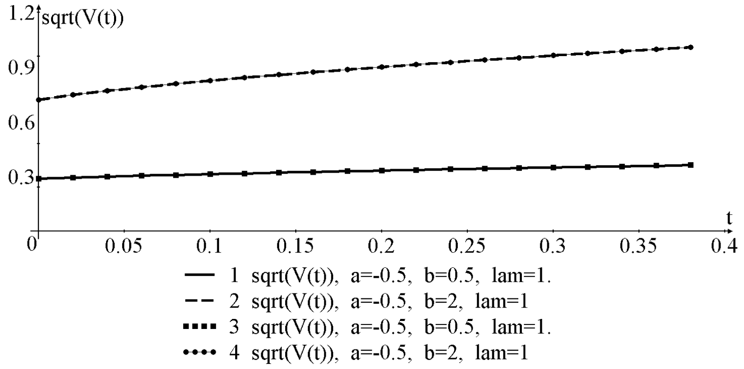


Figure 4: Graph of the function $\sqrt{D[Y(t)]}$ calculated by the formula (5) and by the model samples for the cases: $-b = a = 0.5$ – 1,3, $a = 0.5$, $b = 2.0$ – 2,4 on the time interval (0,0.4).

The expression for the dispersion $D[Y(t)]$ of the process $Y(t)$ has the form:

$$\begin{gathered}
D[Y(t)] = E[Y^2(t)] - (E[Y(t)])^2 = \\
= E[(X^2_{\nu(t)} - 2\,X_{\nu(t)}X_{\nu(t)-1} + X^2_{\nu(t)-1})Q^2_{\nu(t)-1}(t) + \\
+2(X_{\nu(t)} - X_{\nu(t)-1})Q_{\nu(t)-1}(t)X_{\nu(t)-1} + X^2_{\nu(t)-1}] = \\
= \left(E[X^2_{\nu(t)}] \,- 2E[X_{\nu(t)}]\,E[X_{\nu(t)-1}] \,+ E[X^2_{\nu(t)-1}]\right) E[Q^2_{\nu(t)-1}(t)\,] + \\
+2\left(E[X_{\nu(t)}]E[X_{\nu(t)-1}] - E[X^2_{\nu(t)-1}]\right)\, E[Q_{\nu(t)-1}(t)] \,+ E[X^2_{\nu(t)-1}] - (E[X_{\nu(t)-1}(t)])^2 = \\
= D[X_{k+1}]\left(2(E[Q^2_k(t)\,] - E[Q_k(t)]) + 1\right).
\end{gathered}$$

Consider two examples.

1. If $Y_0 = \alpha_0$, $Y_n = \alpha_n\,, n \geq 1$, α_n - mutually independent random variables not dependent on X_i, and uniformly distributed in the interval $[a\,, b]$ with a density distribution

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b], \\ 0, & x \notin [a,b], \end{cases}$$

Then in this case the expression for the mean value of the process $Y(t)$ takes the form:

$$E[Y(t)] \,= E[\,\alpha_{\nu(t)-1}\,] = (b+a)/2,$$

And the expression for the dispersion $D[Y(t)]$ of the process $Y(t)$ has the form:

$$\begin{gathered}
D[Y(t)] = \tfrac{1}{36}\left((11a^2 + 14ab + 11b^2) + (a-b)^2(1 + \lambda t(7 + 2\lambda t))e^{-\lambda t} - \right. \\
\left. -(a-b)^2\lambda t(6 + \lambda t(9 + 2\lambda t))\Gamma(0,\lambda t)\right) - \tfrac{(a+b)^2}{4}.
\end{gathered}$$

Note that:
$\lim\limits_{t\to 0} D[Y(t)] = \frac{(a-b)^2}{12}$ and $\lim\limits_{t\to\infty} D[Y(t)] = \frac{(a-b)^2}{18}$.

2. If $Y_0 = \alpha_0$, $Y_n = \alpha_n\,, n \geq 1$, α_n - mutually independent random variables

not dependent on X_i, and exponential distributed with parameter λ_1 then expression for the mean value of the process $Y(t)$ takes the form:

$$E[Y(t)] \ = E[\,\alpha_{\nu(t)-1}\,] = 1/\lambda$$

And the expression for the dispersion $D[Y(t)]$ of the process $Y(t)$ has the form:

$$D[Y(t)] = \tfrac{1}{3\lambda^2}e^{-\lambda t}\left(1 + 5e^{\lambda t} + 7\lambda t + 2\lambda^2 t^2 - \ e^{\lambda t}\lambda t(6 + 9\lambda t + 2\lambda^2 t^2)\Gamma(0, \lambda t)\right) - \tfrac{1}{\lambda^2} =$$
$$= \tfrac{1}{3\lambda^2}e^{-\lambda t}\left(1 + 2e^{\lambda t} + 7\lambda t + 2\lambda^2 t^2 - \ e^{\lambda t}\lambda t(6 + 9\lambda t + 2\lambda^2 t^2)\Gamma(0, \lambda t)\right).$$

At that $\lim\limits_{t\to 0} D[Y(t)] = \ \frac{1}{\lambda^2}$ and $\lim\limits_{t\to\infty} D[Y(t)] = \frac{2}{3\lambda^2}$.

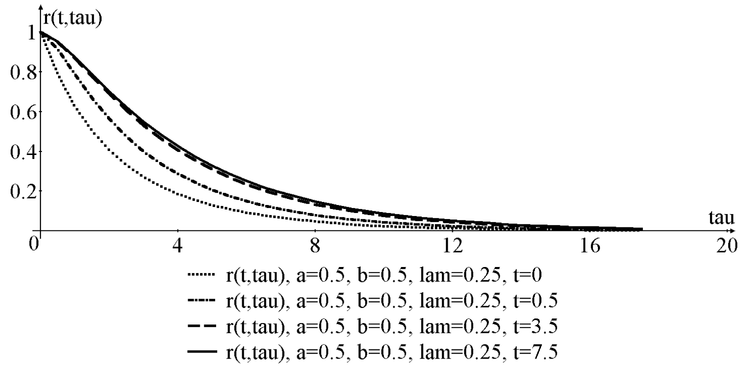


Figure 5: The dependence of the correlation function $r(t, \tau)$ of the process $Y(t)$ from time t, $f(x) \ = \lambda_1 \exp(-\lambda_1 x)$, $\quad \lambda_1 = 0.25$.

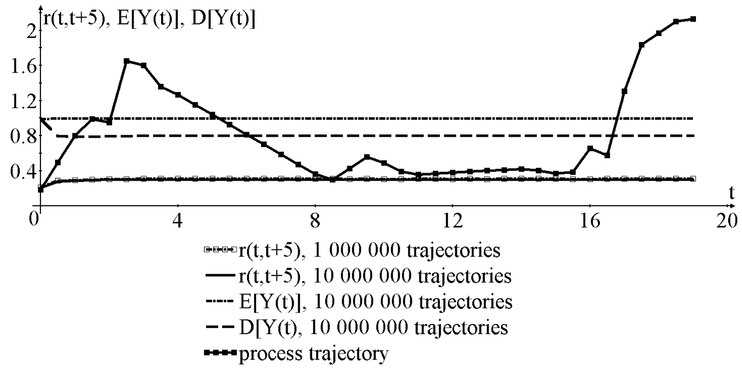


Figure 6: Correlation function of the process $Y(t)$ here $f(x) \ = \lambda \exp(-\lambda x)$ and $\lambda_1 = 0.25, \quad 0.125, \quad 0.0625.$

These examples show that the process of (1) to consider the distributions of Y_n is asymptotically stationary on the mean and variance. The correlation structure of the processes of the form (1) in this paper was investigated numerically on the basis of model samples. The calculation results for the case $F(x) \ = 1 - \exp(-\lambda_1 x), \quad \lambda_1 = 1$ are shown in Fig. 5.6. Fig. 5 shows the dependence of the correlation function $r(t, \tau)$

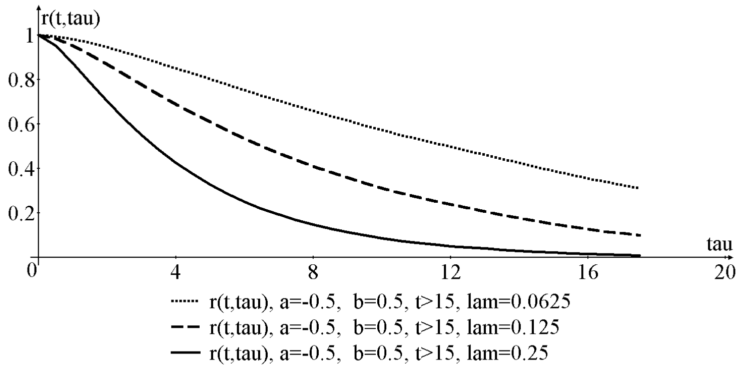


Figure 7: Correlation function of the process $Y(t)$ here $f(x) = \lambda \exp(-\lambda x)$ and $\lambda_1 = 0.25, \quad 0.125, \quad 0.0625$.

of the process $Y(t)$ from time. For a given process parameters and for $t > 15$ the process becomes close to stationary on correlations. The process behaves similarly for the mean and variance (Fig. 7). The dependence of the correlation coefficient $r(t, t + \tau_1)$, $\tau = 10$ from t is presented in Fig. 7

The examples of correlation function $r(t, \tau)$ of process $Y(t)$ for values $t = 20$ and $\lambda_1 = 0.25, \quad 0.125, \quad 0.0625$ is presented in Fig. 7. The inflection point for each of the functions which are shown in the fig. 7 is a characteristic feature of this process. This distinguishes them from the correlation functions of piecewise constant processes on Poisson point flows and flows of Palma [4], which is a characteristic feature is the bulge down.

Conclusions

In conclusion, it should be noted that the algorithms are modifications of algorithms for simulation of piecewise constant processes on point flows, in particular, considered in [4]. Specificity of asymptotically stationary processes discussed in this paper is that their correlation functions have a point of inflection and dimensional distribution of the process does not coincide with the distribution of the random variables in Poisson reference points. If you use the Poisson flow as a periodic function of the time the process becomes asymptotically periodically correlated.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grants No. 15-01-01458-a, 15-01-08988-a) and Program “Leading Scientific Schools of Russian Federation” (NSh-5111.2014.1).

References

[1] V.A.Ogorodnikov, A.V. Novikov (2002) Stochastic model of price series. Proceedings of the International Conference on Computational Mathematics. Novosibirsk, P. 243-248.

[2] V.A.Ogorodnikov, L.Ya.Savel'ev, O.V.Sereseva (2007) Numerical stochastic models of piecewise-linear random processes // Rus. J. Numer. Analys. Math. Modeling. V. 22. No 5. P. 505-514.

[3] L.Ya.Savel'ev, V.A.Ogorodnikov, O.V.Sereseva (2007) Stochastic model of piecewise-linear random process. Vestnik Syktyvkar university. V.1. No 7. P 67-76. (in Russian)

[4] G.A. Mikhailov. Optimisation of weight Monte-Carlo Methods. springer,New York (1992)

An Algorithm for Numerical Simulation of Isotropic Random Fields and its Meteorological Application

Nina A. Kargapolova and Vasily A. Ogorodnikov
Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk State University, Novosibirsk, Russia
e-mail: nkargapolova@gmail.com, ova@osmf.sscc.ru

Abstract

In this paper several properties of special type of correlation functions are considered. This paper presents also a model of monthly mean temperature fields in the Lake Baikal area. Suggested model is based on real data and on studied in the paper simulation algorithm for isotropic random fiels.

Keywords: isotropic field, stochastic simulation, monthly mean temperature, Lake Baikal.

Introduction

Solution of various applied problems related to the study of actual time series (for example, meteorological, oceanologic or hydrological) rather often requires numerical simulation of isotropic or homogeneous random fields. Existing general algorithms are computer memory- and time-consuming. That is why special algorithms for simulation of random fields with different types of correlation dependence are developed [1, 2].

To simulate isotropic discrete Gaussian scalar $m-$dimensional random fields with correlation function

$$R(r) = \int_a^b \exp\left(-xr^2\right) f(x)\, dx,$$

where $r-$Cartesian distance between 2 points in R^m, $f(x)-$ distribution density on $[a;b]$, $a>0$, a method, suggested in [5], may be used. Formally, this method is not "precise" one, in the sense of distribution – simulated field does not have Gaussian distribution, but distribution of the field is rather close to Gaussian. Suggested method is low cost in sense of memory- and time-consumption because it is based on modification of "on rows and columns" algorithm, which was developed for simulation of fields with correlation function $\exp\left(-xr^2\right)$ [2].

In this paper properties of correlation function $R(r)$ are considered. This paper presents also a model of monthly mean temperature fields in the Lake Baikal area. Suggested model rely heavily on real data and special case of $R(r)$.

1 Flex points of correlation function

Here are several examples of function $R(r)$.

Example 1. Let $f(x)$ be density of a right truncated exponential distribution with parameter $\lambda > 0$, i.e.

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}}{1-e^{-\lambda b}}, & x \in (0,b), \\ 0, & x \notin (0,b). \end{cases}$$

In that case

$$R_1(r) = \int_0^b e^{-xr^2} \frac{\lambda e^{-\lambda x}}{1-e^{-\lambda b}} dx = \frac{\lambda}{1-e^{-\lambda b}} \frac{1}{r^2+\lambda} \left(1 - e^{-\left(r^2+\lambda\right)b}\right),$$
$$R_1(0) = 1.$$

Example 2. If $f(x)$ is parabolic distribution density on the interval $[a,b]$, $a > 0$, i.e.

$$f(x) = \begin{cases} \frac{6(x-a)(b-x)}{(b-a)^3}, & x \in [a,b], \\ 0, & x \notin [a,b], \end{cases}$$

then

$$R_2(r) = \frac{6}{(b-a)^3} \left(e^{-br^2} \left[\frac{2}{r^6} - \frac{a}{r^4} + \frac{b}{r^4}\right] - e^{-ar^2} \left[\frac{2}{r^6} + \frac{a}{r^4} - \frac{b}{r^4}\right]\right).$$

Example 3. Let $f(x)$ be Simpson distribution density on the interval $[a,b]$, $a > 0$, i.e.

$$f(x) = \begin{cases} \frac{4(x-a)}{(b-a)^2}, & x \in \left[a, \frac{a+b}{2}\right], \\ \frac{4(b-x)}{(b-a)^2}, & x \in \left[\frac{a+b}{2}, b\right]. \end{cases}$$

In that case function $R(r)$ is given by

$$R_3(r) = \frac{4}{(b-a)^2} \frac{1}{r^4} \left(e^{-ar^2} - 2e^{-\frac{a+b}{2}r^2} + e^{-br^2}\right).$$

Correlation functions $R_i(r)$, $i = \overline{1,3}$ are shown in Figure 1.

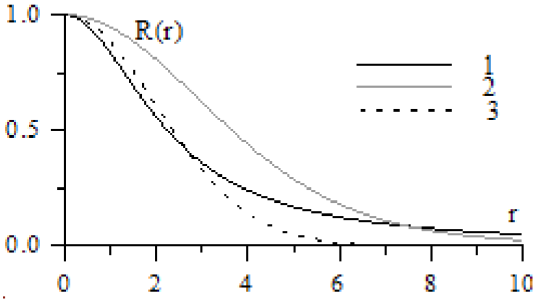


Figure 1: Functions $R_i(r)$, $i = \overline{1,3}$. Curve 1 – function $R_1(r)$ with $b = 10$, $\lambda = 5$; curve 2 – function $R_2(r)$ with $a = 0.01$, $b = 0.1$; curve 3 – function $R_3(r)$ with $a = 0.1$, $b = 0.15$.

It is well seen from Figure 1 that considered correlation functions have common characteristic: each of them has a flex point. It turns out that by the following Proposition every correlation function of considered type has a flex point.

Proposition. If $a > 0,\ \ b < +\infty$, distribution density $f(x)$ is continuous function on $[a,b]$ and $f(x) > 0,\ \ x \in [a;b]$, then function

$$F(y) = \int\limits_a^b e^{-xy^2} f(x)\,dx,\ \ y \in \mathrm{R}$$

has flex points.

To prove this proposition the following lemma is required

Lemma. Under the conditions of Proposition function $(2x^2y^2 - x)\,f(x)\exp(-xy^2)$ tends uniformly on $[a,b]$ to $(2x^2M^2 - x)\,f(x)\exp(-xM^2)$ if $y \to M$.

Proof of the Lemma. Let us denote $G(x,y) = (2x^2y^2 - x)\,e^{-xy^2} f(x)$. According to [3] it is necessary to prove that

$$\forall \varepsilon_1 > 0\ \ \exists \delta(\varepsilon_1):\ \ |G(x,y_1) - G(x,y_2)| < \varepsilon_1$$

for all $x \in [a,b]$ as soon as $|y_1 - M| < \delta,\ \ |y_2 - M| < \delta$. By Cantor Theorem $G(x,y)$ is uniformly continuous function on $[a;b] \times [-M;M]$, i.e.

$$\forall \varepsilon > 0\ \ \exists \delta:\quad |G(x,y) - G(x_0,y_0)| < \varepsilon$$

for all $(x,y),\ \ (x_0,y_0)$ simultaneously, if $|x - x_0| < \delta,\ \ |y - y_0| < \delta$. If we consider such points $(x,y),\ \ (x_0,y_0)$, that $x = x_0,\ \ y_0 = M,\ \ y = y_i,\ \ i = 1,2$, then $\forall \delta > 0\ \ |x - x_0| = 0$. From definition of uniform continuity it follows that

$$\forall \frac{\varepsilon_1}{2} > 0\ \ \exists \delta_i(\varepsilon_1):\ \ |G(x,y_i) - G(x,M)| < \frac{\varepsilon_1}{2}$$

for all points $(x,y_i),\ \ (x,M)$ simultaneously, if $|y_i - M| < \delta_i$. Thus,

$$\begin{gathered}\forall \varepsilon_1 > 0\ \ \exists \delta_1,\delta_2:\ \ |G(x,y_1) - G(x,y_2)| \le \\ \le |G(x,y_1) - G(x,M)| + |G(x,y_2) - G(x,M)| \le \tfrac{\varepsilon_1}{2} + \tfrac{\varepsilon_1}{2} = \varepsilon_1\end{gathered}$$

for all x simultaneously, if $|y_1 - M| < \delta_1,\ \ |y_2 - M| < \delta_2$. This implies that, if with given $\varepsilon_1 > 0$ we chose $\delta = \min\{\delta_1,\delta_2\}$, than inequality

$$|G(x,y_1) - G(x,y_2)| < \varepsilon_1$$

holds for all x simultaneously, if $|y_1 - M| < \delta,\ \ |y_2 - M| < \delta$.

Proof of the Proposition. It is necessary to prove that point, where $\partial^2 F(y)/\partial y^2$ is equal to 0 and changes its sign, exists. For convenience, we'll study function

$$I(y) = \frac{1}{2}\frac{\partial^2 F(y)}{\partial y^2} = \int\limits_a^b \left(2x^2y^2 - x\right)\exp\left(-xy^2\right) f(x)\,dx,\ \ y \in \mathrm{R}.$$

We show that for all $a,b:\ \ 0 < a < b < +\infty$ such $M_1 > 0, M_2 > 0$ exist, that $I(M_1) > 0, I(M_2) < 0$. By Lemma and by theorem about passage to the limit under the integral depending on a parameter:

$$\lim_{y \to M} I(y) = \int\limits_a^b \left(2x^2M^2 - x\right)\exp\left(-xM^2\right) f(x)\,dx.$$

At the same time, equality

$$I(M) = \lim_{y \to M} I(y)$$

follows from continuity theorem for parameter-dependent proper integral (integrand is a continuous function).

Hence,

$$I(M) = \int\limits_a^b \left(2x^2M^2 - x\right) \exp\left(-xM^2\right) f(x)\,dx.$$

If $M^2 > \frac{1}{2a}$, then $2x^2M^2 - x > 0$ for all $x \in [a;b]$. This is almost obvious. On the one hand, $x > 0$, and so $2x^2M^2 - x > 0 \;\Leftrightarrow\; M^2 > \frac{1}{2x}$. On the other hand, just as $0 < a < b < +\infty$, so $\frac{1}{2b} \leq \frac{1}{2x} \leq \frac{1}{2a}$. If $M^2 > \frac{1}{2a}$, then $M^2 > \frac{1}{2x}$ and thus $2x^2M^2 - x > 0$ for all $x \in [a;b]$. Integrand is a positive function, limits of integration are greater than 0, so every real number greater than $1\big/\sqrt{2a}$ may be used as M_1. Similarly, inequality $2x^2M^2 - x < 0$ holds for all $x \in [a;b]$, if $M^2 < \frac{1}{2b}$. Hence, every real number less than $1\Big/\sqrt{2b}$ may be used as M_2. Thus, $I(y)$ is continuous function and $\exists\ M_1, M_2 :\ \ I(M_1) > 0,\ \ I(M_2) < 0$. These facts implies that $\exists\ M_0 > 0 :\ \ I(M_0) = 0$. So, $I(y) = \frac{1}{2}\frac{\partial^2 F(y)}{\partial y^2}$ is equal to 0 if $y = M_0$ and changes sign. It means that $M_0 > 0$ is a flex point of function $F(y)$.

Remark 1. Proposition holds if $f(x) \geq 0,\ \ x \in [a;b]$. In this case it is necessary to chose such points M_1, M_2 that $f(M_1) \neq 0,\ \ f(M_2) \neq 0$.

Remark 2. Proposition holds even if $f(x)$ is a distribution density on $[a; +\infty)\,,\ \ a > 0$. In this case proof is a little bit trickier.

It was shown in this section, that all correlation functions of isotropic random fields, that admit considered representation, have flex points.

2 Stochastic model of monthly mean temperature spatio field

Statistical structure description of monthly mean temperature fields (MMTF) in the Lake Baikal area is given in [4]. Study was based on long-term real data from 33 weather stations. These stations are situated in $14 \cdot 10^5\ km^2$ area. Map of considered area is given in Figure 2.

It was shown that MMTF is isotropic, but parameters of one-dimensional distribution are different at various locations. Because of small sample size it is practically impossible to estimate probabilistic characteristic of rare events on basis of real data. For estimation of rare events probabilistic characteristic it is necessary to develop and use a simulation model of MMTF. Such model was developed under the following conditions:

1. Random field was simulated in nodes of a regular grid with rectangle cells. Size of cells was $35\,km \times 25\,km$;

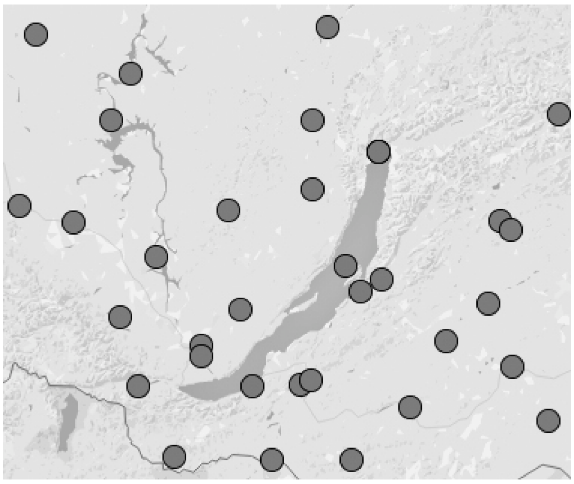


Figure 2: Map of considered area. Dots show locations of weather stations.

2. In every node one-dimensional distribution of the field was assumed to be a Gaussian one. Parameters of Gaussian distribution $N\left(\mu_i, \sigma_i^2\right)$ were taken according to following formulas:

$$\begin{gathered}\mu_i = p_{i1} M_{i1} + p_{i2} M_{i2}, \\ \sigma_i^2 = p_{i1} \Sigma_{i1}^2 + p_{i2} \Sigma_{i2}^2,\end{gathered}$$

   where M_{i1}, M_{i2} – sample means on nearest and second to nearest to node No. "i" weather stations, Σ_{i1}^2, Σ_{i2}^2 – sample variance on these stations, p_{i1}, p_{i2} – weights that are inversely proportional to distances between node No. "i" and stations;

3. Random field was assumed to be isotropic.

Figure 3 shows values μ_i for all nodes, when September mean temperature was considered. Adequacy of such choice of parameters was checked using data from 2 weather stations, which are situated exactly in grid nods and were not used during parameters estimation.

For the construction of the model we need an approximation of the empirical correlation function by a certain special function describing the field structure at arbitrary points of the considered domain. Set of experiments with different types of correlation functions was carried to define best approximating function. As a criteria for goodness of approximation minimality condition for the mean square difference between actual and approximating functions was used. Close approximation for all months was obtained with function

$$corr\left(r\right) = \lambda \big/ \left(r^2 + \lambda\right),$$

where r is Cartesian distance between points. This function is a special case of function $R\left(r\right) = \int\limits_a^b \exp\left(-xr^2\right) f\left(x\right) dx$, considered in previous section, when $f\left(x\right) =$

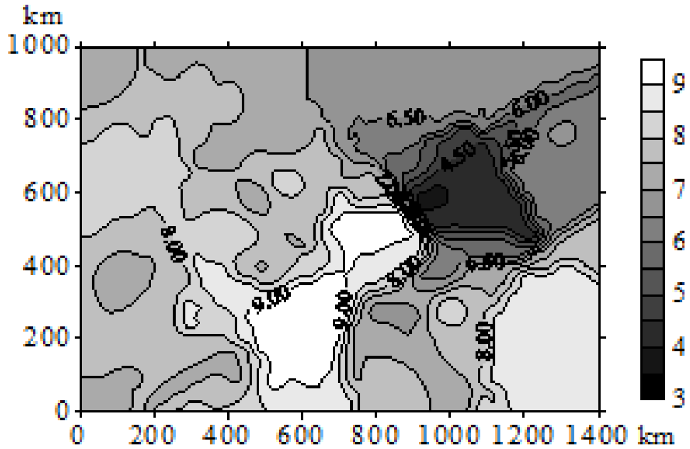


Figure 3: First parameter of Gaussian distribution $N(\mu_i, \sigma_i^2)$. September.

$\lambda \exp(-\lambda x)$, $x \geq 0$, $\lambda > 0$. To illustrate correlation structure of real and simulated fields Figure 4 is given. It shows both correlation coefficients between weather station, situated in village Chervjanka, and other stations and approximating function.

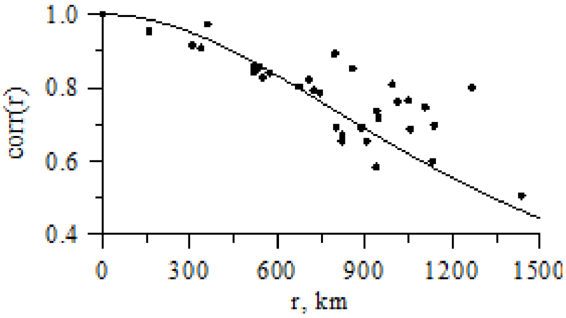


Figure 4: Correlation coefficients estimated on real data (black dots) and approximating function (black curve). August.

Depending on month parameter $\lambda > 0$ changes radically. Some values of this parameter are presented in Table 1.

Table 1: Parameters of the approximating correlation function.

Month	λ
February	2339810
April	4836330
July	1144200

Simulation of random field with correlation function $corr(r) = \lambda/(r^2 + \lambda)$ was done using algorithm suggested in [5].

Using the numerical model of monthly mean temperature we examine certain characteristics of the MMMF. Figure 5 shows dependence between share of territory, where monthly mean temperature is below given level, and this level. 10000 samples of random field were used

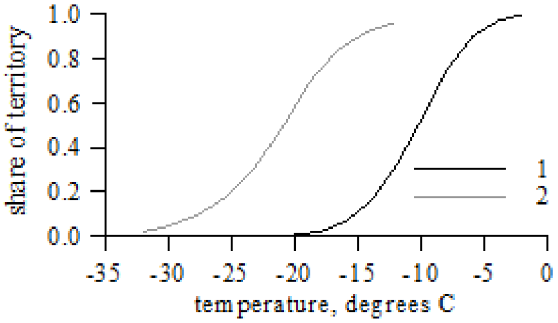


Figure 5: The level-dependence of share of territory, where monthly mean temperature is below level. Curve 1- March, curve 2 – December.

Using the model probability of very cold/warm month on given territory was estimated. Recall that month is considered as very cold/warm if its monthly mean temperature is 4 degrees Celsius less/greater than so calld "normal mean temperature". These probabilities, estimated for 875 km^2-area around Irkutsk, are given in Table 2.

Table 2: Probabilities of very cold/warm month.

Month	Pr. of very cold month	Pr. of very warm month
January	0.005	0.122
April	0.013	0.016
December	0.020	0.049

Many other meteorological characteristics may be estimated on basis of considered model. Obtained results may also be used as a basis of both spatio-temporal model of monthly mean temperature with due regard for annual cycle of real processes and joint model "monthly mean temperature and monthly precipitation sums".

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grants No. 15-01-01458-a, 15-01-08988-a) and Program "Leading Scientific Schools of Russian Federation" (NSh-5111.2014.1).

References

[1] Ambos A.Ju., Mikhailov G.A. (2011) Statistical Modeling of the exponentially correlated multivariate random field // Rus. J. Numer. Analys. Math. Modeling. V. 26. No 3. P. 213–232.

[2] Ermakov S.M., Mikhailov G.A. (1982) Statistical modeling. – Moscow: Nauka. (in Russian)

[3] Fikhtenholts G.M. (1970) A course of differential and integral calculus. V.1,2. – Moscow: Nauka .(in Russian)

[4] Kargapolova N.A. (2015) Analysis of statistical structure of spatial-time fields of average monthly temperature and monthly precipitation sums in the lake Baikal region // Int. scientific congr."Interekspo GEO-Siberia"-2015, Int. scientific conference "Remote sensing methods and photogrammetry, environmental monitoring, geo-ecology": sourcebook. V. 1. P. 191-195. (in Russian)

[5] Ogorodnikov V.A. (1988) Simulation of one type of isotropic Gaussian fields // Collected papers "Theory and applications of Statistical Modeling". – Novosibirsk.

The Accuracy of Spectral Models for the Sea Surface Simulation

Kristina V. Litvenko[1], Sergei M. Prigarin[1,2] and
Evgeniya R. Sagoyakova[2]
[1] *Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: litchristina@gmail.com

Abstract

In this paper, numerical errors for models of the sea surface undulation based on spectral decomposition of the stochastic field of the water level are studied. Such errors depend on the number of random harmonics in a spectral model and on the size of the domain, for which the spectral model is constructed. Numerical errors are studied for temporal and spatial spectral models.

Keywords: simulation of random fields, spectral models, sea surface undulation, numerical error.

Introduction

Spectral numerical models of random processes and fields are constructed using the spectral expansions theory. Spectral models are used to study various stochastic objects and phenomena such as turbulence, atmospheric cloudiness, sea surface, etc. An important research was carry out by G. A. Mikhailov, see [4, 5], namely, the method of spectrum partitioning and randomization was proposed. This initiated a series of further studies of theoretical and applied character.

The sea surface roughness can be quite adequately described by a Gaussian homogeneous random field whose statistical properties are estimated from observations, see [1, 3]). Spectral models of the sea surface were used to solve series of applied problems by Monte Carlo method [2, 9, 11]. However, the problem of evaluating the spectral models remains insufficiently studied. In this paper, we use the approach proposed in [7] to study the errors of spectral models of the sea surface. This approach is based on the calculation of the error of correlation functions reproduction for non-randomized models and on the estimation of the mean deviation for correlation functions in randomized models.

Let us assume that a real Gaussian homogeneous random field $w(x)$, $x \in \mathbb{R}^k$, has zero mean, unit variance, the correlation function $R(x) = \mathbf{M}w(x+y)w(y)$ and the spectral density $f(\lambda)$. In this case, the spectral model of the random field and its correlation function can be written down in the form

$$w(x) = \int\limits_{\mathbb{P}} \cos\langle x, \lambda\rangle \xi(d\lambda) + \int\limits_{\mathbb{P}} \sin\langle x, \lambda\rangle \eta(d\lambda), \tag{1}$$

$$R(x) = \int\limits_{\mathbb{P}} \cos\langle x, \lambda\rangle f(\lambda) d\lambda, \tag{2}$$

where $\mathbb{P}$ is a measurable set such that $\mathbb{P}\cap(-\mathbb{P})=\{0\}$, $\mathbb{P}\cup(-\mathbb{P}) = \mathbb{R}^k$, $\xi(d\lambda)$, $\eta(d\lambda)$ are real orthogonal Gaussian stochastic measures on the half-space $\mathbb{P}$, $f(\lambda)$ is the spectral density of the random field $w(x)$, and $\langle ., .\rangle$ denotes the scalar product in $\mathbb{R}^k$. See [8] for a detailed information about the above discussed spectral models.

A simple non-randomized spectral model is an approximation of stochastic integral (1) by a finite sum of harmonics:

$$w_n(x) = \sum_{j=1}^{n} a_j^{1/2} \big[\xi_j \cos\langle \lambda_j, x\rangle + \eta_j \sin\langle \lambda_j, x\rangle\big], \quad a_j^2 = \int\limits_{Q_j} f(\lambda)(d\lambda) \tag{3}$$

where ξ_j, η_j are independent Gaussian variables, $\mathbf{M}\xi_j = \mathbf{M}\eta_j = \mathbf{M}\xi_j\eta_k = 0$, $\mathbf{M}\xi_j^2 = \mathbf{M}\eta_j^2 = 1$, the vectors $\lambda_j \in \mathbb{P}$ belong to the corresponding sets Q_j:

$$\mathbb{P} = \sum_{j=1}^{n} Q_j, \quad Q_j \cap Q_i = \emptyset \quad i \neq j.$$

The random field $w_n(x)$ in (3) is homogeneous Gaussian with the correlation function

$$R_n(x) = \sum_{j=1}^{n} a_j \cos\langle \lambda_j, x\rangle.$$

The following value can be naturally considered as the error of the spectral model:

$$\Delta(w_n, w) = \|q(x)[R(x) - R_n(x)]\|_F\,, \tag{4}$$

where $q(x)$ is a certain non-negative weight function and $\|.\|_F$ is the norm in the functional space F.

Let us consider the randomized model of a homogeneous Gaussian random field $w(x)$ with zero mean, unit variance, and the spectral density $f(\lambda)$. In this case, λ_j are random vectors independent of $(\xi_j,\, \eta_j)_{j=1\ldots n}$ and distributed in Q_j according to the probability densities $f(\lambda)/a_j^2$, $\lambda \in Q_j$.

Realizations of randomized spectral models are homogeneous random fields with a spectrum concentrated at the points λ_j taken randomly. The following values are considered as errors of randomized spectral model (3):

$$\mathbf{M}\Delta(w_n, w), \tag{5}$$

where $\Delta(w_n, w)=\big\|q(x)[R(x)-R_n(x)]\big\|_F$ is a random variable, $R_n(x)=\sum a_j^2 \cos\langle \lambda_j, x\rangle$ and the mathematical expectation is taken in (5) accoding to the joint distribution of the vectors λ_j.

1 Spectral models of the sea surface roughness

Spatial spectral models are of strong interest because stationary processes can be simulated using conventional autoregression schemes and moving average which present a series of advantages in comparison with spectral models of random processes. We consider randomized and non-randomized spatial spectral models representable as the sum of $2n^2$ harmonics:

$$\begin{aligned}&w_n(x_1,x_2)=\sum_{i=1}^{n}\sum_{j=1}^{n}a(i,j)\times\\&\Big(\xi(i,j)\cos[\lambda_1(i,j)x_1+\lambda_2(i,j)x_2]+\eta(i,j)\sin[\lambda_1(i,j)x_1+\lambda_2(i,j)x_2]+\\&\xi(i,-j)\cos[\lambda_1(i,-j)x_1+\lambda_2(i,-j)x_2]+\eta(i,-j)\sin[\lambda_1(i,-j)x_1+\lambda_2(i,-j)x_2]\Big).\end{aligned} \tag{6}$$

Here $\xi(.,.)$ and $\eta(.,.)$ are independent standard normal random variables.We can take some real value $A > 0$ and consider the following sets in $\mathbb{R}^2$:

$$\Lambda_{ij}=((i-1)A/n,iA/n)\times((j-1)A/n,jA/n),\quad i,j=1\cdots n.$$

We consider (6) as non-randomized *model N*:

$$a(i,j)^2=\iint\limits_{\Lambda_{ij}}f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2,\quad a(i,-j)=a(i,j),\quad i,j=1\cdots n. \tag{7}$$

$$\begin{aligned}&\lambda_1(i,j)=(i-0.5)A/n,\quad \lambda_2(i,j)=(j-0.5)A/n,\\&\lambda_1(i,-j)=\lambda_1(i,j),\quad \lambda_2(i,-j)=-\lambda_2(i,j),\quad i,j=1\cdots n.\end{aligned} \tag{8}$$

In this case, the correlation function of random field (6) has the form

$$R_n(x_1,x_2)=\sum_{i=1}^{n}\sum_{j=1}^{n}[a(i,j)]^2\Bigg(\cos\Big(\lambda_1(i,j)x_1+\lambda_2(i,j)x_2\Big)+\cos\Big(\lambda_1(i,j)x_1-\lambda_2(i,j)x_2\Big)\Bigg). \tag{9}$$

We consider (6), (7) as randomized *model R1* with such a distinction from the non-randomized model N, that the vectors $(\lambda_1(i,j),\lambda_2(i,j))$ are independently simulated in the corresponding sets Λ_{ij} from the distributions induced by the spectral density f. In addition, consider randomized *model R2* without spectrum partitioning similar to that used in [10]:

$$a(i,j)^2=\frac{1}{2n^2},\quad i=1,\cdots,n,\quad j=\pm1,\cdots,\pm n, \tag{10}$$

and the vectors $(\lambda_1(i,j),\lambda_2(i,j))$ for $i=1,\cdots,n$, $j=\pm1,\cdots,\pm n$ are distributed in the domain $(0,A)\times(-A,A)$ according to the density proportional to f and mutually independent.

2 The study of errors of spectral models

All the three spectral models N, R1, R2 are determined by the two parameters A and n. First, we consider a non-randomized model N of the sea roughness with the spectrum that takes into account three intervals of gravitational field ([1], pp. 19-20). For this spectrum we used the following parameters: the wind speed at 10 m above the sea level $v = 5$ and the frequency of the spectral peak $\mu_{max} = 0.4$. Figure 1 presents the correlation functions for the spectral model N with the parameters $A = 0.1$, $n = 25$.

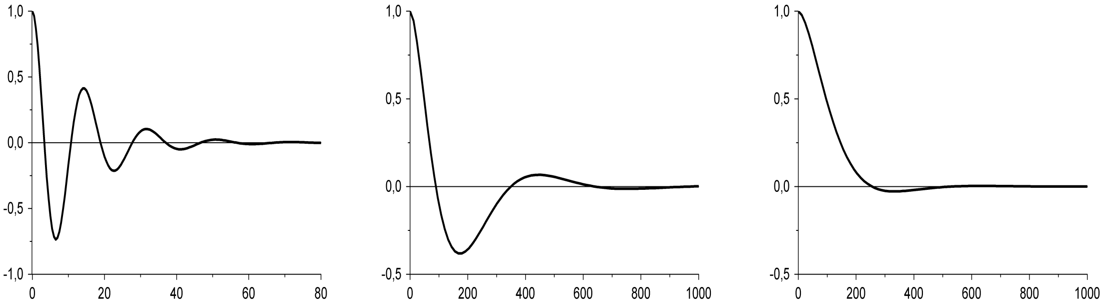


Figure 1: Normalized correlations functions R^t (a), R^x (b), R^y (c) of the random sea roughness field with the spectrum and the parameters $v = 5$ m/sec, $\mu_{max} = 0.4$. The values on the horizontal axes are given in seconds for R^t and in meters for R^x and R^y.

The correlation functions of the spectral model well approximate the corresponding correlation functions of the sea roughness on the initial sections. However, at large distances, where the values of correlation functions of the simulated random field are close to zero the values of the correlation functions of spectral models essentially deviate from zero values. For the correlation functions R^x, R^y such deviations are close to one in absolute value. This can be explained by the fact that the correlation functions R^x, R^y are represented by sums of harmonics (9). These sums are almost periodic functions, which is clear from the periodic character of realizations of spatial spectral models of the sea surface, see Figure 2. In order to increase the section where the periodicity of implementations is seen, we have to increase the number of harmonics.

Table 1 presents the distances such that the correlation of spectral models adequately represents the correlations of a random field depending on the number of harmonics. The Table also presents the values of the half-periods P^x, P^y of the spatial spectral models, i.e., the distances to the closest negative peaks of the correlation functions $R^x_n = R_n(x, 0)$, $R^y_n = R_n(0, y)$, approximating the value -1 (see (9) and Figure 3).

Concerning the error of the temporal correlation function, an increase in the number of harmonics also increases the distance to the region with essential deviations

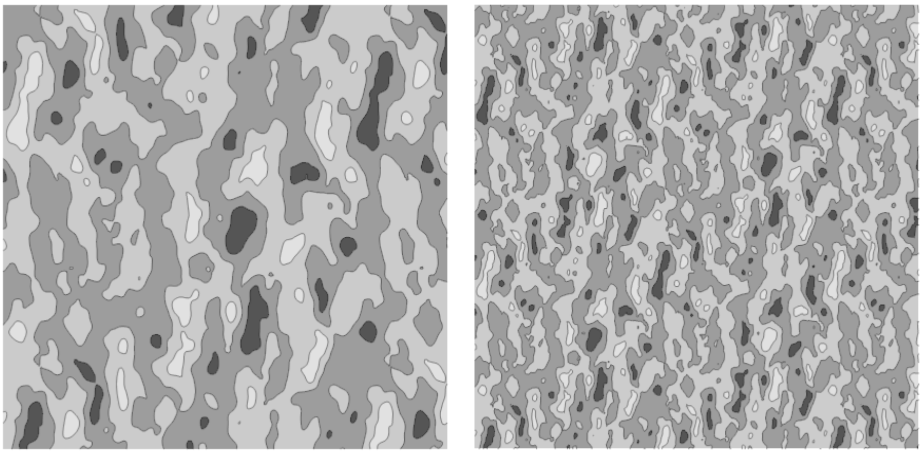


Figure 2: Realization of spatial spectral model (6), (7), (8) with the parameters $A = 0.1$, $n = 25$ (1,250 harmonics) of the sea surface with the spectrum , $v = 5$ m/sec, $\mu_{max} = 0.4$. The left picture presents the surface area of 2×2 km^2, the right picture presents the area of 4×4 km^2. One can easily see a periodic character of the pattern on the right.

from zero values, and the deviations themselves decrease in absolute value, Figure 3.

Let us proceed to randomized spectral models R1, R2. We consider these models for $A = 0.1$ for the same random field of the sea roughness with the spectrum with parameters $v = 5$ and $\mu_{max} = 0.4$. Introduce the following notations:

$$\delta^t_m(T) = \mathbf{M} \sup_{t \in (0,T)} |R^t(t) - R^t_n(t)|, \tag{11}$$

$$\delta^x_m(X) = \mathbf{M} \sup_{x \in (0,X)} |R^x(x) - R^x_n(x)|, \tag{12}$$

$$\delta^y_m(Y) = \mathbf{M} \sup_{y \in (0,Y)} |R^y(y) - R^y_n(y)|. \tag{13}$$

Here $R^t(t)$, $R^x(x) = R^{xy}(x, 0)$, $R^y(y) = R^{xy}(0, y)$ are normalized correlation functions of the simulated random field. The functions

$$R^t_n(t), \quad R^x_n(x) = R^{xy}_n(x, 0), \quad R^y_n(y) = R^{xy}_n(0, y)$$

are defined by expressions (9), mathematical expectation is taken over the joint distribution of the vectors λ_j and the index m equals 1 for the randomized model R1 with partitioning of its spectrum and equals 2 for the randomized model R2 without partitioning of the spectrum. Values (11) - (13) are presented in Table 2. Each value was calculated by Monte Carlo method from 100,000 independent realizations of the spectral model (in this case, the mean square deviation of the corresponding estimates is less than the values of the estimated parameters by three orders). In particular, the Table shows that the randomized model R1 with partitioning of its

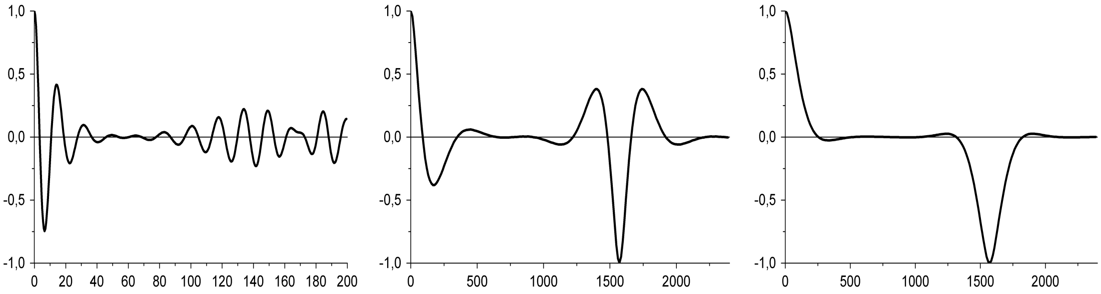


Figure 3: Normalized correlations functions R^t (a), R^x (b), R^y (c) of non-randomized spectral model N with the parameters $A = 0.1$, $n = 25$ (1,250 harmonics) for the random sea roughness field with the spectrum considered ($v = 5$ m/sec, $\mu_{max} = 0.4$).The values on the horizontal axes are given in seconds for R^t and in meters for R^x and R^y.

Table 1: The distances D^t, D^x, D^y, where spectral models (6), (7), (8) well represent (with the absolute error not exceeding 0.02) the normalized correlation functions R^t, R^x, R^y of the random sea roughness field with the spectrum with the parameters $v = 5$ and $\mu_{max} = 0.4$ for $A = 0.1$ depending on the number of harmonics. The values of the half-periods P^x, P^y of the spatial spectral models are presented as well:

n	number of harmonics	D^t, s	D^x, m	D^y, m	P^x, m	P^y, m
25	1250	90	1000	1300	1570	1510
50	5000	180	2550	2750	3150	3150
100	20000	420	5750	6000	6300	6300

spectrum represents the correlation structure of the simulated field more precisely than model R2 without partitioning, i.e., the values δ_1 are approximately half of the value δ_2.

Along with values (11) – (13), Table 2 presents the errors of reproduction of the correlation functions of the non-randomized spectral model N:

$$\Delta^t(T) = \sup_{t\in(0,T)} |R^t(t) - \check{R}^t_n(t)|, \tag{14}$$

$$\Delta^x(X) = \sup_{x\in(0,X)} |R^x(x) - \check{R}^x_n(x)|, \tag{15}$$

$$\Delta^y(Y) = \sup_{y\in(0,Y)} |R^y(y) - \check{R}^y_n(y)|, \tag{16}$$

where

$$\check{R}^t_n(t), \quad \check{R}^x_n(x) = \check{R}^{xy}_n(x,0), \quad \check{R}^y_n(y) = \check{R}^{xy}_n(0,y)$$

Table 2: The accuracy of representation of the correlation structure of the sea surface roughness with the spectrum considered ($v = 5$, $\mu_{max} = 0.4$) with respect to time and the spatial variables depending on the number of harmonics for spectral models R1, R2, and N).

n	number of harmonics	T, s	$\delta_1^t(T)$	$\delta_2^t(T)$	$\Delta^t(T)$
25	1250	60	0.082	0.174	0.016
50	5000	120	0.049	0.103	0.016
100	20000	240	0.028	0.058	0.015
n	number of harmonics	X, m	$\delta_1^x(X)$	$\delta_2^x(X)$	$\Delta^x(X)$
25	1250	750	0.081	0.170	0.017
50	5000	1500	0.048	0.102	0.017
100	20000	3000	0.029	0.058	0.016
n	number of harmonics	Y, m	$\delta_1^y(Y)$	$\delta_2^y(Y)$	$\Delta^y(Y)$
25	1250	750	0.081	0.156	0.011
50	5000	1500	0.046	0.094	0.011
100	20000	3000	0.026	0.054	0.011

are the correlation functions for model (6) – (8). Deviations (14) – (16) of the correlation functions for non-randomized spectral models are essentially less than the corresponding averaged values (11) – (13) for randomized models. At the same time, randomized models allow one to represent the spectrum and correlation structure of the simulated field exactly from an ensemble of implementations. This property is often useful when solving of applied problems (see [6]).

Conclusion

In this paper, we demonstrate the methodology of the error estimation for spectral models of Gaussian random fields on an example of the sea surface roughness. In particular, it was shown that a sufficiently good representation of the correlation structure of the sea surface in a region of several square kilometers in a period of several minutes may require dozens of thousands of harmonics.

Acknowledgements

The authors are grateful to Prof. G. A. Mikhailov and Prof. V. A. Ogorodnikov for their valuable remarks. The work was supported by the Russian Foundation for Basic Research (Project 15-01-01458), by the Basic Research Program N43 of the Presidium of the Russian Academy of Sciences "Fundamental Problems of Mathematical Modeling", and by the Program for Leading Scientific Schools of Russia (Grant NSh-5111.2014.1).

References

[1] Davidan I. M., Lopatukhin L. I., and Rozhkov V. A. (1985). *Wind-Driven Undulation in the World Ocean.* Gidrometeoizdat, Leningrad, [in Russian].

[2] Kargin B.A., Prigarin S.M. (1992). *Simulation of the sea undulation surface and study of its optical properties by Monte Carlo method.* Atmospheric and Oceanic Optics, Vol.5, No.3, pp.186-190.

[3] Krylov Yu.M. (1966). *Spectral Methods of Investigation and Calculation of Wind-Driven Waves.* Gidrometeoizdat, Leningrad [in Russian].

[4] Mikhailov G.A. (1978). *Numerical construction of a random field with given spectral density.* Doklady Akad. Nauk SSSR 238, No.4, pp.793–795, [in Russian].

[5] Mikhailov G.A. (1983). *Approximate models of random processes and fields.* Zh. Vychisl. Matem. Matem. Fiz. 23, No.3, pp.558–566, [in Russian].

[6] Mikhailov G.A. (1992). *Optimization of Weighted Monte Carlo Methods.* Springer-Verlag, Berlin.

[7] Prigarin S.M. (2001). *Spectral Models of Random Fields in Monte Carlo Methods.* VSP, Utrecht.

[8] Prigarin S.M. (2005). *Numerical Modeling of Random Processes and Fields.* Inst. of Comp. Math. and Math. Geoph. Publ., Novosibirsk, [in Russian].

[9] Prigarin S.M., Litvenko K.V. (2012). *Conditional spectral models of extreme ocean waves, Russian Journal of Numerical Analysis and Mathematical Modelling.* V.27, No.3, pp.289-302

[10] Shalygin A. S., Palagin Yu. I. (1986). *Applied Methods of Statistical Modelling.* Mashinostroenie, Leningrad, [in Russian].

[11] Tovstik P. E., Tovstik T. M., and Shekhovtsov V. A. (2012). *The influence of the form of the spectral density of a random undulation on oscillations of a stationary platform.* Vestnik St. Petersburg Gos. Univ. Ser. 1, No.2, pp.61-68.

Subgrid Modeling of Propagation of Acoustic Waves in Multiscale Random Media

Soboleva O.
Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk State Technical Univwersity, Novosibirsk , Russia
e-mail: olgasob@gmail.com

Abstract

In this paper we study effective coefficients in the problem of propagation of the acoustic waves in a multiscale heterogeneous medium. The correlation fields of the density and of the elastic stiffness have been mathematically represented by Kolmogorov's multiplicative cascades in three-dimensional space. The wavelength is assumed to be large as compared with the scale of heterogeneities of the medium. We obtain the effective acoustic equation using a subgrid modeling approach. Theoretical results are compared to the results of direct 3D numerical modeling.

Keywords: Subgrid modeling, multiplicative cascades, effective acoustic equation, random media.

Introduction

Wave propagation in heterogeneous media is a fundamental phenomenon of great scientific and practical interest. It is relevant to such important problems as detecting underground nuclear explosions, understanding the scale structure of oil, gas, and geothermal reservoirs. Seismic wave propagation and reflection are used not only to estimate the hydrocarbon content of a potential oil reservoir, but also the spatial distributions of its fractures, faults, and porosity [6]. In order to compute the displacement fields in an arbitrary medium, one must numerically solve a system of elasticity equations or an acoustic equation. The large-scale variations of coefficients as compared with wavelength are taken into account in these models with the help of some boundary conditions. The numerical solution of the problem with variations of parameters on all the scales requires high computer costs. The small-scale heterogeneities are taken into account by the effective parameters. In this case, equations are found on the scales that can be numerically resolved. The spatial geometry of small-scale heterogeneities is not exactly known. It has been shown that the irregularity of elastic parameters, density, permeability, porosity increases as the scale of measurements decreases [6], [3]. It is customary to assume these parameters to be random fields characterized by the joint probability distribution functions. However, it is difficult to measure higher order statistical moments for the geophysical parameters. At best, only the mean values and correlation functions of the second order are known. Geophysical parameters, for example, porosity, density, elastic modules can be well approximated by fractals and multiplicative hierarchical cascade models with non-Gaussian distributions [3]. The effective permeability coefficients were derived in [4]. The permeability was approximated by a hierarchical cascade model with log-normal and log-stable distributions. As the first step toward the goal of finding effective coefficients in the problem of propagation of elastic waves in strongly heterogeneous solids, in this paper we study the propagation of acoustic waves in the same type of media in which local elastic parameters have essentially all variations of scales from a a certain interval at each spatial point. In the present paper, the density of a medium and the elastic stiffness are approximated by a multiplicative continuous cascade. We obtain effective coefficients for the estimating the first statistical moment of the displacement in the acoustic equation if wavelength essentially exceeds maximum scale of heterogeneity, using the subgrid modeling method. If a medium is assumed to satisfy the improved Kolmogorov similarity hypothesis [2], the effective coefficients take especially a simple form. The derived formulas are verified by a direct numerical modeling.

1 Statement of the problem

The propagation of acoustic waves in heterogeneous medium is described by the equation

$$\rho\left(\mathbf{x}\right)\frac{\partial^2 u\left(\mathbf{x},t\right)}{\partial t^2}-\frac{\partial}{\partial x_i}\left(\lambda\left(\mathbf{x}\right)\frac{\partial}{\partial x_i}u\left(\mathbf{x},t\right)\right)=F(\mathbf{x},t), \tag{1}$$

where t is the time, $\mathbf{x}$ is the vector of spatial coordinates, $\rho(\mathbf{x})$ is the density of medium, $\lambda\left(\mathbf{x}\right)$ is the elastic stiffness, $u\left(\mathbf{x},t\right)$ is the displacement, $F(\mathbf{x},t)$ is the source with the dominant frequency ω_0 and the pulse width ω_1. Here and later, the summation over repeated indices is assumed. The wavelength is assumed to be large as compared with the maximum scale of heterogeneities L.

An increase in the randomness and intermittency in the behavior of the physical fields with a decrease in the scale of measurements has led to using hierarchic models for the physical parameters [4]. For the approximation of the coefficients $\rho(\mathbf{x})$, $\lambda\left(\mathbf{x}\right)$ we use the approach described in [4].

Let, for example, the field $\lambda\left(\mathbf{x}\right)$ be known. This means that the field is measured on a small scale l_0 at each point $\mathbf{x}$, $\lambda_{l_0}\left(\mathbf{x}\right)=\lambda\left(\mathbf{x}\right)$. Following Kolmogorov [2], we consider a dimensionless field ψ, which is equal to the ratio of two fields obtained by smoothing the field $\lambda\left(\mathbf{x}\right)_{l_0}$ on two different scales l,l'. Let $\lambda_l\left(\mathbf{x}\right)$ denote the parameter $\lambda_{l_0}\left(\mathbf{x}\right)$ smoothed on the scale l. Then $\psi(\mathbf{x},l,l')=\lambda(\mathbf{x})_{l'}/\lambda(\mathbf{x})_l$, $l'<l$. Expanding the field ψ into a power series in $(l-l')$ and retaining the first order terms of the series, at $l'\to l$, we obtain the equation:

$$\frac{\partial\ln\lambda_l(\mathbf{x})}{\partial\ln l}=\varphi(\mathbf{x},l), \tag{2}$$

where $\varphi(\mathbf{x},l')=(\partial\psi(\mathbf{x},l',l'y)/\partial y)\mid_{y=1}$. The small scale fluctuations of the field φ are observed only in the interval (l_0,L). The solution of equation (2) is as follows

$$\lambda_{l_0}(\mathbf{x})=\lambda_0\exp\left(-\int_{l_0}^{L}\varphi(\mathbf{x},l_1)\frac{dl_1}{l_1}\right), \tag{3}$$

where λ_0 is the constant. The field φ determines the statistical properties of the elastic stiffness. According to the central limit theorem for sums of independent random variables if the variance of $\varphi(\mathbf{x}, l)$ is finite, the integral in (3) tends to a field with a normal distribution as the ratio L/l_0 increases. If the variance of $\varphi(\mathbf{x}, l)$ is infinite and there exists a non-degenerate limit of the integral in (3), the integral tends to a field with a stable distribution. In this paper, it is assumed that the field $\varphi(\mathbf{x}, l)$ is statistically homogeneous with a normal distribution. The density coefficient $\rho(\mathbf{x})$ is constructed by analogy with the elastic stiffness coefficient:

$$\rho_{l_0}(\mathbf{x}) = \rho_0 \exp\left(-\int_{l_0}^{L} \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right). \tag{4}$$

The function $\chi(\mathbf{x}, l)$ is assumed to have the normal distribution and statistically homogeneous. For such field as the density, the cascade model must be the conservative model, i.e. the following equality should be satisfied

$$\langle \rho_l(\mathbf{x})\rangle = \rho_0, \tag{5}$$

for any scale l, where $\langle\rangle$ means statistical averaging. Condition (5) follows from physical essence of the field ρ. The measured on two different scales fields $\varphi(\mathbf{x}, l)$, $\chi(\mathbf{x}, l)$ are considered to be statistically independent

$$\begin{aligned}
&< \ \varphi(\mathbf{x},l)\ \varphi(\mathbf{y},l') > - < \varphi(\mathbf{x},l) >< \varphi(\mathbf{y},l') >= \Phi^{\varphi\varphi}(\mathbf{x}-\mathbf{y},l,l')\delta\left(\ln l - \ln l'\right) \quad (6)\\
&< \ \chi(\mathbf{x},l)\ \chi(\mathbf{y},l') > - < \chi(\mathbf{x},l) >< \chi(\mathbf{y},l') >= \Phi^{\chi\chi}(\mathbf{x}-\mathbf{y},l,l')\delta\left(\ln l - \ln l'\right),\\
&< \ \varphi(\mathbf{x},l)\ \chi(\mathbf{y},l') > - < \varphi(\mathbf{x},l) >< \chi(\mathbf{y},l') >= \Phi^{\varphi\chi}(\mathbf{x}-\mathbf{y},l,l')\delta\left(\ln l - \ln l'\right).
\end{aligned}$$

This supposition is usually assumed in the scaling models and reflects the decay of statistical dependence when the scales of fluctuations become different in the order of magnitude. The latter was proposed in [2]. To derive subgrid formulas to calculate effective coefficients, this assumption may be ignored. However, this assumption is important for the numerical simulation of the field ρ, λ. If the minimum scale l_0 in formulas (3), (4) tends to zero, the parameters tend to continuous multifractals. Hence the parameters are described by extremely irregular fields that are close to continuous multifractals. If the fields are statistically invariant to the scale transform, then the following equality is valid for any positive K:

$$\Phi^{\varphi\varphi}(\mathbf{x}-\mathbf{y},l) = \Phi^{\varphi\varphi}(K\left(\mathbf{x}-\mathbf{y}\right),Kl),\ \ \Phi^{\chi\chi}(\mathbf{x}-\mathbf{y},l) = \Phi^{\chi\chi}(K\left(\mathbf{x}-\mathbf{y}\right),Kl).$$

For simplicity, we use the same notation Φ in the right-hand side. Choosing $K = 1/l$, we obtain

$$\Phi^{\varphi\varphi}(\mathbf{x}-\mathbf{y},l) = \Phi^{\varphi\varphi}\left(\frac{\mathbf{x}-\mathbf{y}}{l}\right),\ \ \Phi^{\chi\chi}(\mathbf{x}-\mathbf{y},l) = \Phi^{\chi\chi}(\frac{\mathbf{x}-\mathbf{y}}{l}),$$

when $\mathbf{x} = \mathbf{y}$ the functions $\Phi^{\varphi\varphi}$, $\Phi^{\chi\chi}$ are equal to the constants $\Phi_0^{\varphi\varphi}$, $\Phi_0^{\chi\chi}$. If condition (5) is satisfied in scale-invariant medium, then $\Phi_0^{\chi\chi} = 2\left\langle\chi\right\rangle$.

2 Subgrid modeling

The density and elastic stiffness $\rho(\mathbf{x}) = \rho_{l_0}(\mathbf{x})$, $\lambda\left(\mathbf{x}\right) = \lambda_{l_0}\left(\mathbf{x}\right)$ are divided into two components with respect to the scale l. The large-scale (ongrid) components $\lambda\left(\mathbf{x}, l\right)$, $\rho\left(\mathbf{x}, l\right)$ are obtained, respectively, by statistical averaging over all $\varphi(x, l_1)$ and $\chi(x, l_1)$ with $l_0 < l_1 < l$, $l - l_0 = dl$, where dl is small. The small-scale (subgrid) components are equal to $\rho'(\mathbf{x}) = \rho(\mathbf{x}) - \rho(\mathbf{x}, l)$, $\lambda'(\mathbf{x}) = \lambda(\mathbf{x}) - \lambda(x, l)$. Applying (3), (4), (5) yields the formulas:

$$\begin{aligned}
\rho(\mathbf{x}, l) &= \rho_0 \exp\left[-\int_l^L \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right] \\
\rho'(\mathbf{x}) &= \rho(\mathbf{x}, l)\left[\exp\left[-\int\limits_{l_0}^{l} \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right] - 1\right], \ \langle \rho'(\mathbf{x})\rangle = 0, \\
\lambda(\mathbf{x}, l) &= \lambda_0 \exp\left[-\int_l^L \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\left\langle \exp\left[-\int_{l_0}^l \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\right\rangle \\
\lambda'(\mathbf{x}) &= \lambda(\mathbf{x}, l)\left[\frac{\exp\left[-\int\limits_{l_0}^{l} \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]}{\left\langle \exp\left[-\int\limits_{l_0}^{l} \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\right\rangle} - 1\right], \ \langle \lambda'(\mathbf{x})\rangle = 0.
\end{aligned} \tag{7}$$

From formulas (7) with the second order of accuracy in dl/l it follows that

$$\rho(\mathbf{x}, l) = \rho_l(\mathbf{x}), \ \lambda(\mathbf{x}, l) \simeq \left[1 - \langle\varphi\rangle\frac{dl}{l} + \frac{1}{2}\Phi^{\varphi\varphi}\left(0, l\right)\frac{dl}{l}\right]\lambda_l(\mathbf{x}). \tag{8}$$

$$\left\langle \lambda'\left(\mathbf{x}\right)\lambda'\left(\mathbf{x}'\right)\right\rangle \simeq \Phi^{\varphi\varphi}\left(\mathbf{x} - \mathbf{x}', l\right)\lambda\left(\mathbf{x}, l\right)\frac{dl}{l}, \ \left\langle \rho'\left(\mathbf{x}\right)\rho'\left(\mathbf{x}'\right)\right\rangle \simeq \Phi^{\chi\chi}\left(\mathbf{x} - \mathbf{x}', l\right)\rho\left(\mathbf{x}, l\right)\frac{dl}{l},$$

$$\left\langle \rho'\left(\mathbf{x}\right)\lambda'\left(\mathbf{x}'\right)\right\rangle \simeq \Phi^{\chi\varphi}\left(\mathbf{x} - \mathbf{x}', l\right)\rho\left(\mathbf{x}, l\right)\lambda\left(\mathbf{x}, l\right)\frac{dl}{l}.$$

Consider the temporal Fourier transform of equation (1)

$$\omega^2\rho\left(\mathbf{x}\right)u\left(\omega, \mathbf{x}\right) + \frac{\partial}{\partial x_i}\left(\lambda\left(\mathbf{x}\right)\frac{\partial}{\partial x_i}u\left(\omega, \mathbf{x}\right)\right) = -F. \tag{9}$$

The large-scale (ongrid) component of the displacement $u\left(\omega, \mathbf{x}, l\right)$ is obtained by averaging the solutions to equation (9)

$$\omega^2\rho\left(\mathbf{x}, l\right)u\left(\omega, \mathbf{x}, l\right) + \omega^2\left\langle \rho' u'\right\rangle + \frac{\partial}{\partial x_i}\left[\left(\lambda\left(\mathbf{x}, l\right)\frac{\partial}{\partial x_i}u\left(\omega, \mathbf{x}, l\right)\right) + \left\langle \lambda'\frac{\partial}{\partial x_i}u'\right\rangle\right] = -F. \tag{10}$$

The subgrid terms $\left\langle \lambda'\left(\mathbf{x}\right)\frac{\partial}{\partial x_i}u'\left(\mathbf{x}\right)\right\rangle$ in equation (10) are unknown. These terms cannot be neglected without preliminary estimation. The form of these terms in (10) determines a subgrid model. The subgrid terms are estimated using the perturbation theory. Subtracting system (10) from system (9) and taking into account only the first order terms of smallness obtain the subgrid equation:

$$\omega^2\rho\left(\mathbf{x},l\right)u'\left(\omega,\mathbf{x}\right)+\lambda\left(\mathbf{x},l\right)\frac{\partial^2 u'\left(\omega,\mathbf{x}\right)}{\partial x_j^2}=-\omega^2\rho'\left(\mathbf{x}\right)u\left(\omega,\mathbf{x},l\right)-\frac{\partial}{\partial x_j}\lambda'\left(\mathbf{x}\right)\frac{\partial u\left(\omega,\mathbf{x},l\right)}{\partial x_j}. \quad (11)$$

The variable $u\left(\omega,\mathbf{x},l\right))$ in the right-hand side of (11) is assumed to be known. Using the solution of equation (11) for isotropic media provided that $L^2\omega^2\rho\left(\mathbf{x},l\right)/\lambda\left(\mathbf{x},l\right)\ll 1$, we obtain

$$\omega^2\left\langle\rho'\left(\mathbf{x}\right)u'\left(\mathbf{x}\right)\right\rangle\simeq 0,\ \left\langle\lambda'\left(\mathbf{x}\right)\frac{\partial}{\partial x_i}u'\left(\mathbf{x}\right)\right\rangle\simeq-\frac{1}{3}\Phi^{\varphi\varphi}\left(0,l\right)\frac{dl}{l}\lambda\left(\mathbf{x},l\right)\frac{\partial}{\partial x_i'}u\left(\omega,\mathbf{x},l\right)(.12)$$

Substituting (12) into the ongrid equation(10) and given the formulas from (8) we arrive at the equation

$$\omega^2\rho\left(\mathbf{x},l\right)u\left(\omega,\mathbf{x},l\right)+\frac{\partial}{\partial x_i}\left(\lambda_{l0}\int_l^L\varphi(\mathbf{x},l_1)\frac{dl_1}{l_1}\frac{\partial}{\partial x_i}u\left(\omega,\mathbf{x},l\right)\right)=-F\left(\omega,\mathbf{x}\right),$$

$$\lambda_{l0}\ =\ \left(1-\frac{1}{3}\Phi^{\varphi\varphi}\left(0,l\right)\frac{dl}{l}\right)\left(1-\left\langle\varphi\right\rangle\frac{dl}{l}+\frac{1}{2}\Phi^{\varphi\varphi}\left(0,l\right)\frac{dl}{l}\right). \quad (13)$$

With the second order of accuracy in (dl/l) the coefficient λ_{l0} satisfies the equation

$$\lambda_{0l}=\left(1-\left\langle\varphi\right\rangle\frac{dl}{l}+\frac{1}{6}\Phi^{\varphi\varphi}\left(0,l\right)\frac{dl}{l}\right)\lambda_0$$

As $dl\to 0$, we obtain the effective equation for λ_{0l}, ρ_{0l}:

$$\rho_{0l}=\rho_0,\ \frac{d\ln\lambda_{0l}}{d\ln l}=\frac{1}{6}\Phi^{\varphi\varphi}\left(0,l\right)-\left\langle\varphi\right\rangle,\ \ \lambda_{0l_0}=\lambda_0 \quad (14)$$

In the scale-invariant media the solution of equation (14) has a simple form: $\lambda_{0l}=\lambda_0\left(\frac{l}{l_0}\right)^{\frac{1}{6}\Phi_0^{\varphi\varphi}-\left\langle\varphi\right\rangle}$.

By virtue of formulas (14) in the isotropic case, the form of the correlation functions does not affect on the effective coefficients. The anisotropic case study demands on knowledge of the form of the correlation functions. Wave propagation analysis in randomly layered media has shown that the form of the correlation function has a little effect on the effective coefficient [1]. A similar result have obtained for the filtration problem in the porous medium [7]. In the numerical calculations we use the correlation function

$$\Phi_1^{\varphi\varphi}=\Phi_0^{\varphi\varphi}\left(l\right)\exp\left[-\frac{\left(x_1'-x_1\right)^2+\left(\ x_3'-x_3\right)^2}{\alpha_1^2l^2}-\frac{\left(\ x_2'-x_2\right)^2}{\alpha_2^2l^2}\right]. \quad (15)$$

We assume, that $l_1=\alpha_1 l$ is the scale by coordinates $x_1,\ x_3,\ l_2=\alpha_2 l$ is the scale by coordinate x_2, and the mass density is constant. Taking into account equation

(1) in the limit $l \to l_0$, we come to the expression for the effective coefficients, which correctly describes the expectation of the displacement:

$$\frac{d\ln\lambda_{0l}^{i}}{d\ln l} = \frac{\Phi_0^{\varphi\varphi}(l)}{2} + \eta_{11}\Phi_0^{\varphi\varphi}(l) - \langle\varphi\rangle, \ i = 1,3, \ \frac{d\ln\lambda_{0l}^{2}}{d\ln l} = \frac{\Phi_0^{\varphi\varphi}(l)}{2} + \eta_{12}\Phi_0^{\varphi\varphi}(l) - \langle\varphi\rangle, \tag{16}$$

where $\lambda_{ef}^{i}(\mathbf{x}) = \lambda_{0l}^{i}\exp\left(-\int_l^L \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right)$ is the coefficient before $\frac{\partial u}{\partial x_i}$in equation (1). For $\alpha_1 > \alpha_2$ the coefficients are equal to:

$$\begin{aligned}
\eta_{11} &= \frac{1}{2}\frac{\alpha_1^2}{\left(\alpha_3^2-\alpha_1^2\right)}\Phi_0^{\varphi\varphi}(l)\left(\frac{\alpha_2}{2\sqrt{\alpha_2^2-\alpha_1^2}}\ln\frac{\alpha_2+\sqrt{\alpha_2^2-\alpha_1^2}}{\alpha_2-\sqrt{\alpha_2^2-\alpha_1^2}} - \frac{a_2^2}{\alpha_1^2}\right), \ i = j, \ i = 1,3 \qquad (17)\\
\eta_{12} &= \frac{\alpha_1^2}{\left(\alpha_2^2-\alpha_1^2\right)}\Phi_0^{\varphi\varphi}(l)\left[1 - \frac{\alpha_2}{2\sqrt{\alpha_2^2-\alpha_1^2}}\ln\frac{\alpha_2+\sqrt{\alpha_2^2-\alpha_1^2}}{\alpha_2-\sqrt{\alpha_2^2-\alpha_1^2}}\right], \quad i = 2.
\end{aligned}$$

If $\alpha_2 \to \alpha_1$, we obtain the isotropic case and $\eta_{11} = -\frac{1}{3}\Phi_0(l)$, $\eta_{12} = -\frac{1}{3}\Phi_0(l)$.

3 Numerical simulations

The following numerical problem was solved in order to verify the formulas obtained above. We have carried out the numerical simulation of the 3D problem by solving equation (1), using the finite-difference method (FD) with second-order discretization for time and the spatial variables. We used 512×1024×512 grids (where x_2 is the main direction of wave propagation). The domain of integration is separated into three subdomains. In the subdomains $0 < x_1 \le 512h$, $0 < x_2 \le 450h$, $0 < x_3 \le 512h$ and $0 < x_1 \le 512h$, $962h < x_2 \le 1024h$, $0 < x_3 \le 512h$ the coefficients ρ, λ are equal to $\rho = \rho_0 = 2000\text{kg/m}^3$, $\lambda = \lambda_0 = 1.8 * 10^{10}\text{Pa}$. On the plane boundaries $x_1 \times x_2$ at $x_3 = 0$, $x_1 \times x_2$ at $x_3 = 512h$ and $x_2 \times x_3$ at $x_1 = 0$, $x_2 \times x_3$, $x_1 = 512h$, the partial derivatives $\partial u(t, \mathbf{x})/\partial x_2$ are equal to zero; on the plane boundary $x_1 \times x_3$ at $x_2 = 1024h$ the displacement u is equal to zero. In the subdomain $0 < x_1 \le 512h$, $450h < x_2 \le 962h$, , $0 < x_3 \le 512h$ the spatial distributions of ρ, λ are simulated by formulas (3),(4), in which the integrals are approximated by the sums. The normal fields $\varphi(\mathbf{x}, l)$, $\chi(\mathbf{x}, l)$ are generated separately for each l using the method, described in [5]. We use the following pulse wave source: $f(t) = \left(1 - 2\pi^2(t_0 - t)^2\right)\exp\left(-\pi^2(t_0 - t)^2\right)$, where $t_0 = 0.8$, the dominant frequency is 1Hz. The pulse wave source is located at every node of the plane $x_1 \times x_3$ at $x_2 = 0$. Such a boundary condition has ensured the generation of a smooth initial wave front. Using a point source will not change the results that we present below, but it would require a large number of realizations for obtaining reliable statistics. We combine the spatial averaging over the planes $x_1 \times x_3$ for each value of x_2 with the ensemble averaging.

In Figures 1, 2 the averaged results obtained by the numerical modeling are compared with the solution of the effective equation and the solution obtained with the mean value of the coefficients ρ, λ in the subdomain $0 < x_1 \le 512h$, $450h < x_2 < 962h$, , $0 < x_3 \le 512h$. In Figures 1, one can see that the parameters of density do not affect the effective coefficient (curves 3, 4) that is consistent with the theoretical

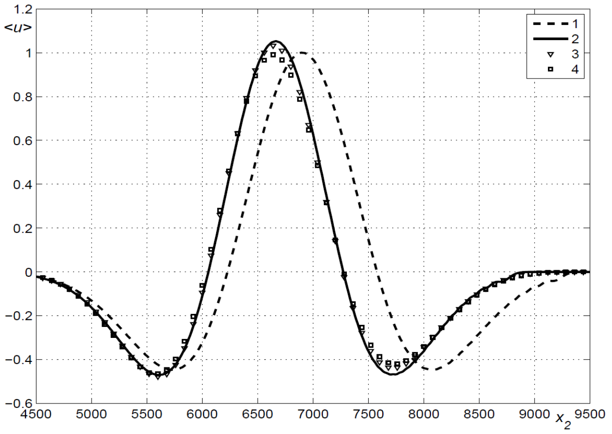


Figure 1: The isotropic case. The average of the displacement along the axis x_2, $\Phi_0^{\varphi\varphi} = 0.3$, $\varphi_0 = 0.15$, $t = 3.1 sec$. 1 –the result obtained for $\rho = \rho_0$, $\lambda = \lambda_0$; 2 – the result obtained by the effective equation; 3 – the result of numerical modeling with λ calculated by formula (3) for three scales, $\rho = \rho_0$. 4– the result of numerical modeling with ρ and λ calculated by formulas (3), (4) with the coefficient of correlation $\nu = 0.9$ for three scales: $l_j = 8h, 16h, 32h$. A minimum scale is $l_0 = 1/64$ of the wavelength, while a maximum scale is $L = 1/16$.

prediction. The results were averaged over 45 realizations. Curves 2, 3 in Figures 1, 2 slightly differ in magnitude from curve 1 for one wavelength. Such deviations will have a significant influence over a distance containing many wavelengths. Figure 2 presents the results obtained for $\rho = \rho_0$ and the anisotropic λ.

Conclusion

We have presented the effective coefficients for the wave equation if its parameters are described by extremely irregular small-scale fields that are close to multifractals. The multifractals can be obtained if a minimum scale l_0 in formulas (3), (4) tends to zero. To approximate the medium, we have started from the modified Kolmogorov theory in terms of the ratios of smoothed fields. As a minimum scale is finite, any singularities are absent, therefore we use only the theory of differential equations and the theory of stochastic processes. We have shown that small-scale heterogeneities affect the acoustic wave propagation, and indicate that parameters distribution of density does not affect the average displacement in the first order of scale heterogeneities. The numerical testing illustrates the efficiency of the approach proposed when the scales of heterogeneities are much less than the size of the wavelength.

The paper was supported by the RFBR N15-01-01458.

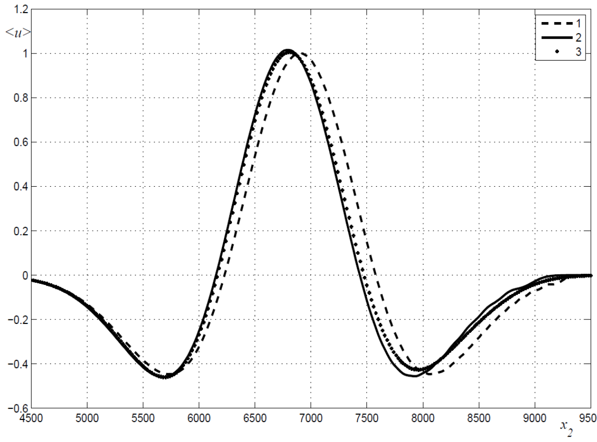


Figure 2: The anisotropic case. The average of the displacement along the axis x_2, $\Phi_0^{\varphi\varphi} = 0.45$, $\varphi_0 = 0.225$, $t = 3.1 sec$; $\alpha_1/\alpha_2 = 4$, $\rho = \rho_0$; λ calculated by formula (3) for two scales: 1/64, 1/32 of the wavelength along the axes x_1, x_3 , 1/16, 1/8 along the axes x_2. 1 – the result obtained for $\rho = \rho_0$, $\lambda = \lambda_0$; 2 –the result obtained by the effective equation; 3 – the result of numerical modeling.

References

[1] Fouque J.P, Garnnier J., Papanicolaou G., Solna K.(2007) *Wave Propagation and Time Reversal in Randomly Reversal in Randomly Layered Media.volume 56 of Stochastic Modelling and Applied Probability*. Springer.

[2] Kolmogorov A. N. (1962) A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number*Journal Fluid Mech.*, Vol. **13**, pp. 82-85.

[3] Koohi lai Z., Vasheghani Farahani S., Jafari G.R. (2013). Non-Gaussianity effect of petrophysical quantities by using q-entropy and multi fractal random walk. *Physica A.* Vol. **392**, pp. 3039–3044.

[4] Kuz'min G.A., Soboleva O.N. (2002). Subgrid modeling of filtration in porous self-similar media. *App. Mech. and Tech. Phys.* Vol. **43**, pp. 583–592.

[5] Ogorodnikov V. A., Prigarin S. M. (1996) *Numerical Modeling of Random Processes and Fields: Algorithms and Applications*, Utrecht, The Netherlands.

[6] Sahimi M. (1993). Flow phenomena in rocks: from continuum models, to fractals, percolation, cellular automata, and simulated annealing. *Rev. Modern Phys.* Vol. **65**, pp. 1393–1534.

[7] Shvidler M. I. (1985) *Statistical Hydrodynamics of Porous Media.* Nedra, Moscow, in Russia.

Numerical Statistical Modeling of the Thermal State of Aircraft Honeycomb coatings

Sergey A. Gusev[1] and Vladimir N. Nikolaev[2]
[1] *ICM&MG SB RAS, NSTU, Novosibirsk, Russia*
[2] *S.A. Chaplygin Siberian Aeronautical Research Institute, Novosibirsk, Russia*
e-mail: sag@osmf.sscc.ru, nikvla50@mail.ru

Abstract

The paper is devoted to construction a method for evaluation of the thermal state of the honeycomb structures which are part of the aircraft fuselage. The considered problem is described by a boundary value problem for a parabolic equation with discontinuous coefficients. The generalized solution of this problem can be approximated by a solution of a parabolic equation with smoothed coefficients. The smoothing coefficients in the paper is made by the integral averaging. The statistical estimation of the solution of the problem with smoothed coefficients is obtained by using the numerical solution of stochastic differential equations.

Keywords: heterogeneous structures, parabolic boundary value problem, discontinuos coefficients, integral averaging, stochastic differential equations.

Introduction

Using inhomogeneous structures such as cell is a promising direction in aeronautical engineering. This is due to the fact that these structures combine the properties most suitable for creating modern aircraft: lightweight, strength and low thermal conductivity. Some applications honeycomb structures can be found in [1], [2]. This paper is devoted to construction a method for evaluation heat transfer in structures such as honeycomb. We consider sealed panels, whose sheets and cell frame is made of carbon fiber and the maximum size of a cell channel does not exceed 1 cm. The heat transfer in the panel can be described by a parabolic boundary value problem with discontinuous coefficients. It is known that this kind of problem has a unique generalized solution in the sense of satisfying to the integral identity [3]. At the same time this generalized solution can be approximated by a solution of problem whose coefficients are approximated by smooth functions. The approximation (smoothing) of the coefficients is done by the integral averaging with a smooth compactly supported kernel. The statistical estimating this approximate solution is done by numerical solution of stochastic differential equations (SDE).

1 Mathematical modeling of heat transfer in heterogeneous bodies

Heterogeneous body is a body consisting of a mixture of chemically different substances. A typical example of a heterogeneous body is concrete. A homogeneous substance containing air voids is also heterogeneous body. The description and physical properties of heterogeneous media can be found in [4]. We consider as mathematical model of heat exchange in heterogeneous bodies as a parabolic boundary value problem with discontinuous coefficients.

We introduce the following denominations: $G \subset \mathbb{R}^3$ is a bounded domain with the boundary ∂G, and G is partitioned into M subdomains $G = \bigcup\limits_{k=1}^{M} G^{(k)}$; $Q_T = G \times (0,T)$ is a cylinder in $\mathbb{R}^4$; $S_T = \partial G \times [0,T]$ is the lateral surface of the cylinder. It is assumed that the domains $G^{(k)}$ are separated by piecewise smooth surfaces Γ.

The heat transfer in the heterogeneous body is described by the following parabolic equation

$$\frac{\partial u}{\partial t} - \sum_{i,j=1}^{3} a_{ij}(x,t) \frac{\partial^2 u}{\partial x_i \partial x_j} = 0, \quad (x,t) \in Q_T \quad , \tag{1}$$

where the coefficients a_{ij} are Lipshits continuous with respect to x in subdomains $G^{(k)}$, k=1,..., M. At the same time, a_{ij} can be discontinuous on Γ. It is assumed also that there exist $\mu, \eta > 0$ that the following condition holds uniformly with respect to $(x,t) \in Q_T$

$$\mu \sum_i \xi_i^2 \le \sum_{i,j} a_{ij}(x,t) \xi_i \xi_j \le \eta \sum_i \xi_i^2 .$$

We require that the unknown function u must satisfy to the following conditions:

$$u|_{t=0} = \phi(x) \tag{2}$$

and one of the following two boundary conditions on ∂G:
the first boundary condition

$$u(x,t)|_{x \in \partial G} = \psi(x,t) \tag{3}$$

or the third one

$$\left. \sum_{i,j} a_{ij} n_i \frac{\partial u}{\partial x_j} + \eta(x,t) u + \gamma(x,t) \right|_{x \in \partial G} = 0 \quad , \tag{4}$$

where n_i is the i-th coordinate of an inward normal vector on ∂G.

It is proved in [3] the existence of a generalized solution of the problems (1) – (3) or (1), (2), (4). This solution can be approximated by a solution of the boundary value problem for a parabolic equation with coefficients that are approximations of the initial discontinuous coefficients. Thus, we can obtain an approximate solution of the problem by suitably smoothing discontinuous coefficients. In the paper we estimate of the approximate solution of the problem with smoothed coefficients by using a statistical method based on numerical solution of SDE. To smooth the coefficients we use the integral averaging [5]

$$f^{(\rho)}(x)=\rho^{-3}\int\limits_{|x-y|<\rho}\omega(|x-y|)f(y)dy \tag{5}$$

with an infinitely differentiable averaging kernel such that $\omega(|\xi|)=0$ when $|\xi|\geq 1$ and $\int\limits_{|\xi|\leq 1}\omega(|\xi|)d\xi=1$.

2 Estimating solutions of parabolic problems

It is well known (see, for example [6]) that the solution of a parabolic equation can be represented as expectation of functional of SDE solution. This fact is often used to obtaining statistical estimates of solutions of parabolic equations by using numerical solution of SDE. One advantage of this method is that you do not need to build a grid in the spatial variables and solving large systems of linear algebraic equations. We apply this method to estimating solution of the parabolic equation with smoothed coefficients by the integral averaging (5).

We find approximate solution of (1) as solution of the following equation

$$\frac{\partial u}{\partial t}-\sum_{i,j=1}^{3}a_{ij}^{(\rho)}(x,t)\frac{\partial^2 u}{\partial x_i\partial x_j}=0 \quad , \tag{6}$$

where $a_{ij}^{(\rho)}$ are smoothed coefficients of (1) at the neighborhood of Γ.

For a point $(x,t)\in Q_T$ we define a random process $X_\bullet$ which is a solution of the following vector SDE

$$X_v=x+\int\limits_{T-t}^{v}\sigma\left(X_r,r\right)dW_r \quad , \tag{7}$$

where $W_\bullet$ is a Wiener process, σ is a 3×3 matrix such that $2\sigma\sigma^T=A$, $A=(a_{ij}^{(\rho)})$.

Let us denote by $\mathrm{E}_{t,x}$ the mathematical expectation with respect to the probability measure $P_{t,x}$ corresponding to a random process that begins from the point x at the time point t. Then the solution of (6) at the point (x,t) satisfying the conditions (2), (3) can be obtained by the formula

$$u\left(x,t\right)=\mathrm{E}_{T-t,x}[\phi\left(X_T\right)\mathbf{1}_{\tau>T}+\psi\left(X_\tau,\tau\right)\mathbf{1}_{\tau<T}]_, \tag{8}$$

where $\tau=\inf(v|X_v\notin G)$ is the first exit time of the process $X_\bullet$ from G, $\mathbf{1}_S$ is the indicator function of the set S.

So, we can obtain estimates of the solution of the problem (6), (2, (3) by modeling trajectories of the process $X_\bullet$ numerically. For this purpose we use the Euler method, according to which approximate trajectories of $X_\bullet$ are calculated as ([7])

$$x_{i+1}=x_i+\sqrt{h}\sigma(x_i,t_i)\zeta_i, \quad t_{i+1}=t_i+h \quad , \tag{9}$$

where h is the integration step, ζ_i are 3D vector with independent N(0,1) random variables.

In the case of the problem (6), (2), (4) we construct a reflected diffusion process $X_\bullet$ in the form ([8])

$$X_v = x + \int\limits_{T-t}^{v} \sigma\left(X_r, r\right) dW_r + \int\limits_{T-t}^{v} n^A\left(X_r, r\right) d|k_r|, \tag{10}$$

where n^A is the normalized inner co-normal vector, i.e. $n^A = An/|An|$, $|k_v| = \int\limits_t^v \mathbf{1}_{\partial G}(X_r) d|k_r|$ is a nonnegative stochastic process that increase when the process $X_\bullet$ is on the boundary. Then the solution of (6) at the point (x,t) satisfying the conditions (2), (4) can be expressed in the form of the following expectation

$$u\left(x,t\right) = \mathrm{E}_{T-t,x}\left[\phi\left(X_T\right)\exp\left(\int\limits_{T-t}^{T} \eta\left(X_r, r\right) d|k_r|\right) + \int\limits_{T-t}^{T} \gamma\left(X_r, r\right) d|k_r|\right]. \tag{11}$$

We obtain estimates of the solution of the problem (6), (2), (4) by numerical modeling trajectories of the process $X_\bullet$. For this purpose we use the Euler method in the form

$$x_{i+1} = x_i + \sqrt{h}\sigma(x_i, t_i)\zeta_i + \left(\Delta_{i+1}K\right) n_i^A, \quad t_{i+1} = t_i + h\ , \tag{12}$$

$$\Delta_{i+1}K = \left[d\left(x_i + \sqrt{h}\sigma(x_i, t_i)\zeta_i\right)\right]^- \ , \tag{13}$$

where n_i^A is the unit inner co-normal vector at the point x_i, if x_i is on ∂G; $[a]^- = \max\{0, -a\}$; $d(x)$ is a nonpositive real function satisfying for any point $x \notin G$ to the following equation

$$\mathrm{x{=}\rho(x){+}d(x)n^A(\rho(x)).} \tag{14}$$

We use in (14) the following designations: $\rho(x)$ is a projection of a point $x \notin G$ on $\overline{G}$ in the conormal direction, $n^A\rho(x)$ is the unit inner conormal vector at $\rho(x)$. Note that $d(x) = 0$, if $x \in \overline{G}$.

To obtain an estimate of the solution of the problem (6), (2), (4) we also need to compute the following functions defined in the time grid nodes

$$y_i = \begin{cases} 1, \quad i = 0, \\ \exp\left(\sum\limits_{k=0}^{i-1} \eta(x_k, t_k)\mathbf{1}_{\partial \mathrm{G}}(x_k)\Delta_{k+1}K\right), i \geq 1, \end{cases}$$

$$z_i = \begin{cases} 0, \quad i = 0, \\ \sum\limits_{k=0}^{i-1} \left(\gamma(x_k, t_k)\mathbf{1}_{\partial G}(x_k)\Delta_{k+1}K\right) y_k, i \geq 1. \end{cases}$$

Estimation of the solution of the problem (6), (2), (4) is calculated by the formula

$$\hat{u}(x,t) = \mathrm{E}_{T-t,x}\left[\phi\left(x_N\right) y_N + z_N\right], \tag{15}$$

where $N = {}^{t}\!/_{h}$.

3 Calculation of heat transfer in honeycomb structures

A honeycomb panel consists of two sheets and a honeycomb core between them filled with a filler of low thermal conductivity. We considered a honeycomb panel which frame made from carbon that filled with air Fig.1. Air is the thermal-protective filler of the panel.

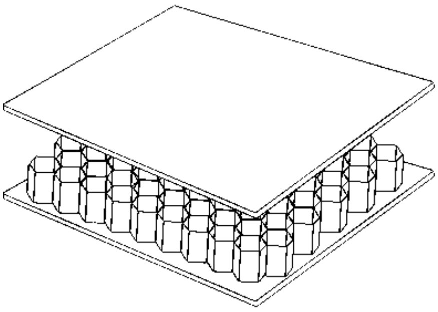


Figure 1: Honeycomb structure

This panel has the following characteristics: the total thickness of the panel is 0.035 m; the thicness of each sheet is 0.001 m; the thickness of the cell frame is $6 \cdot 10^{-5}$m; the edge length of the cell is 0.0042 m; the thermal diffusivities of carbon and air are equaled $8 \cdot 10^{-4}$ m^2/s and $2.36 \cdot 10^{-5}$ m^2/s correspondingly. The effective thermal conductivity of this honeycomb panel was by thermophysical experiments, and it is equal to 0.08 $W/(m \cdot K)$. The thermal diffusivity of the honeycomb panel was also determined and it is equaled to $1 \cdot 10^{-3}$ m^2/s.

The calculation and experiment were carried out for the boundary conditions of the third kind for the cold type of climate Fig. 2.

The heat transfer equations of the skin multi-layer honeycomb for boundary conditions of the third kind are represented as one-dimensional heat-conduction equations describing the heat transfer process in the multi-layer honeycomb:

$$C(x)u_i = (\lambda(x,u)u_x)_x, \quad 0 < x < l; \tag{16}$$

$$\begin{gathered}\lambda(0,u)F_k u_x = \alpha_{k,out}(t)F_k(u(t,0) - u_e(t)) + Q_{k,out} + Q_{k,in} - \\ -\sigma_0 \varepsilon_{k,out} F_{k,out} u^4(x), \quad x = 0;\end{gathered} \tag{17}$$

$$\begin{gathered}\lambda(l,u)F_k u_x = \alpha_{k,in}(t)F_k(u_e(t) - u_{air}(t,0)) + \sum_{i=1}^{k} G_{i,k} u_i^4/u_{ref}^4 - \\ -\sigma_0 \varepsilon_{k,in} u^4(t) + Q_{k,out} + Q_{k,in}, \quad x = l;\end{gathered} \tag{18}$$

$$u(0,x) = u_0(x), \quad 0 < x < l, \tag{19}$$

where $C(x) = C_i, \lambda(x,t) = \lambda_{i,0} + \lambda_{i,1}u,\ l_{i-1} \leq x < l_i,\ (i = 1, \ldots, k-1), \quad C(x) = C_k, \lambda(x,t) = \lambda_{k,0}+\lambda_{k,1}u$ with $l_{k-1} \leq x \leq l_k$, i. e. the coefficients C, λ depend on which layer is examined concerning the heat transfer. Here $0 = l_0 < l_1 < \ldots < l_k = l$. In equations (1) - (4) the following notations are used: $C(x)$ is the multi-layer honeycomb of the skin or windows volumetric heat capacity (the product of specific heat capacity by density); $\lambda(l,u)$ is the heat-conduction coefficient of the multi-layer structure; $\alpha_{k,out}$ is the heat transfer coefficient of the outer structural surface; $\alpha_{k,in}$ is the heat transfer coefficient of the inner structural surface; F_k is the area of the construction at the outer and inner heat transfer; $Q_{k,out}$ are the heat energy of external sources; $Q_{k,in}$ are the heat energy of internal sources; σ_0 is the Stefan-Boltzmann constant; $\varepsilon_{k,in}$ is the emissivity of the multi-layer structure inner surface; k is the number of blocks in the section; $G_{i,k}$ is the radiation transfer coefficient of the system; u_e is the recovery temperature; u_{air} is the air temperature in the section or in the part of a section; $u(x,t)$ is the temperature of the multi-layer structure; u_{ref} is the reference temperature; u_i is the temperature of the i-th unit of the section; u_x is first order derivative of u; $u_{x,x}$ is the second order derivative of u; l is the thickness of the multi-layer structure.

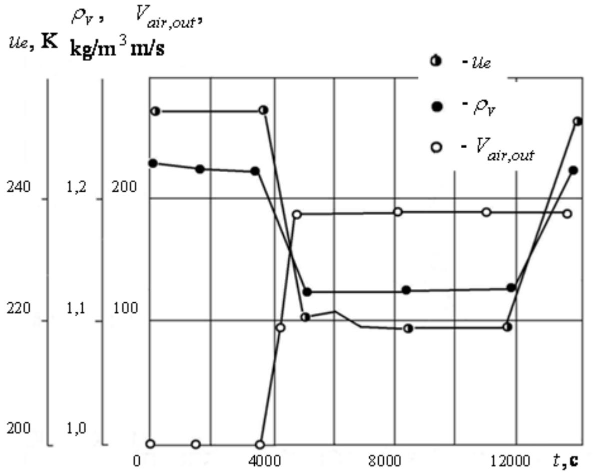


Figure 2: Parameters of the flight mode and the ambient air overboard for the cold type climate: u_e is the recovery temperature; ρ_V is the density of the ambient air overboard; $V_{air,out}$ is the airspeed

The heat transfer coefficient of the multi-layer structure outer surface and the heat transfer coefficient of the multi-layer structure inner surface will be calculated according to the procedures described respectively in the works [9] and [10]. The values of the heat transfer coefficient on the skin outer side for the first 150 s of the flight are shown in Fig. 3. The temperature of the ambient air at the inner surface of the skin was constant at 283 K. At initial conditions the linear temperature distribution over the honeycomb thickness was accepted. The beginning of the Cartesian coordinate system is a point that is located on the panel bottom edge (the inside of the aircraft skin) in the center of the plate, which coincided with the center of the hexagon, was taken. Axes and are located in the plane of the lower edge of the plate, bounding the honeycomb frame, the axis is directed from the panel bottom edge toward the panel upper edge (the outside of the skin).

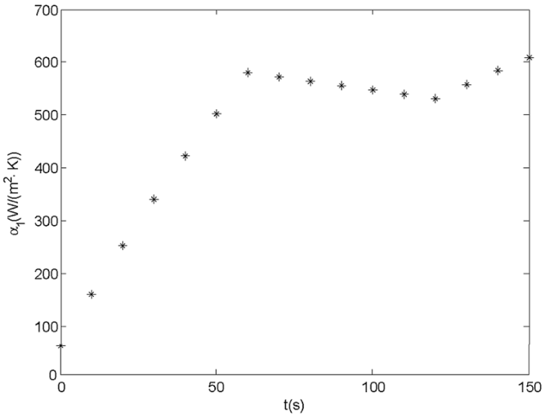


Figure 3: The heat transfer coefficient on the skin outer side

As the averaging kernel at the smoothing of discontinuous coefficients the function $\omega_\rho(x) = \gamma_\rho(\rho^2 - |x|^2)$ was used in the calculation, where - γ_ρ was a normalizing factor. For the panel thermal state calculation the parallel program in Fortran 90 was developed. The parallelization in the program is carried out under the scheme Master-Slave. In this scheme, one compute core is considered the main one, and it distributes the full amount of work on the simulation of random paths on all cores involved in the work. Upon completion of all paths simulating, all cores pass the calculation results to the leading core to compute the functional expectation, giving an assessment of temperature. When writing the parallel program the software Intel MPI, Version 4.1 was used.

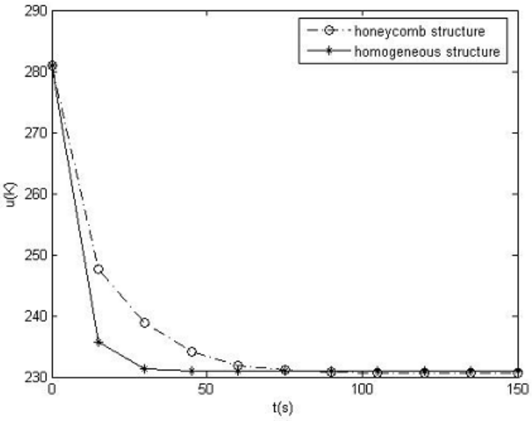


Figure 4: The temperatures on the inner side of the insulating casing

Simulating of the random process paths was carried out using the Gaussian random variables parallel generator from the library Intel MKL [11]. The calculations were performed in the Siberian Supercomputer Center on the hybrid cluster HKC-30T + GPU with the use of 36 quad-core processors E5540 on 2,53 GHz. Temperature calculations were performed near the panel bottom edge at the coordinates where 0, 0.0001 m. The step size in the Euler's method was taken and the sample number - 4000 of random paths. Yet the confidence interval of the honeycomb temperature did not exceed 1.2 K at a confidence level of 95% in non-stationary and 0.7 K in stationary conditions. Firstly we estimated temperature values of the honeycomb panel using our method. After that, analogous calculations were performed for the panel considered as homogeneous with experimentally obtained value of effective thermal conductivity. The obtained results are demonstrated in Fig. 4.

Acknowledgements

This work was supported by Russian Leading Science Schools Programme grant 5111.2014.1, Russian Foundation for Basic Research grant 14-01-00340-a

References

[1] Koch L.C., Pagel L.L. (1978). High heat flux actively cooled honeycomb sandwich structural panel for a hypersonic aircraft. *NASA Contractor Report*, No. 2959, p. 161.

[2] Tsihosh E. (1983). *Supersonic aircrafts: Reference manual. Trans. from Polish.* Mir, Moscow.

[3] Ladyzhenskaya O.A., Solonnikov V.A., Uraltseva N.N. (1967) *Linear and quasi-linear equations of parabolic type.* Nauka, Moscow.

[4] Misnar A. (1968) *The thermal conductivity of solids, liquids, gases and their compositions.* Mir, Moscow.

[5] Sobolev S.L. (1988) *Applications of functional analysis in mathematical physics.* Nauka, Moscow.

[6] Gikhman I. I., Skorokhod A. V. (1977) *Introduction to the theory of random processes.* Nauka, Moscow.

[7] Kloeden P. E., Platen E. (1992) *Numerical solution of stochastic differential equations.* Springer-Verlag, Berlin.

[8] Saisho Y. (1987) Stochastic differential equations for multi-dimensional domain-with reflecting boundary. *Probab. Theory Rel. Fields.* Vol. **74**, pp. 455-477.

[9] Voronin G. I. (1973) *Air-conditioning systems onboard the aircrafts.* Mashinostroenie, Moscow.

[10] Dulnev G. N., Tartakovsky N. N. (1971). *Thermal conditions of electronics.* Energy, Leningrad.

[11] Handbook of Intel MKL:
http://software.intel.com/sites/products/documentation/doclib/mkl_sa/11/
/mklman/index.htm

Research of Wiener Type System Nonparametric Models

N.V. Koplyarova[1] and A.V. Medvedev[2]
[1] *Siberian Federal University*
[2] *Siberian State Aerospace University*
Krasnoyarsk, Russian Federation
e-mail: koplyarovanv@mail.ru

Abstract

The task of nonlinear dynamical systems of Wiener type identification is considered in this thesis. The linear dynamical element of the system is in nonparametric uncertainty conditions. The type of nonlinearity is assumed to be known with the set of vector of parameters. The systems with a quard and link saturation nonlinearity are considered. The proposed method of dynamic objects modeling is based on the nonparametric estimation of linear and nonlinear parts of the system. Presented algorithm allows to design the models, that describes the system with sufficient accuracy.

Keywords: nonlinear system, nonparametric, Wiener model, estimation, dynamical systems.

Introduction

The problem of nonlinear dynamical system identification is one of the most important one in the theory of control. In spite of the existing a lot of methods for dynamical systems identification, there is no universal theory that allows to design the models of such systems.

Most of the methods of nonlinear system identification are difficult to apply in practice or they do not take into account all the properties of the investigated object. Besides, the task of identification in the most methods is considered “in the narrow sense”, it is corresponds to the case when the object structure is known with a vector of parameters[1]. In this paper the dynamic systems identification "in the broad sense" is considered. In this case the parametrization of the investigated object model is not available or one can partially parameterized the model on the base of available a priori information. We consider the nonlinear systems in the form of a sequence connected linear dynamic and nonlinear static blocks. A structure and parameters of linear dynamic block of such system is unknown, but the type of the nonlinear element is known with the set of parameters. Thus, we consider the problem of modeling of the nonlinear dynamical processes under conditions of partial parameterization of model structure.

1 Identification problem

We consider the nonlinear system in the form of a sequence connected linear dynamic and nonlinear blocks. Such systems are called a models of Wiener type [2]. It is required to design the mathematical model of the stochastic object according to the measures of process. That would describes objects behavior at arbitrary input effects and additive noise the presence on the output. The total scheme of nonlinear dynamical system identification is shown in Fig. 1.

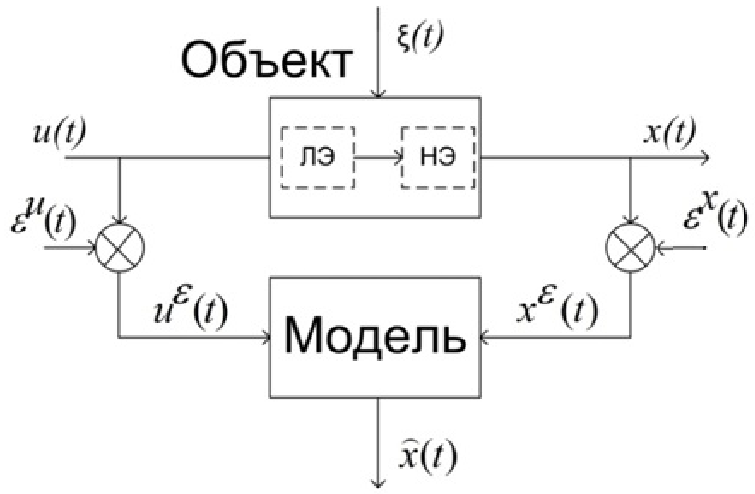


Figure 1: The general scheme of the identification problem

where $u(t)$ and $x(t)$- input and output variables of the object, u_t^ξ, x_t^ξ appropriate observation of process variables, $\xi(t)$ – unobserved random effects, $\varepsilon^u(t), \varepsilon^x(t)$ – random noise in measure channels, $\hat{x}(t)$ – output of the object model. Available priory information is uneven sample of input and output variables of the object's measures of s size – $\{u_i, x_i, i = 1, s\}$.

The structure of the linear dynamical part of the system is unknown. The common type of the nonlinear function is assumed to be known with the set of parameters. In the paper the following nonlinear elements are considered: quard and link saturation.

Problem of nonlinear system identification can be divided into two tasks. At first part we construct a nonparametric model of linear dynamical block of system. Then, on the base of some estimation, a model of nonlinear dymanical system can be designed.

2 Nonparametric identification of Wiener type system

Lets consider the system that can be represented as a model of Wiener type (Fig. 2)[3]

Let the order and parameters of a linear dynamic block of the system are unknown, and the nonlinear element structure is defined up to a set of parameters α. The

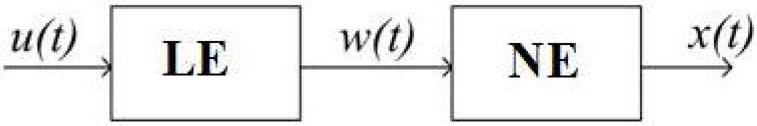


Figure 2: Wiener model, LE -– linear dynamical and NE – nonlinear parts of the system, $u(t)$, $x(t)$ – input and output action, $w(t)$ – intermediate part output(is not measured)

main idea of the proposed algorithm for constructing a such system models is to use a nonparametric estimates to describe a system relations, details of which, for some reason are unknown (in this case a linear dynamic block output $w(t)$ is not measured), and to apply a nonparametric approach for functions estimation.

As we can see from the figure, the object output is calculated as a function of the $w(t)$ value:

$$x(t) = f(w(t), \alpha), \tag{1}$$

where $x(t)$ – the nonlinear object output, $w(t)$ – the linear part output (is not measured); $f(w)$ – nonlinear function this unknown parameter α.

The reaction of linear dynamical block of system $w(t)$ to the input signal $u(t)$ is described with the Duhamel integral[3]:

$$w(t) = k(0)u(0) + \int_0^t h(t-\tau)u(\tau)d\tau = k(0)u(0) + \int_0^t k\prime(t-\tau)u(\tau)d\tau, \tag{2}$$

where $h(t)$-impulse response (weight function) of the system, and $k(t)$-step response of this system (transient function).

That, the model is:

$$x(t) = f\left(\int_0^t h(t-\tau)u(\tau)d\tau, \alpha\right). \tag{3}$$

The mathematical model of the nonlinear object can be represented as the equation (3), in that, instead of the weight function $h(t)$ and the nonlinearity parameters α are used their statistical estimations. To obtain this estimation it is necessary to generate the sample $\{u_i, w_i, i = 1, s\}$.

A response function value $k(t)$ represents a reaction of linear dynamical system to the step input action $u(t) = 1$, that is $k(t) = w(t/u(t) = 1)$. But the value of $w(t)$ is not available for measurement. After feeding to the system input the Heviside function $u(t) = 1$, it is possible to measure only the output of nonlinear process $x(t)$, which will have a value of $x(t) = f(h(t))$.

In the case when for some classes of nonlinear elements, the equation (1) can be solved for $w(t)$,we have[2]:

$$k(t) = w(t) = f^{-1}(x_1(t), \alpha), \tag{4}$$

where k_i – calculated values of linear dynamic element step response, $f^{-1}(x(t))$ – the inverse function to the description of the nonlinear element, x_i^1 – sample values of the investigated object output, if the input action is equal to $u(t) = 1(t)$, α – parameter of the nonlinearity function.

Further, on the base of the sample of discrete values can be estimated step response function of the system in the form of stochastic approximation nonparametric regression type as follows:

$$k_s(t) = \frac{1}{sC_s} \sum_{i=1}^{s} k_i H\left(\frac{t-t_i}{C_s}\right), \tag{5}$$

where k_i – sample values of step response of linear dynamical system (LDS), $H()$ – Kernel function and c_s – bandwidth parameter are satisfied the conditions of convergence[4].

$$c_s > 0, s = 1, 2..., \lim_{s\to\infty} c_s = 0, \lim_{s\to\infty} cc_s = \infty,$$

$$\int_{\Omega(u)} H\prime(u)du = 0, c_s \int_{\Omega(u)} H\prime(u)udu = -1, u = \frac{\tau - t}{c_s}, \tag{6}$$

$$\lim_{s\to\infty} c_s^{-1} H\left(\frac{\tau-t}{c_s}\right) = \delta(\tau - t),$$

The weight function of the system $h(t)$ is a time derivative of the step response function $k(t)$, i.e. $h(t) = k\prime(t)$. Therefore the nonparametric estimation of the impulse function can be described as follows:

$$h_s(t) = k\prime(t) = \frac{1}{sC_s} \sum_{i=1}^{s} k_i H\prime\left(\frac{t-t_i}{C_s}\right). \tag{7}$$

Linear dynamic block of the system can be described with the following mathematical formula:

$$w_s(t) = \frac{1}{sC_s} \sum_{i=1}^{s} \sum_{j=1}^{t/\Delta\tau} k_i H\prime\left(\frac{t-t_i-\tau_j}{C_s}\right) u(\tau_j)\Delta\tau. \tag{8}$$

In this case, the nonparametric model of nonlinear object is the following:

$$\hat{x}(t) = f(\hat{w}(t), \alpha), \tag{9}$$

$$\hat{w}(t) = \int_0^t \hat{k}\prime(t-\tau)u(\tau)d\tau, \tag{10}$$

Or, if we omit the $w(t)$, we obtain the formula[5]:

$$\hat{x}(t) = f\left(\int_0^t \hat{h}(t-\tau)u(\tau)d\tau, \alpha\right), \tag{11}$$

where $\hat{x}(t)$ -– nonlinear function estimation, $\hat{k}(t)$ -– estimation of the step response function of the system linear element, $x(t)$ — system output signal; $u(t)$ — input signal of the system; α -– estimation of the nonlinear function parameters.

Thus, we get the algorithm for modeling of Wiener type nonlinear dynamical systems.

3 Identification of nonlinear system with a quad

Let's consider a system that is represented as the Wiener model. The nonlinear part of the system is a quad. The object output is calculated as follows: $x(t) = aw(t)^2, a - const$. If the value of input action $u(t) = 1$, then the output of nonlinear system $x^1(t) = ak(t)^2$.

That is, the step response estimation of a linear element can be represented by the output of the process as follows:

$$\hat{k}(t) = \sqrt{x^1(t)/a}) \tag{12}$$

For an arbitrary input action the output of linear part of the system is described by the Duhamel integral. Considering (12) the output of the linear element is:

$$w(t) = \frac{1}{sC_s}\sum_{i=1}^{s}\sum_{j=1}^{t/\Delta\tau}\sqrt{x(t)/a}H\prime\left(\frac{t-t_i-\tau_j}{C_s}\right)u(\tau_j)\Delta\tau. \tag{13}$$

Then the model of the nonlinear dynamic object of Wiener type is:

$$\hat{x}(t) = \left(\frac{1}{sC_s}\sum_{i=1}^{s}\sum_{j=1}^{t/\Delta\tau}\sqrt{x}^{1}(t)H\prime\left(\frac{t-t_i-\tau_j}{C_s}\right)u(\tau_j)\Delta\tau)\right)^2, \tag{14}$$

where x_i^1 – the reaction of a nonlinear system (if $u(t) = 1$), $u(t)$ – a test input action.

Example. Consider a nonlinear dynamical system consisting of a quad (parameter $a = 0.7$) and the differential equation (simulating object):

$$2x\prime\prime(t) + 0.3x\prime(t) + 1.5x(t) = u(t).$$

The noise in measure channels is generated as follows:

$$x_i^{sh} = x_i + c\xi,$$

where x_i – object output (without interference), ξ – normally distributed additive noise with zero mean and unit variance. Constant value c determines the noise intensity, it calculated according to a given value p, determining the signal-noise ratio (if $p = 10$ noise 10, if $p = 1$, noise 100):

$$p = \frac{\sqrt{\frac{1}{s}\sum_{i=1}^{s}x_i^2}}{c},$$

The quality of the model is estimated using the average relative error of simulation.

$$W = \frac{1}{s}\sum_{i=1}^{s}\frac{|x_i-\hat{x}_i|}{|x_{max}-x_{min}|},$$

where x_i – object output, $\hat{x}_i$ – output of model.

The object modelling result is presented in the figure:

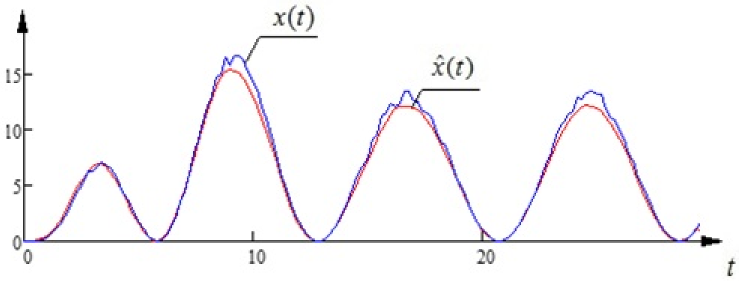


Figure 3: The model of system: $\hat{x}(t)$ – a model of a nonlinear system, $x(t)$ – the system output, sample size s = 250, sampling interval $h = 0.2$, noise 5%, input action $u(t) = 2\cos(0.4t)$

4 Identification of a system with link saturation nonlinearity

We consider the system with a nonlinear element is described by the function of link saturation type.

$$f(w) = \begin{cases} w(t), & \text{if } w(t) < b; \\ a, & \text{if } w(t) > b; \\ -a, & \text{if } w(t) < -b. \end{cases} \tag{15}$$

where a, b – unknown parameters. The function graph is shown in the figure: In this case if $w(t) < b$, then the object output is equal to the output of its linear dynamic part. Otherwise, the output of the object is a constant, which can be determined

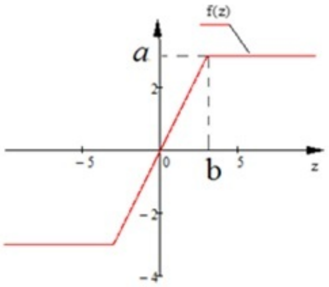


Figure 4: The graph of a link saturation, z – an arbitrary argument

experimentally by several static experiments. We get the following algorithm to construct the model:

1. to carry out some experiments (m) under the following conditions: input actions $u_j = c_j$, $c_j = const$. The result is a sample $\{u_j, x_j\}$, $j = 1, m.$, where $u_j = c_j$ – constant input action, $x_j = x_j^y$ – stand value of the output

2. to find the distance between two consecutive measurements: $\delta x = |x_j - x_{j-1}|/\delta u$, where δu – sampling interval.

3. estimation of parameters:

$\hat{b} = x_j$, if $\delta_j < \varepsilon, \varepsilon > 0$.

$\hat{a} = M(x_j)$, $j = j_0, m$ if $x_{j_0} = \hat{b}$, where $M(x)$ – estimation of mathematical expectation.

4. to get a step response (apply to the object input a step function, the amplitude of that is less than b) and then construct a linear part model in the form of the Duhamel integral.

5. to build a model of the object, the output of which is calculated as the value of the function describing the nonlinear element, whose argument is the output of the linear model of the object.

Example. Consider a nonlinear dynamical system consisting of a link saturation (with parameters $b = 1.34, a = 1.5$) and the differential equation (simulating object):

$$7.4x\prime\prime(t) + 2.5x\prime(t) + 2.43x(t) = u(t)$$

The object modelling results with different input actions are presented in the figure:

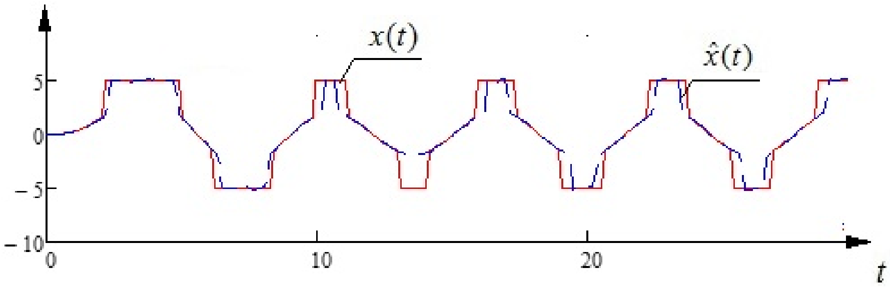


Figure 5: The model of system: $\hat{x}(t)$ – a model of a nonlinear system, $x(t)$ – the system output, sample size s = 250, sampling interval $h = 0.15$, noise 5%, input action $u(t) = 5\sin(t)$, the relative average error of simulation 5.2%

Making the analysis the model of nonlinear dynamic objects with a link saturation and quad nonlinearity, we can conclude, that the nonparametric model sufficiently describes the nonlinear dynamical systems Wiener type.

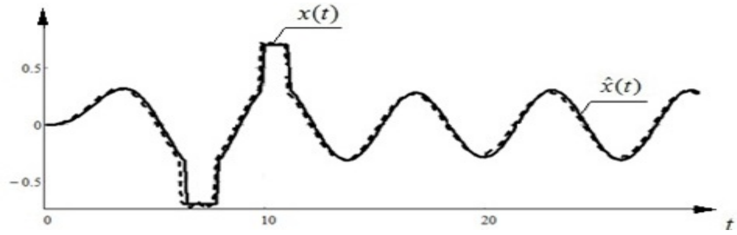


Figure 6: The model of system: $\hat{x}(t)$ – a model of a nonlinear system, $x(t)$ – the system output,input action $u(t) = 0.35\sin(2t)$, the relative average error of simulation 3.4%

Conclusions

In this paper we consider the problem of nonlinear dynamical systems identification. The investigated objects are presented as a consequent combination of linear dynamic and nonlinear static blocks (Wiener model). In this case a structure and parameters of linear dynamic block of such system is unknown, but the type of the nonlinear element is assumed to be known with the set of parameters.

The problem of nonlinear system identification can be divided into two tasks. At first the nonparametric model of linear dynamical element is considered. Presented methods of the nonlinear system identification are based on the combining the models of linear dynamic and nonlinear static processes in the overall model of the system. These techniques do not require the presence of full a priori information about the structure of the object.

The practical part presents the results of numerical experiments, in that were designed the models nonlinear dynamical processes of Wiener type in the cases of quad and link saturation nonlinearity.

References

[1] Eykoff P. (1975). *The basis of control systems identification.* Springer-Verlag.

[2] Chaika S.N. (1989). *Identification of dynamic systems with partially parameterized model structure.*Izd Gorky State. University Press, Gorky.

[3] Medvedev A.V., (1979). - *Nonparametric algorithms of nonlinear dynamical systems identification. / Sat. Stochastic control system.*. Nauka, Novosibirsk.

[4] Medvedev A.V., (1983). - *Nonparametric system adaptation.* Nauka, Novosibirsk.

[5] Popkov U.S. (1976). *Identification and optimization of nonlinear stochastic systems,* - Pub. M: Energiya, Moscow.

Some Remarks on H-models Identification of Noninertial Processes

Korneeva A., Mihov E.
Siberian Federal University, Krasnoyarsk, Russia
e-mail: anna.korneeva.90@mail.ru, edmihov@mail.ru

Abstract

A modeling of discrete-continuous processes with “tubular” structure in the space of “input-output” variables. Modeling of this process differs significantly from the class of conventional parametric models representing the same surface area. In the construction of students parametric models “tubular” processes require the use of appropriate non-parametric indicators. Some special examples of modeling “tubular” processes, from which it follows that the processes take place in the spaces of fractional dimension. Cited the case of functions of several variables and analyzed the situation when in the course of time, these variables can “disappear” and “occur” again. It is shown that the calculation of the fractional dimension space can be done in different ways.

Keywords: aprioristic information, identification, nonparametric model, nonparametric algorithms, H-models, space of fractional dimension.

Introduction

Many stochastic objects’s identification is often reduced to static systems with delay identification. It is caused some output variables of object are controlled through much big intervals of time, than entrance and significantly exceed an object time constant. For example, a number of variables is measured in the electric way (in this case discretization of control Δt can be rather small), and other variables are controlled as a result of the chemical analysis or physicomechanical tests (in this case discretization of control ΔT – is great, i.e. $\Delta T >> \Delta t$). The most general scheme of the research discrete continuous process can be submitted in the following figure [1]:

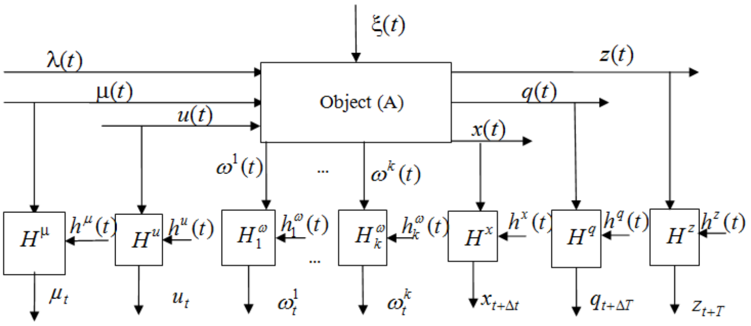


Figure 1: The general scheme of the studied process

In Figure 1, the notation: A - the object with unknown structure, $x(t)$, $z(t)$, $q(t)$ – the output process variables, $u(t)$ - manipulated, $\mu(t)$ – uncontrolled, but the measured entrance process variable, $\lambda(t)$ - uncontrolled and measured entrance process variable, $\xi(t)$ - a random effect, $\omega^i(t)$ – the process variables controlled the length of the object, (t) - continuous time H^μ, H^x, H^u, H^q, H^ω – communication channels corresponding to different variables, including the controls, instruments for measuring the observed variables μ_t, u_t, x_t, $z_{\Delta T}$, q_T, ω_t – means measuring $\mu(t)$, $u(t)$, $x(t)$, $z(t)$, $q(t)$, $\omega(t)$ in a discrete time, $h^\mu(t)$, $h^u(t)$, $h^x(t)$, $h^z(t)$, $h^q(t)$, $h^\omega(t)$ – marked with top – random noise measurement relevant process variables.

We note a significant difference between the output variables $z(t)$, $q(t)$ and $x(t)$ presented in the Figure 1. The output variable $x(t)$, as well as input, monitored at intervals controlled by substantially larger intervals Δt, $q(t)$ - through T, $T >> \Delta T >> \Delta t$. From a practical point of view, for the process is often the most important, is the control variables $z(t)$. In this case, the output variables depend on the input and (more information), as follows:

$$x_t = A(u(t), \mu(t), \omega(t), \lambda(t), \xi(t), t) \tag{1}$$

Modeling similar processes, considering various sampling of control of measurements $x(t)$, $q(t)$ and $z(t)$ forecasting $q(t)$ and $z(t)$ it is natural to use all set of the variables influencing the forecast $x(t), q(t), z(t)$

$$\hat{x}_t = \hat{A}(u(t), \mu(t), \omega(t), t) \tag{2}$$

$$\hat{q}_t = \hat{A}(u(t), \mu(t), \omega(t), \hat{x}(t), t) \tag{3}$$

$$\hat{z}_t = \hat{A}(u(t), \mu(t), \omega(t), \hat{x}(t), \hat{q}(t), t) \tag{4}$$

Considering great values between ΔT and T, considerably exceeding object time constants, when modeling it is necessary to consider that processes belong to the class static with delay that considerably raises their role and value in problems of identification and management of stochastic systems.

For a further statement, without violation of a community, we “will curtail” all entrance and output variables in corresponding a vector. Then the studied object can be presented static with delay. Expediently on the respective canal to present such process in the form:

$$x(t) = f(u(t-\tau), \xi(t)) \tag{5}$$

here is x(t) is the output variable of object, $u(t-\tau)$ is the cumulative entrance variable, τ is the delay, $\xi(t)$ is the casual indignation operating on object, t is the continuous time.

1 H-models

Let $u = (u_1, ..., u_k)$, $x \in \Omega(x) \subset R^1$. Everyone vector component $u_i \subset [a_i; b_i]$, $i = \overline{i, k}$, and $x \subset [c; d]$. At research of real processes of value of coefficients $\{a_i, b_i, c, d\}$,

$i = \overline{i,k}$, are always known. In technological processes of value of these coefficients are regulated by production schedules (card). Further, without violation of a community, we will accept these intervals single [1], then $\Omega(u)$ – a single hyper cube, i.e. $u \subset [0;1]$, $\Omega_{k+1} = [0;1]$, $(u,x) \subset \Omega_{k+1}(u,x)$. The adaptive model in this case will look as follows:

$$\hat{x}_s(u) = \hat{f}(u, \alpha_s) \tag{6}$$

The "weakest" place is the choice of parametrical structure of model here. If at the first stage rather gross blunder is made, as a result, the received model will hardly be satisfactory. This problem was rather in detail discussed in [2, 3]. We will pay attention that models of a class (6) represent hypersurfaces in space of "entrance-output" variables of object, i.e. $(u,x) \subset \Omega(u,x) \in R^{k+1}$.

If the studied process has a "tubular" structure [2], the model (6) needs to be corrected as follows:

$$\hat{x}_s(u) = I_s(u)\hat{f}(u, \alpha_s) \tag{7}$$

or:

$$\hat{x}_s(u) = I_s(u)\sum_{j=1}^{N} \alpha_{sj}\phi_j(u) \tag{8}$$

here is $\phi_j(u)$ – system linearly-independent functions, the indicator $I_s(u)$ is described by:

$$I_s(u) = \begin{cases} 1, if u \subset \Omega_s^H(u), \\ 0, if u \notin \Omega_s^H(u). \end{cases} \tag{9}$$

We will notice only that, generally speaking, the area $\Omega^H(u)$ to us isn't known, and only selection $\{x_i, u_i, i = \overline{i,s}\}$ is known. If the indicator is equal to zero, the assessment $\hat{x}_s(u)$ can't be calculated, i.e. at such values a vector component $u \in \Omega(u)$ process can't proceed. If the indicator $I_s(u)$ at any value $u \in \Omega(u)$ is equal to unit, the model (7) coincides with (6). As an assessment of the indicator $I_s(u)$ it is possible to accept the following approach:

$$I_s(u) = sgn\sum_{i=1}^{s} \Phi(c_s^{-1}(x_s(u) - x_i))\prod_{j=1}^{k} \Phi(c_s^{-1}(u^j - u_i^j)), \tag{10}$$

here is:

$$x_s(u) = \sum_{i=1}^{s} x_i \prod_{j=1}^{k} \Phi(c_s^{-1}(u^j - u_i^j)/\sum_{i=1}^{s} x_i \prod_{j=1}^{k} \Phi(c_s^{-1}(u^j - u_i^j), \tag{11}$$

and the parameter of blurring c_s and bell-shaped function $\Phi(.)$ meet some conditions [2]. Thus, at known value $u = u`\in \Omega(u)$ at first the assessment $x_s(u = u`)$ on a formula (11) is under construction, then the indicator $I_s(u)$ is calculated, and only at the following stage models (7) or (8) if the indicator was equal to unit are used. If the indicator is equal to zero, it means that though, $u` \in \Omega(u)$ but, $u` \notin \Omega^H(u)$

i.e. components of a vector $u = u` = (u`_1, ..., u`_k)$ are defined not truly, otherwise, really proceeding "tubular" process there doesn't correspond to set of preset values the vector component $u = u`$. The reasons of it can consist that components of a vector $u = u` = (u`_1, ..., u`_k)$ are chosen incorrectly, or are measured with a considerable error like "emission". Of course, it is fair only provided that we have representative selection $\{x_i, u_i, i = \overline{i, s}\}$. It is necessary to notice that use of traditional models (6) type will allow to receive an assessment $\hat{x}(u = u`)$ which, naturally, will be far from reality.

2 Dimension of the process

We will give the following example concerning identification of inertialess system. We will consider the following simple special case. Let the object be described by the equation:

$$x(u) = f(u_1, u_2, u_3), \tag{12}$$

here is $u = (u_1, u_2, u_3) \in R^3$ the three-dimensional vector is an entrance variable, and $x \in R^1$ – an output variable. The traditional way of creation of model of the process described (12) consists in definition of a class of parametrical dependences $\hat{x}(u) = \hat{f}(u_1, u_2, u_3, \alpha)$ and the subsequent assessment of parameters α one way or another on selection of supervision $(u_i, x_i), i = \overline{1, s}$, where s – selection volume. We will analyse this example from the different points of view. Let components of a vector of entrance variables $u = u_1, u_2, u_3$ stochastic be not connected in any way, i.e. are independent. In this case it is natural to use the standard traditional practice described above. Now we will assume that objectively components of a vector of entrance variables are functionally connected, for example:

$$u_2 = \phi_1(u_1), u_3 = \phi(u_2) = \phi_2(\phi_2(u_1)). \tag{13}$$

Naturally, the researcher doesn't know about existence of dependences (13). Otherwise it would be possible to make substitution (13) in (12) and to receive the following dependence x already from one variable u1 of a look

$$x(u) = f(u_1, \phi_1(u_1), \phi_2(\phi_2(u_1))). \tag{14}$$

Thus, dependence (12) in the conditions given above can be reduced to one-dimensional dependence x from u1. In case dependence u3 from u2 objectively is absent, (12) is easily given to a look

$$x(u) = f(u_1, \phi_1(u_1), u_3). \tag{15}$$

i.e. to two-dimensional dependence x from u_1, u_3. From here it is possible to conclude that with functional dependence between vector u components we receive dependence x from u, in this case, one - two-three-dimensional. We will emphasize once again that between components of a vector of entrance variables the researcher doesn't know of existence of functional dependences. Simply we analysed a case: "If ...". And now we will analyse the most interesting case directly related to H-processes [1]. Let u_3, u_2 and, though an unknown image, but stochastic are connected [2]. We will emphasize – stochastic, but it isn't functional. We will return once again to the analysis of that occurred. First, if components of a vector u are independent, the studied process is described by function of three variables. If two components of a vector of entrance variables u are connected by functional dependence, process is described by function of two variables. At last, if two variables are connected stochastic, process is described by function more than two variables, but less than three?! It is possible to consider that we come to dependence on fractional number of variables and, therefore, to space of fractional dimension. For example, B. Mondelbrot in [4] notices: "The blood system of the person – pulsing live – has dimension 2.7". Fractional dimension of spaces, apparently, was for the first time noted in Hausdorff and Bezikovich's works.

We will consider the following situation. From simplicity of reasons, let the process interesting us is described (19). In case of stochastic dependence between the u_2 variables u_1 , $u_3(u_1$ on the available training selections it is possible to calculate a square error of the forecast $u_{2s}(u_1), u_{3s}(u_1)$. Here $u_{2s}(u_1), u_{3s}(u_1)$ are nonparametric estimates [1].

$$\delta_{21} = \sum_{i=1}^{s} (u_2 - u_{2s}(u_1))^2 / \delta_{u_2}^2 \tag{16}$$

$$\delta_{31} = \sum_{i=1}^{s} (u_3 - u_{3s}(u_1))^2 / \delta_{u_3}^2 \tag{17}$$

"Force" of stochastic communication λ between two any variables can be calculated, for example, on a formula:

$$\lambda = 1 - \delta \tag{18}$$

From here it is visible that the strongest stochastic communication (functional) is equal 1, lack of communication takes place at ? = 0, and at stochastic dependence between entrance variables $0 < \lambda < 1$.

If to keep mathematical "shape" of interpretation of function of many variables as a point of multidimensional space, we come to existence of space of fractional dimension F^λ. Calculation of dimension F^λ. can be carried out, for example, so:

$$dim F^\lambda = (n+1) - \sum_{i=1}^{n-1} \lambda_{i,i+1} \tag{19}$$

here is n – dimension of a vector of u, and $\lambda_{i,i+1}$ means "force" of stochastic communication between u_i and u_{i+1}. Also other schemes of calculation of dimension of space can be offered. For example,

$$dim F_1^\lambda = (n+1) - \sum_{i=1}^{n-1} \lambda_{1,i+1} \tag{20}$$

here is $\lambda_{1,i+1}$ – dependence of all u vector component from one components u1. In rather attentive analysis of decomposition of functions in ranks pertinently to remember V. I. Arnold's phrase from the book "Theory of Accidents" [5]: "Calculation in these applied researches were usually carried out without the general theory due to the correct rejection of one members of a number of Taylor, and leaving of others, the most important. From the physicists who were especially systematically applying the theory of accidents before its emergence once especially L. D. Landau allocates. In his hands to reject art "insignificant" members of a number of Taylor, keeping members, smaller in size "the physically important", gave a lot of the accidents of results included in the theory".

3 Computing experiments

Let process be described by function $x = f(u_1, u_2)$ and is under the influence of a hindrance $\xi(t)$. We will accept the training selection equal 500, entrance variables – are independent (Figure 2), also we will show dependence of dimension of space of F^λ on s (Figure 3).

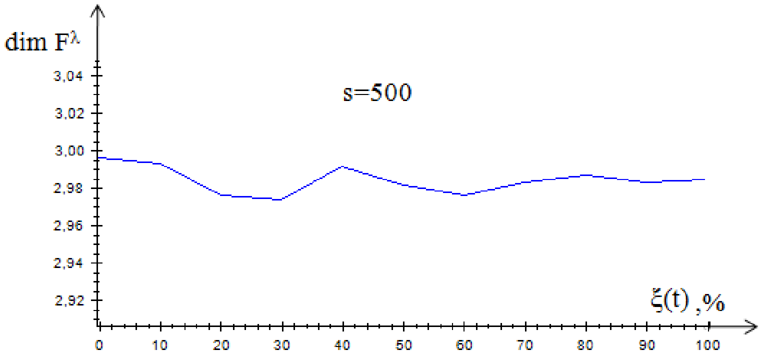


Figure 2: Dependence of dimension of space of F^λ depending on the level of hindrances

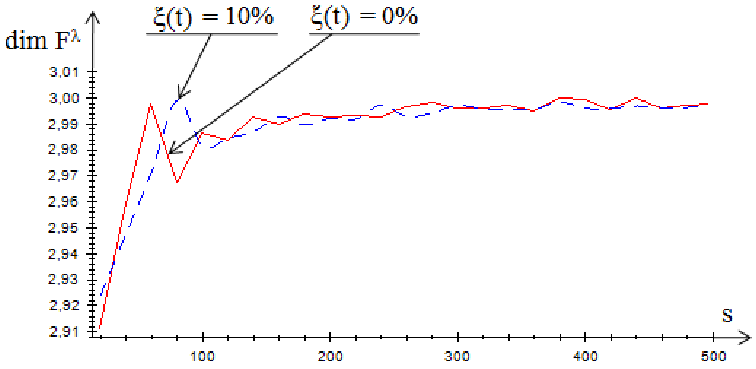


Figure 3: Dependence of dimension of space of F^λ depending on selection volume

In figure 2 it is clearly that at independent entrance variables, dimension of process is close to 3. Figure 3 illustrates that at small selection, dimension of F^λ decreases, but at increase in selection dimension of space of F^λ is close to 3. We will consider the process having “tubular” structure, that is H-process.

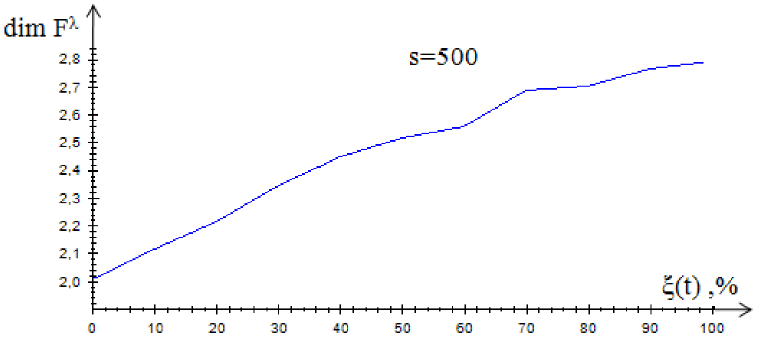


Figure 4: Dependence of dimension of space of F^λ depending on the level of hindrances

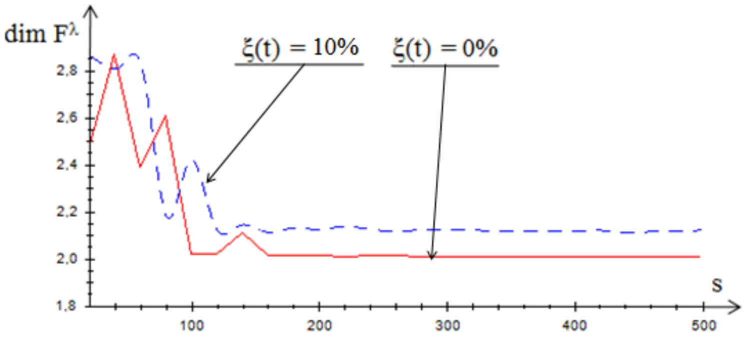


Figure 5: Dependence of dimension of space of F^λ depending on selection volume.

It is clearly that at H-process, in this case, dimension of space is close to two (Figure 4). We will explain, the reason of this phenomenon.

Conclusion

The analysis of the situations arising when modeling processes of “tubular” structure which takes place always if components of a vector of entrance variables of process stochastic are dependent was carried out. In this case traditionally used models of static systems with delay are inapplicable or can lead to considerable mistakes. The most interesting is that fact that we come to need of consideration of space of fractional dimension. Certainly, the disappearance fact, and emergence of influence of some entrance variables during various periods of time for values of output variables of process that is closely connected not so much with space of fractional dimension, how many with space of the changing dimension is interesting.

Acknowledgements

The authors wish to express their gratitude to the supervisor, Prof. Dr. Medvedev A.V. who was abundantly helpful and offered invaluable assistance, support and guidance.

References

[1] Medvedev A.V. (2014). Some remarks to N-models of inertialess processes with delay.*Vestnik SibGAU*. Vol. **54**, pp. 42-34.

[2] Medvedev A.V. (1995). The analysis of data in a problem of identification. Computer analysis of data of modeling. *BGU*. Vol. **2**, pp. 201-206.

[3] Medvedev A.V. (2012). H-models for inertialess systems with delay.*Vestnik SibGAU*. Vol. **45**, pp. 84-89.

[4] Mondelbrot V. (2010). *Fractal geometry of the nature.* Research Center regular and chaotic dynamics, Moscow.

[5] Arnold V.I. (1990). *Theory of accidents.* Science, Moscow.

About the Method of Observations Supplements the Source Data

Ekaterina A. Chzhan and Natalia A. Sergeeva
Siberian Federal University, Krasnoyarsk, Russia Federation
e-mail: ekach@list.ru

Abstract

The problem of identification of noninertial stochastic processes with delay is discussed. Quality of solving the problem of identification depends on the quality of input data. This work is dedicated to the elimination of such drawbacks in the original sample observations as sparse and the area to the lack of observations. The proposed algorithms improve the quality of the model several times.

Keywords: sample, nonparametric identification, data analysis.

Introduction

The problem of input sample quality is considered. Usually there is sufficient a priori information about the object being studied so it is necessary to apply the methods of identification in the “narrow” sense. These methods include nonparametric estimation of regression function from observations.

The quality of solving the problem of identification depends on the quality of input data. It is advisable to conduct a preliminary analysis of data to identify and address all the deficiencies in the sample. Under the preliminary analysis of the data is taken to mean filling gaps in observations and eliminating emissions. However, the sample may have other flaws, that will be discussed below, which adversely affect the accuracy of estimation, and, in some cases, lead to the fact that the resulting model will be inadequate to the investigated process. If the point of the original sample in the field of process located patchy, there are low-pressure range and lack of observations, in the areas of reconstruction accuracy is low. Due to the properties of nonparametric estimation, which belongs to the class of local approximations, projections can not be given at the lack of observations subdomain. To resolve all these shortcomings we propose an algorithm to obtain a working sample by generating new points in regions where the density is low in comparison with other areas. After generating new working sample, the quality of estimation is significantly improved.

1 Posing of the problem

The general scheme of the researched process is shown in Fig.1.

The table of symbols are accepted in fig. 1: A is an unknown object operator, $x(t) \in \Omega(x) \subset R^1$ is an output variable of the process, $u(t) = (u_1(t), u_2(t), ..., u_m(t)) \in \Omega(u) \subset R^m$ is a control action, $\xi(t)$ is a vector random action, (t) is continuous time, G^u, G^x are channels of connection corresponding to different variables and including control tools, $g^u(t), g^x(t)$ are random noises of measurements corresponding to variables of the process with zero means and limited variance, $u_i, x_it, i = 1, 2, ..., m$ are obtained by measurements of process variables at discrete time.

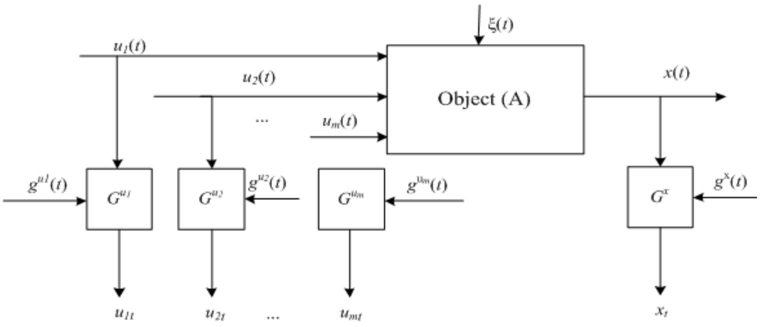


Figure 1: The general scheme of multidimensional object

So, we have a sample of observations of input and output variables process $\{x_i, u_i, i = \overline{1, s}\}$, where s means sample size.

Usually model is based on measurements of the input and output process variables where some algorithms are used to estimate the object output. In most cases, the researcher has a small amount of a priori information, the mathematical description of the object is unknown so it is advisable to use methods of nonparametric statistics.

2 Nonparametric identification

The nonparametric estimation of regression functions on observations Nadaraya-Watson refers to methods of nonparametric identification [1].

Let us assume that observations $\{x_i, u_i, i = \overline{1, s}\}$ of random values x, u distributed with the unknown density of probability $p(x, u), p(u) > 0 \forall u \in \Omega(u)$ Nonparametric estimates [1] are used for the backing up $\tilde{x} = M\{x|u\}$

$$x_s(u) = \sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi(c_s^{-1}(u^j - u_i^j)) / \sum_{i=1}^{s} \prod_{j=1}^{m} \Phi(c_s^{-1}(u^j - u_i^j)), \tag{1}$$

where the kernel function $\Phi(c_s^{-1}(u^j - u_i^j))$, $i = \overline{1, s}$, $j = \overline{1, m}$ and the smoothing factor c_s^{-1} have convergence properties [1]. In this case the triangular kernel was used as the bell-shaped function $\Phi(c_s^{-1}(u^j - u_i^j))$, $i = \overline{1, s}$, $j = \overline{1, m}$:

$$\Phi(c_s^{-1}(u^j - u_i^j)) = \begin{cases} 1 - |c_s^{-1}(u^j - u_i^j)|, if c_s^{-1}(u^j - u_i^j) \leq 1, \\ 0, if c_s^{-1}(u^j - u_i^j) > 1. \end{cases} \tag{2}$$

The smoothing parameter is defined as a solution of minimization of a square criterion which shows the equivalence between object and model outputs compliance and it is based on the method of "sliding examination", i.e. the i-observation is not considered in the model [2]:

$$R(c_s) = \sum_{k=1}^{s}(x_k - x_s(u_k, c_s))^2 = \min_{c_s}, k \neq j. \tag{3}$$

If every component of a vector c_s corresponds to every component of a vector u then in many real problems it is possible to accept that c_s is a scalar if components of a vector u are transformed into the same interval, for example, with centering and rationing operations.

The quality of the estimation mainly depends on the quality of the source data: a sample could have a number of drawbacks, which ultimately lead to poor estimate. Consider the case where the sample has such shortcomings as the "sparsity and "gaps". As an example, for reasons of simplicity of illustration, consider a three-dimensional object, a field correlation input variables is shown in Fig. 2. In the present sample sparsity - areas with a small number of points and gaps - areas where observations are missing. In these areas, the estimate will be of low quality or not be able to get.

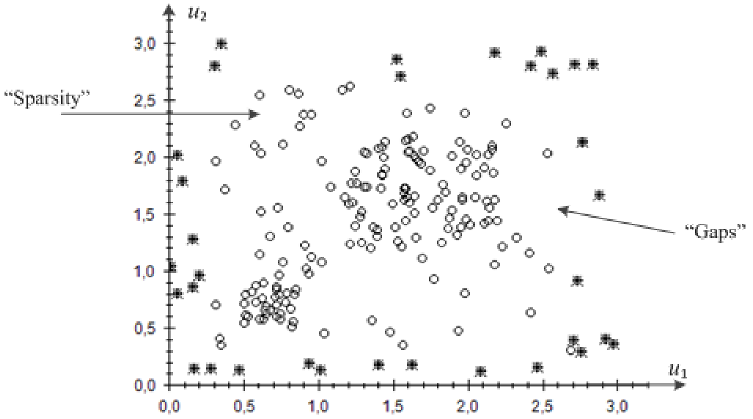


Figure 2: Field correlation of input variables for source sample

The algorithm supplements the source sample by generating new observations is suggested below.

3 The method of observations supplements the source data

An algorithm for generating the working sample on the basis of the original obtained by measuring the input and output variables of the process is considered. The proposed algorithm is based on the generation of new observations in the tolerance range of variables. These observations are used only for reconstruction assessment model output in real sampling points of the original or new points that need to receive the value of the forecast.

The first stage is determined by observing the relative density of the number of neighboring elements. Then, in the sub-regions, where the density is low are generated by new observations. The value of the output variable x is calculated using a non-parametric estimation of (1). This increases the size of the original sample due to the generated elements.

4 Computational experiment

Let's consider the results of the simulation of non-linear object. Let the object described by the following equation:

$$x = u_1^2 - 2\sin(u_2) + \xi, \tag{4}$$

where inputs $u_1, u_2 \in [0;3]$, ξ - normally distributed disturbance:

$$\xi = k\sigma, \tag{5}$$

where k - noise level, σ - a random variable distributed normally with zero expectation in the range $[-1;1]$.

Mathematical description (4) is given only for the generation the source sample since no opportunity to operate with real data. After generating the source sample, we assume that the dependence of (4) we do not know.

Let generate a sample size of 100 elements ($s = 100$). In the sample "sparsity" is located, where is a small number of sample elements, and "gaps" in the data where is no elements. Thus, the density of sample points is non-uniform. Field correlation of input variables is shown in Figure 2.

The object output is recovered with the method of identification in the "narrow" sense, using a non-parametric estimation (1). If there is a situation of uncertainty, i.e. no point does not fall under the bell (2), the value assigned to the forecast of 6 - as the maximum possible value of the output variable x. The relative error of approximation shows the quality of recovery:

$$W = \sqrt{\frac{\frac{1}{s}\sum_{i=1}^{s}(x_s i - x_i)^2}{\frac{1}{s-1}\sum_{i=1}^{s}(x_i - \hat{m}_x)^2}}, \tag{6}$$

As a result of the above algorithm a working sample is generated , which includes the elements of the source sample and generated artificially. The size of new sample is 453 points. The relative (6) error for the source sample is $0,838$. Then the estimation (1) is recovered for the source sample wy the working sample, the error is $0,237$. The quality of recovery increases 4 times.

Field correlation of input variables for the working sample is shown in Figure 3. As can be seen, the density of the points are now uniform now.

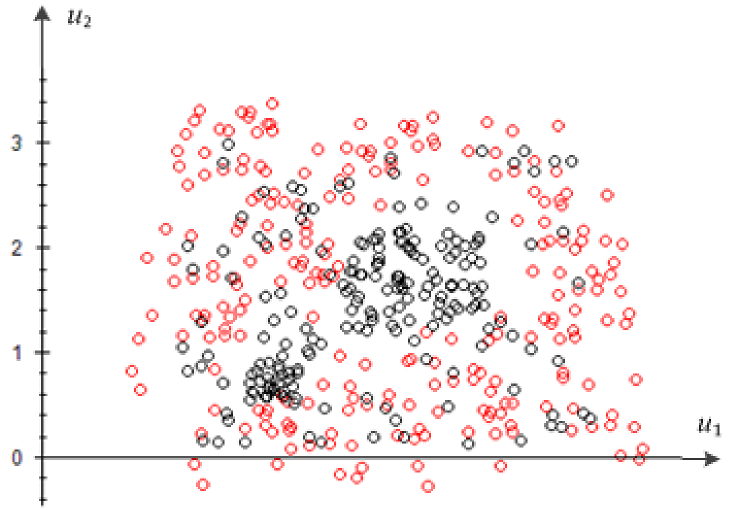


Figure 3: Field correlation of input variables for working sample

Conclusions

In this work some of the shortcomings in the source sample of observations are considered (areas of gaps, sparsity), which have a negative impact on the quality of nonparametric estimation recovery. The algorithm that allows to identify such areas and generate them additional elements of the sample is studied. These artificially produced observations do not carry new information about the object, but allows to improve the quality of the simulation several times.

Acknowledgements

The authors wish to express their gratitude to the supervisor Prof. Dr. Medvedev A.V. who was abundantly helpful and offered invaluable assistance, support and guidance.

References

[1] Nadaraya E.A (1983) *Nonparametric estimation of the probability density and the regression curve.*Izdatelstvo Tbiliskogo yniversiteta, Tbilisi.

[2] Korneeva A.A., Sergeeva N.A., Chzhan E.A. About nonparametric data analysis in the identification problem // Vestnik TGU. Control, Computer Science and Informatics. Vol.1 (22). 2013. P.86-96.

Bayesian semiparametric Mixtures in Quantile Regression

Milovan Krnjajić [1] and Athanasios Kottas[2]
[1] *National University of Ireland, Galway, IRELAND*
[2] *University of California at Santa Cruz, USA*
e-mail: milovan.krnjajic@nuigalway.ie, thanos@soe.ucsc.edu

Abstract

Mixture models are typically used to extend flexibility of the shape of standard parametric densities and to allow for increased variability. Bayesian nonparametric (BNP) mixtures employ stochastic processes to randomly generate mixing distributions, resulting in flexibility that goes beyond what is achievable by parametric models. In particular, Dirichlet process mixtures have been successfully used in a variety of settings to model multi-modality and non-standard behavior of density tails. We highlight some details and stages of developing DP mixture models for error distributions in a quantile regression formulation, including a model based on dependent DPM-s that enables error distribution to vary non-parametrically with covariates. Model performance is illustrated on simulated and real datasets.

Keywords: quantile regression, Dirichlet process mixtures, Bayesian nonparametrics.

Introduction

Common regression models focus on the conditional mean of response distribution to summarize relationship between variables. However, an analysis based on these models may be inadeqate in case of data sets with complex interactions of factors manifested in heterogeneous variability of response for different ranges of covariates. Such data appear in a variety of applications, for example, in econometrics and ecology, where distributions of response may be highly skewed and many data points appear as outliers. Quantile regression (QR) is a method suitable for the analysis of such data sets as it estimates relations between covariates and any portion (quantile) of the response distribution, see for example Koenker (2005).

In the standard additive regression formulation, the p-th quantile of the response distribution for observations y_i, and covariate vectors $\boldsymbol{x}_i$, $i = 1, ..., n$, can be written as

$$y_i = h(\boldsymbol{x}_i) + \epsilon_i, \tag{1}$$

where ϵ_i-s are assumed independent from an error distribution with p-th quantile equal to zero, i.e., $\int_{-\infty}^{0} f_p(\epsilon)\mathrm{d}\epsilon = p$, with $f_p(\cdot)$ denoting the error density. The $h(\boldsymbol{x})$ is typically expressed as $\boldsymbol{x}^T\boldsymbol{\beta}$, with error density $f_p(\cdot)$ unspecified except for the requirement that $\int_{-\infty}^{0} f_p(\epsilon)\mathrm{d}\epsilon = p$. To obtain point estimation of $\boldsymbol{\beta}$ classical approach uses optimization of some *loss* function. For example, the point estimates for $\boldsymbol{\beta}$ minimize $\sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$, where $\rho_p(u) = up - u1_{(-\infty,0)}(u)$. For $p = 0.5$, this estimation reduces to the median regression case, which is solved using least absolute deviations (LAD) methods of optimization. A general limitation of optimization based methods is that inference is based on asymptotic assumptions or resampling methods and thus requires large data samples. Interestingly, a version of the LAD regression method was first solved by Ruggiero Boscovich in 1757, a half of a century before Legendre and Gauss developed the least squares method, see Stigler (1984).

We focus on building Bayesian semi-parametric models that belong to Bayesian non-parametrics – a rapidly growing area of Bayesian statistics, that provides a fully probabilistic framework for inference, and develops models that can grow in complexity with sample size. Applications of BNP methods abound in many diverse disciplines such as natural language processing, computer vision, computational biology, medicine, and signal processing. For information on BNP theory, methods and applications, see the book by Hjort et al. (2010) or a recent review paper by Müller and Mitra (2013). Here we give a condensed outline of some aspects of motivation and reasoning in developing a series of Bayesian semi-parametric mixture models for error distribution in quantile regression. We also present a few results (not published previously) of an analysis of a real data set using a model based on dependent Dirichlet process. A full account of models outlined here, their performance and accompanying MCMC algorithms, can be found in Kottas and Krnjajić (2005) and Kottas and Krnjajić (2008).

Nonparametric scale mixture of asymmetric Laplace densities

A parametric model of choice in quantile regression analysis is the family of asymmetric Laplace distributions (ALD) with densities

$$k_p^{AL}(\epsilon;\sigma) = \frac{p(1-p)}{\sigma}\exp\left\{-\frac{|\epsilon| + (2p-1)\epsilon}{2\sigma}\right\}, \tag{2}$$

where $0 < p < 1$, $\sigma > 0$ is a scale parameter and $\int_{-\infty}^{0} k_p^{AL}(\epsilon;\sigma)\mathrm{d}\epsilon = p$.

It is important to note that the parameter, p, determines both skewness and p-th quantile for the density in (2) what limits its flexibility in modeling skewness and tail behavior. In particular, $k_p^{AL}(\cdot;\sigma)$ is right skewed for $p < 0.5$, left skewed for $p > 0.5$ and symmetric for $p = 0.5$, i.e., for the median regression case. This is a very restrictive feature as median regression is typically employed to capture skewness in the response distribution. Moreover, it is impossible for a right skewed ALD ($p < 0.5$) to accurately model left skewed errors for quantiles lower than 0.5, and likewise, a left skewed ALD ($p > 0.5$) can't capture right skewed densities at quantiles higher than 0.5. We refer to (1) with error density $f_p(\cdot) = k_p^{AL}(\cdot;\sigma)$ as the model $\mathcal{M}_0$.

To develop a model with a more flexible tail behavior, we first consider a nonparametric mixture of ALD-s with a Dirichlet process (DP) prior for the mixing distribution (Ferguson, 1974). Specifically, denoting by $\mathrm{DP}(\alpha G_0)$ the DP with precision parameter α and base distribution G_0, we define

$$f_p^{ALD}(\epsilon;G) = \int k_p^{AL}(\epsilon;\sigma)\mathrm{d}G(\sigma), \qquad G \sim \mathrm{DP}(\alpha G_0). \tag{3}$$

Note that mixing in this fashion preserves the quantiles, i.e., $\int_{-\infty}^{0} f_p^{ALD}(\epsilon; G)\mathrm{d}\epsilon = p$. In order to specify the model in hierarchical form we associate a latent mixing parameter σ_i with each y_i:

$$\begin{array}{rcl} Y_i \mid \sigma_i & \overset{ind.}{\sim} & k_p^{AL}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}; \sigma_i), \;\; i = 1, ..., n \\ \sigma_i \mid G & \overset{iid}{\sim} & G, \;\; i = 1, ..., n \\ G \mid \alpha, d & \sim & \mathrm{DP}(\alpha G_0) \end{array} \tag{4}$$

with independent normal priors for the components of $\boldsymbol{\beta}$. We refer to (3), or (4), as model $\mathcal{M}_{ALD}$. Model $\mathcal{M}_{ALD}$ extends model $\mathcal{M}_0$ allowing for increased tail variability in the error distribution. However, the skewness of the ALD kernel implies the same skewness of the mixture and thus forces $\mathcal{M}_{ALD}$ to be as inflexible as $\mathcal{M}_0$ regarding skewness.

Nonparametric scale mixtures of uniform densities

In order to obtain more flexible mixture models we need a more suitable kernel (mixand) distribution than $\mathcal{M}_{ALD}$, given that our mixing distribution, G, being random is already general and flexible. Representation for non-increasing densities on the positive real line is a key result for developing a flexible mixture: for any non-increasing density $f(\cdot)$ on R^+ there exists a distribution function G, defined on R^+, such that $f(t) \equiv f(t; G) = \int \theta^{-1} 1_{[0,\theta)}(t)\mathrm{d}G(\theta)$, i.e., $f(\cdot)$ can be expressed as a scale mixture of uniform densities. A requirement for G is to be general, or random in Bayesian modelling framework, which implies a stochastic (DP) prior for G.

Similarly, any unimodal density on the real line with p-th quantile (and mode) equal to zero, can be represented as $\iint k_p(\epsilon; \sigma_1, \sigma_2)\mathrm{d}G_1(\sigma_1)\mathrm{d}G_2(\sigma_2)$, where G_1 and G_2 are general mixing distributions, supported on R^+, and

$$k_p(\epsilon; \sigma_1, \sigma_2) = \frac{p}{\sigma_1} 1_{(-\sigma_1,0)}(\epsilon) + \frac{(1-p)}{\sigma_2} 1_{[0,\sigma_2)}(\epsilon), \tag{5}$$

with $0 < p < 1$, and $\sigma_r > 0$, $r = 1, 2$. Specifying independent DP priors for G_1 and G_2, we obtain

$$f_p^{DP1}(\epsilon; G_1, G_2) = \iint k_p(\epsilon; \sigma_1, \sigma_2)\mathrm{d}G_1(\sigma_1)\mathrm{d}G_2(\sigma_2), \quad G_r \sim \mathrm{DP}(\alpha_r G_{r0}), r = 1, 2, \tag{6}$$

the model for the error density in (1). In the context of quantile regression, $f_p^{DP1}(\cdot; G_1, G_2)$ successfully captures general forms of skewness and tail behavior. The hierarchical model is

$$\begin{array}{rcl} Y_i \mid \boldsymbol{\beta}, \sigma_{1i}, \sigma_{2i} & \overset{ind}{\sim} & k_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}; \sigma_{1i}, \sigma_{2i}), \;\; i = 1, ..., n \\ \sigma_{ri} \mid G_r & \overset{iid}{\sim} & G_r, \;\; r = 1, 2, \;\; i = 1, ..., n \\ G_r \mid \alpha_r, d_r & \sim & \mathrm{DP}(\alpha_r G_{r0}), \;\; r = 1, 2, \end{array} \tag{7}$$

with independent normal priors for the regression coefficients. Model (6), or (7), will be referred to as model $\mathcal{M}_{DP1}$. The formulation in (6) indicates an alternative nonparametric family of error densities based on a single mixing distribution G assigned a DP prior $\mathrm{DP}(\alpha G_0^*)$. The new model, $\mathcal{M}_{DP2}$, for the random error density is given by

$$f_p^{DP2}(\epsilon; G) = \iint k_p(\epsilon; \sigma_1, \sigma_2) \mathrm{d}G(\sigma_1, \sigma_2), \qquad G \sim \mathrm{DP}(\alpha G_0^*). \tag{8}$$

The hierarchical formulation for $\mathcal{M}_{DP2}$ is analogous to (7), except that now the pairs of latent mixing parameters $(\sigma_{1i}, \sigma_{2i})$, are IID from a single G, and G_0^* is a a bivariate lognormal distribution.

Posterior inference under the models outlined above is done according to well-established simulation methods for DP mixture models that are based on a marginalization of random mixing distributions over their DP priors. Models $\mathcal{M}_{ALD}$, $\mathcal{M}_{DP1}$ and $\mathcal{M}_{DP2}$ were compared on the basis of their predictive performance. Details of MCMC algorithms implemented for models outlined here can be found in the Appendix of Kottas and Krnjajić (2008).

We note the following properties of our modelling approach: (1) it is necessary to fit a model separately for each quantile, p, and (2) quantile lines for extreme percentiles in regions of sparse data may cross. However, quantile regression is usually applied for only a handful of (non-extreme) percentiles so neither of these properties is in the way of successfully applying the proposed models.

Model comparison; censored data

In order to check the modelling methodology we generated data sets from a variety of mixture distributions and compared models' performance in posterior predictive space using formal posterior predictive criteria, such as conditional predictive ordinate (CPO) plots and a posterior predictive loss approach. Here we only illustrate how models fit on a data set generated from a mixture of normals having 0.6-th quantile at zero and a right skewed density with non-standard tail behavior. As expected, the posterior predictive density of model $\mathcal{M}_{ALD}$ fails to capture the data set as depicted in Figure 1 whereas both $\mathcal{M}_{DP1}$ and $\mathcal{M}_{DP2}$ capture well the shape of the data histogram.

All quantile regresion models outlined here need a minor extension to enable them to work with censored data. The only change required when specifying $\mathcal{M}_{DP1}$ and $\mathcal{M}_{DP2}$ is in the likelihood stage to incorporate censored observations. For instance, with y_{i_o} and y_{i_c} denoting, on a logarithmic scale, the observed survival times t_{i_o} and the right censorship times z_{i_c}, respectively, and with $\boldsymbol{x}_{i_o}$ and $\boldsymbol{x}_{i_c}$ denoting the corresponding covariate vectors, the first stage in (7) for model $\mathcal{M}_{DP1}$ becomes

$$\prod_{i_o=1}^{n_o} k_p(y_{i_o} - \boldsymbol{x}_{i_o}^T \boldsymbol{\beta}; \sigma_{1,i_o}, \sigma_{2,i_o}) \prod_{i_c=1}^{n_c} (1 - K_p(y_{i_c} - \boldsymbol{x}_{i_c}^T \boldsymbol{\beta}; \sigma_{1,i_c}, \sigma_{2,i_c}))$$

where $K_p(\cdot; \sigma_1, \sigma_2)$ denotes the distribution function of $k_p(\cdot; \sigma_1, \sigma_2)$. For results of analyses of real data sets with censoring see Kottas and Krnjajić (2008).

Dependent DP mixture models for error distributions

We now propose an extension of the standard modeling framework in (1) to a class of quantile regression models where the error density $f_p(\cdot)$ depends on the covariates. Only a high level outline is provided here, while the details are in Kottas and Krnjajić (2005) and (2008).

For a simpler exposition, we consider only a single continuous covariate x with realized values x_m, $m = 1, ..., M$. For any specified quantile p, the error distribution under (1) is the same for all values of x and hence the response distribution changes with x only through the p-th quantile $\beta_0 + \beta_1 x$. Extension to nonparametric covariate-dependent error distributions requires a nonparametric prior model for the stochastic process of error densities indexed by values x in the covariate space $\mathcal{X}$, i.e., for $f_{p,\mathcal{X}} = \{f_{p,x}(\cdot) : x \in \mathcal{X}\}$, where for each fixed x, $\int_{-\infty}^{0} f_{p,x}(\varepsilon)\mathrm{d}\varepsilon = p$. Hence, in this setting, $f_{p,x}(\cdot)$ and $f_{p,x'}(\cdot)$ are dependent for all $x \neq x'$. In fact, we would typically seek a specification that yields *similar* $f_{p,x}(\cdot)$ and $f_{p,x'}(\cdot)$ for x close to x'. We employ dependent Dirichlet processes (DDPs) to formulate a prior probability model for $f_{p,\mathcal{X}}$, so that for each index value x, $f_{p,x}$ is a (random) DP. The DDP was developed by MacEachern (2000) as a nonparametric prior for a stochastic process of random distributions.

We provide only a sketch of reasoning that leads to a hierarchical formulation of a DDP model based on model $\mathcal{M}_{DP1}$. To allow $f_p^{DP1}(\epsilon; G_1, G_2)$ to change with x, we need mixing distributions G_1 and G_2 that change with x and are still assigned nonparametric priors; that is, we need prior probability models for the stochastic processes $\{G_r(x) : x \in \mathcal{X}\}$, where $G_r(x)$, $r = 1, 2$, are the mixing distributions for covariate value x. An extension of the DP (a prior model for the distribution function G_r) to a DDP (a prior model for the stochastic process $\{G_r(x) : x \in \mathcal{X}\}$) arises by replacing the univariate base distribution function G_{r0}, with a base stochastic process $G_{r0,\mathcal{X}}$ over $\mathcal{X}$ taking values in R. Introducing mixing through independent DDP priors $G_{1,\mathcal{X}}$ and $G_{2,\mathcal{X}}$ yields a prior for the collection $f_{p,\mathcal{X}}$ of quantile regression error densities. In particular, for any x, we obtain model $\mathcal{M}_{DP1}$ as the induced DP mixture model,

$$f_{p,x}^{DP1}(\epsilon; G_{1,x}, G_{2,x}) = \iint k_p(\epsilon; \theta_1(x), \theta_2(x))\mathrm{d}G_{1,x}(\theta_1(x))\mathrm{d}G_{2,x}(\theta_2(x)),$$

with $G_{r,x} \sim \mathrm{DP}(\alpha_r G_{r0}(x))$, $r = 1, 2$, and $G_{r0}(x) = \mathrm{N}(\mu_r, \tau_r^2)$. However, now the random error densities are dependent with the extent of dependence driven by $G_{1,\mathcal{X}}$ and $G_{2,\mathcal{X}}$. More generally, for the vector $\boldsymbol{x}$, we can write

$$f_{p,\boldsymbol{x}}^{DP1}(\boldsymbol{\epsilon}; G_{1,\boldsymbol{x}}, G_{2,\boldsymbol{x}}) = \iint \prod_{m=1}^{M} k_p(\epsilon_m; \theta_1(x_m), \theta_2(x_m))\mathrm{d}G_{1,\boldsymbol{x}}(\boldsymbol{\theta}_1)\mathrm{d}G_{2,\boldsymbol{x}}(\boldsymbol{\theta}_2),$$

where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_M)$, $\boldsymbol{\theta}_r = (\theta_r(x_1), ..., \theta_r(x_M))$, and $G_{r,\boldsymbol{x}} \sim \mathrm{DP}(\alpha_r G_{r0}(\boldsymbol{x}))$, $r = 1, 2$, with $G_{r0}(\boldsymbol{x})$ the M-variate normal distribution. We note that, in practice, learning with DDP priors is facilitated by some form of replication in the response values, i.e, more than one response value for each x_m, $m = 1, ..., M$. Let $\boldsymbol{y}_i = (y_{i1}, ..., y_{iM})$, $i =$

$1, ..., N$, be the i-th response replicate. Using data augmentation methods, the model can also be fitted when some of the y_{im} are missing. Let $\boldsymbol{\theta}_{ri} = (\theta_{ri}(x_1), ..., \theta_{ri}(x_M))$, $r = 1, 2$, be the latent mixing vectors associated with $\boldsymbol{y}_i$ and

$$f_p(\boldsymbol{y}_i; \boldsymbol{x}, (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}) = \prod\nolimits_{m=1}^{M} k_p(y_{im} - (\beta_0 + \beta_1 x_m); \theta_{1i}(x_m), \theta_{2i}(x_m)).$$

Then the quantile regression model is given by

$$\begin{array}{rcl} \boldsymbol{Y}_i \mid (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i} & \overset{ind.}{\sim} & f_p(\boldsymbol{y}_i; \boldsymbol{x}, (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}), \; i = 1, ..., N \\ \boldsymbol{\theta}_{ri} \mid G_{r,\boldsymbol{x}} & \overset{IID}{\sim} & G_{r,\boldsymbol{x}}, \; r = 1, 2, \; i = 1, ..., N \\ G_{r,\boldsymbol{x}} \mid \alpha_r, \mu_r, \tau_r^2, \phi_r & \sim & \mathrm{DP}(\alpha_r G_{r0}(\boldsymbol{x}) = \mathrm{N}_M(\mu_r 1_M, V_r)), \; r = 1, 2, \end{array} \tag{9}$$

with independent priors for all hyperparameters.

Model (9) is a DP mixture model with the M-variate DP priors for $G_{r,\boldsymbol{x}}$ induced by the DDP priors for $G_{r,\mathcal{X}}$. Hence, as for models $\mathcal{M}_{DP1}$ and $\mathcal{M}_{DP2}$, posterior sampling proceeds by marginalizing $G_{r,\boldsymbol{x}}$ over their DP priors, and utilizing an MCMC method for DP mixtures. Regarding predictive inference, interest lies in the posterior predictive density of response at observed covariate values in $\boldsymbol{x}$ as well as at new (unobserved) values, say $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_U)$. The predictive densities are computed based on the samples from the joint posterior distribution of all model parameters: $p(\tilde{y} \mid \boldsymbol{x}, \boldsymbol{y}) = \int_\Theta p(\tilde{y} \mid \boldsymbol{x}, \theta) p(\theta \mid \boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\theta$.

Analysis of data from a genotoxicity experiment

To illustrate the DDP quantile regression model we present some results of analysis of a comet assay dataset from a genotoxicity experiment, see Dunson and Taylor (2005). The data were drawn from a genotoxicity experiment assessing the effect of oxidative damage on the frequency of DNA strand breaks. Samples of cells exposed to different levels of hydrogen peroxide (d_m = 0, 5, 20, 50, 100 uM H_2O_2) were prepared for use in the comet assay. After electrophoresis, the DNA from the nucleus of cells with a high frequency of DNA strand breaks exhibit a comet-style shape, with the nucleus forming the ball-like head and the cut DNA strands the tail. Cells with a low frequency tend to maintain an approximately spherical shape with less tail. The objective of the experiment was to evaluate the sensitivity of the comet assay in detecting genotoxic effects of hydrogen peroxide. Measurements on the % DNA in the comet tail are available for 100 cells in each of the five dose groups. We take the response, y_{im}, for cell i ($i = 1, ..., N = 100$) in dose group m ($m = 1, ..., M = 5$), to be the % DNA in the comet tail divided by 100 (whence $0 \leq y_{im} \leq 1$). The p-th quantile regression function is assumed to be $\beta_0 + \beta_1 x_m$, where $x_m = \log(d_m + 1)$ and β_0, β_1 are the p-th quantile regression coefficients.

The histograms of the response values at the five dose groups (included in Figure 3) suggest that the different shapes for the response distribution will not be captured by a single quantile regression term. Thus, applying the DDP model (9) seems well suited for these data. We obtain results for p = 0.1, 0.25, 0.5, 0.75, and 0.9. Regarding inference for the regression coefficients, of primary interest are the five slope parameters β_1, which indicate how the 0.1, 0.25, 0.5, 0.75, and 0.9 quantiles of the distribution of % DNA in the comet tail change with dose. Posterior medians and 95% central posterior interval estimates for β_1 are 0.0152 and $(0.0097, 0.018)$ for $p = 0.1$; 0.0248 and $(0.0208, 0.0292)$ for $p = 0.25$; 0.0487 and $(0.0442, 0.0564)$ for $p = 0.5$; 0.0671 and $(0.0613, 0.0713)$ for $p = 0.75$; 0.0747 and $(0.0716, 0.0765)$ for $p = 0.9$. These values of the slope show that the frequency of DNA strand breaks increases with increasing dose of hydrogen peroxide, and that this increasing trend is stronger at the higher quantiles.

In the panels of Figure 2, we show the comet assay data and empirical quantiles along with the quantile lines from the DDP model (9). The lower panel summarizes the evidence that the frequency of DNA breaks increases with dose, with this trend stronger at the higher quantiles. Since empirical percentiles suggest a slightly quadratic shape for the quantile lines, we fitted the same DDP model using the p-th quantile function $h(\boldsymbol{x}) = \beta_0 + \beta_1 x + \beta_2 x^2$, and obtained quadratic quantile lines as shown in the upper panel of the same figure, indicating a similar conclusion regarding the dose-response relationship. Moreover, the ability of the DDP mixture to accurately model the error structure with different distributional shapes for different ranges of dose values, is clearly illustrated in Figure 3, which includes posterior predictive densities at the five observed and at three new dose values. It is clear from the shapes of the predictive densities at non-observed values of x $(10, 40, 95)$ that learning occurs from responses at the nearby covariates.

Conclusions

We have developed increasingly more flexible models for the error distribution in quantile regression using a representation for unimodal densities on the real line with a specified quantile equal to zero, and specifying DP priors for mixing distributions. We have illustrated the ability of these classes of nonparametric mixture models to adapt to any shape of unimodal error densities with p-th quantile at zero and therefore exhibit better fitting and predictive properties than parametric models. We have also proposed a model for quantile regression error densities that change with values in the covariate space, using DDP mixing for scale mixtures of uniform densities.

References

[1] Dunson, D, and Taylor, J. (2005). “Approximate Bayesian Inference for Quantiles,” *Journal of Nonparametric Statistics* Vol 17, 3, 2005

[2] Ferguson, T.S. (1974), “Prior Distributions on Spaces of Probability Measures,” *The Annals of Statistics*, 2, 615-629.

[3] Hjort N, et al. (2010). “Bayesian Nonparametrics”, *Cambridge University Press*, 2010.

[4] Koenker, R. (2005). *Quantile Regression*, Cambridge University Press.

[5] Kottas, A, and Krnjajić, M (2005). "Bayesian Nonparametric Modeling in Quantile Regression", Technical Report AMS2005-6, University of California at Santa Cruz (2005).

[6] Kottas, A., and Krnjajić, M. (2009). "Bayesian Semiparametric Modelling in Quantile Regression" *Scandinavian Journal of Statistics*, 36, 297-319.

[7] MacEachern, S.N. (2000). "Dependent Dirichlet Processes," Technical Report, Department of Statistics, The Ohio State University (2000).

[8] Müller P, Mitra R (2013). "Bayesian Nonparametric Inference: Why and How", *Bayesian Analysis*, Volume 8, Number 2 (2013), 269-302.

[9] Reich B, Smith L (2013). "Bayesian quantile regression for censored data" *Biometrics*, 2013 69 (3)

[10] Stigler, S. (1984). "Boscovich, Simpson and a 1760 manuscript note on fitting a linear relation", *Biometrika* 71 (3)

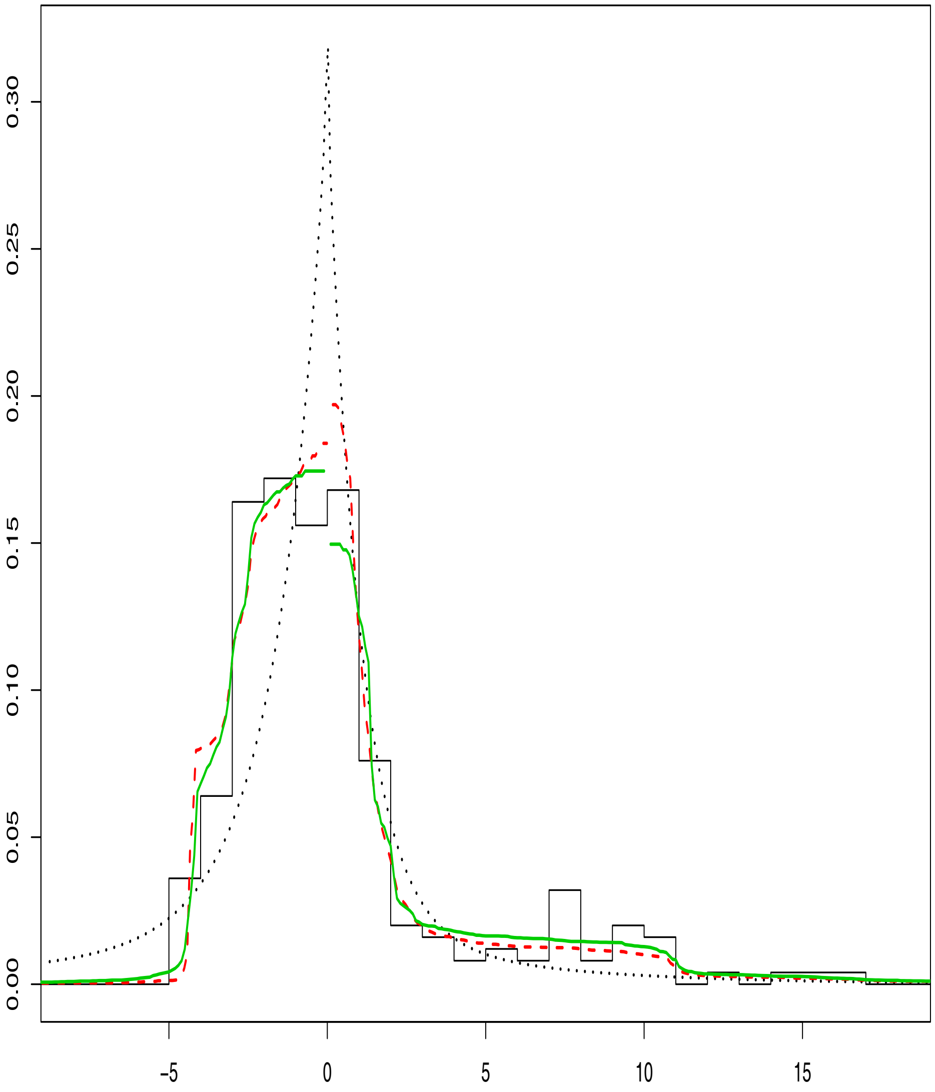


Figure 1: *Simulation study. Posterior predictive densities under model $\mathcal{M}_{ALD}$ (dotted line), model $\mathcal{M}_{DP1}$ (dashed line), and model $\mathcal{M}_{DP2}$ (solid line) for the case of the right skewed normal mixture with 0.6-th quantile at zero. The histogram of the simulated data is also included.*

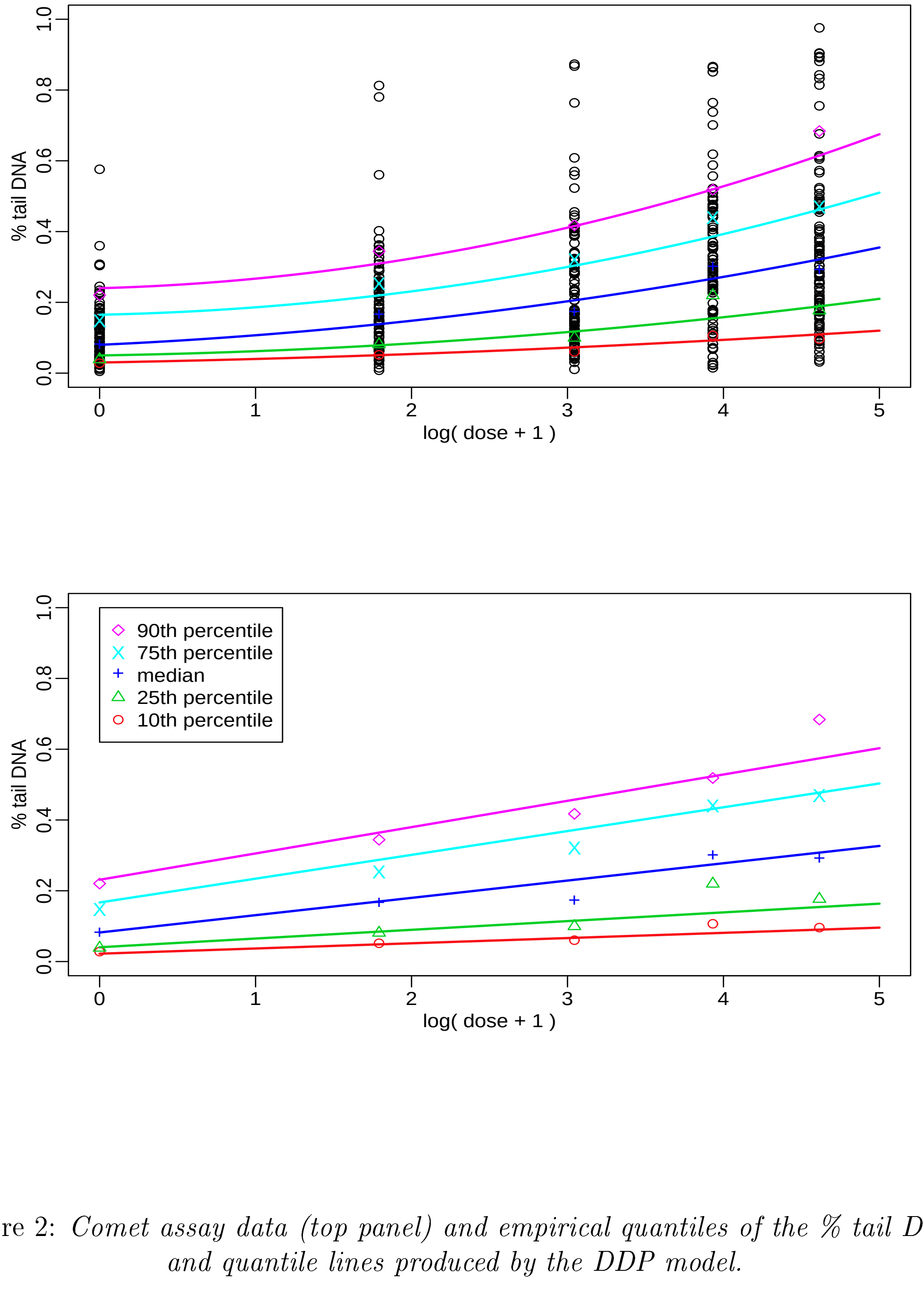


Figure 2: *Comet assay data (top panel) and empirical quantiles of the % tail DNA and quantile lines produced by the DDP model.*

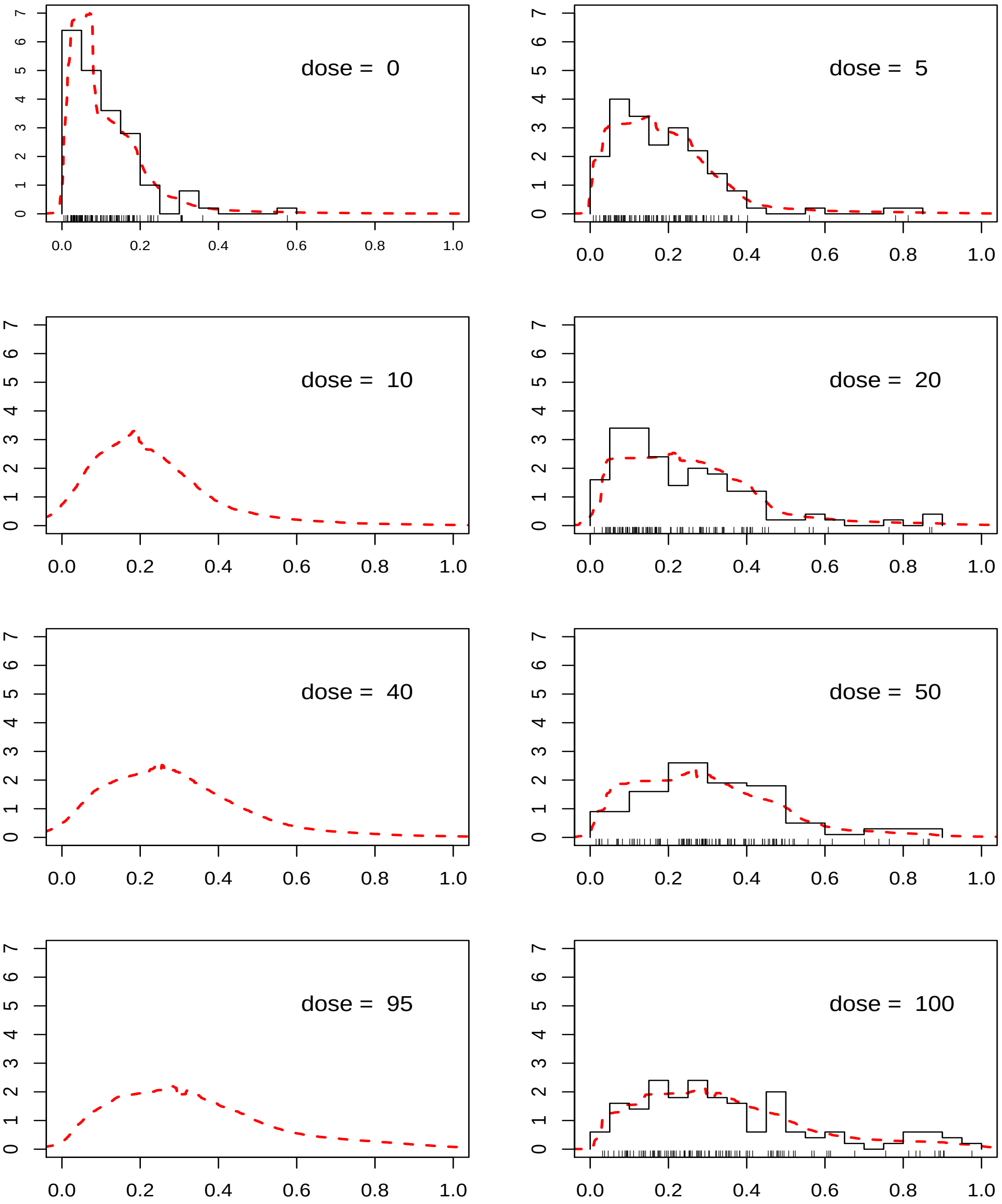


Figure 3: Comet assay data. Posterior predictive densites under the DDP model (dashed lines) at the five observed dose values, overlaid on histograms of the corresponding response observations, and at three new dose values (10, 40, and 95).

Bayesian Approach to Object Recognition under the Conditions of Fuzzy Information

Ganicheva A.V.[1] and Ganichev A.V.[2]
[1] *Tver State Agricultural Academy, Tver, Russia*
[2] *Tver State Technical University, Tver, Russia*
e-mail: alexej.ganichev@yandex.ru

Abstract

Goal of the research: Development of the Bayesian method of recognition of many classes of objects in the absence of a priori probabilities of their occurrence. Research methodology: Derivation of conditions for acceptance of optimal recognition with no data on initial a priori probabilities based on conditional probabilities and a posteriori probabilities from a previous stage. Findings: With a definite number of observations, a posteriori probabilities do not depend on the initial a priori ones and can be used for finding optimal solutions. Applicable scope of the findings: Obtained results can be used in the recognition of many classes under the conditions of clear and fuzzy information about probabilistic characteristics of the objects.

Keywords: object recognition, Bayesian approach, a priori, a posteriori probabilities, clear and fuzzy information, fuzzy number, triangular representation, the principle of the absence of aftereffect.

The Bayesian approach to the recognition of object classes consists in the calculation of conditional a posteriori probabilities and making decisions by comparison of their values [1]. The method is optimal according to the minimum of medium risk and minimum of erroneous decisions criteria [2]. When recognizing the objects, such situations are possible, where a priori probabilities of occurrence of the objects of a relevant class are unknown. It is assumed that it does not seem possible to minimize the value of the medium risk of decision making based on the Bayesian strategy in this case [3]. In [4, 5] a method of recognition of two classes of objects (hypotheses) based on multiple repeated finding of a posteriori probabilities, when a posteriori probabilities that have been calculated at the previous stage are used as a priori probabilities at this stage, is developed. The relations for the number of tests, whereby a posteriori probabilities become independent on the initial conditional probabilities with a set degree of accuracy, are found in the research. The accuracy is set through the initial conditional probability ratio. In this research, similar relations are found for the general case of many classes of objects (many hypotheses).

Let us assume that there are m hypotheses $H_1, H_2, \ldots, H_m$, related to some event A. Assume that $b_i = P(H_i), (i = 1, \ldots, m)$ - are true, but unknown to us probabilities of the hypotheses; $a_i = P(A/H_i), (i = 1, \ldots, m)$ - are known conditional probabilities of the event A within the framework of each of the hypotheses. The case of clearly defined probabilities is considered in [4, 5].

Let us consider the general case, when , b_i, a_i - are fuzzy numbers. In this case, their triangular representation can be applied: $b_i = \lfloor c_i', \overline{b_i}, d_i' \rfloor$, $a_i = \lfloor e_i', \overline{a_i}, u_i' \rfloor$ left

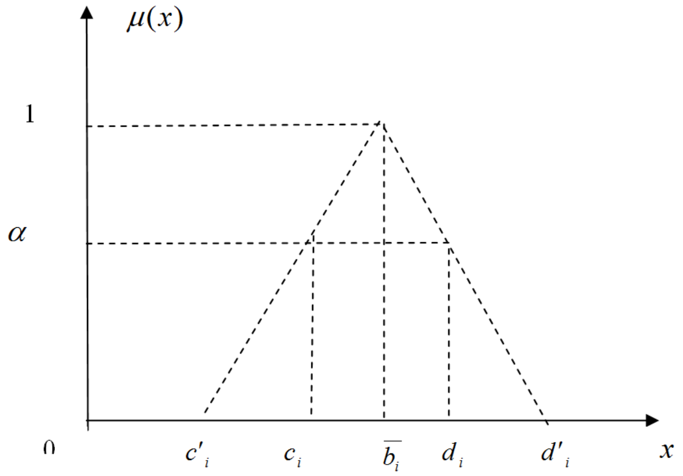


Figure 1

boundaries here correspond to the minimum possible value of the given number, right boundaries - to the maximum possible value, middle numbers - to the most expected value, which is found as the arithmetic average of the boundaries [6]. When the level of membership α is fixed, the specified segments are transformed correspondingly into narrower segments with the same centre. The boundaries of fuzzy numbers will correspond to the abscissae of piercing points of a line corresponding to the level α with the membership function of the given fuzzy number (Figure 1). The said is shown in Figure 1 in relation to the fuzzy number b_i.

The fuzzy number x will be set in the form $[x_1, x_2]$, where x_1 and x_2 - are, correspondingly, the left and the right membership interval boundary for the stated level of membership α. If a clear number is under consideration, then given the set level of membership it will be set via $[x, x]$.

Since the probabilities of the hypotheses are unknown, let us assume them to be equal, 0.5 each, i.e. $P_1(H_i) = 0.5$. These are our subjective (approximate) probabilities, which will be marked with the index 1 below. At that, the clear number 0.5 will be presented in the form $[0.5; 0.5]$. Arithmetic operations with fuzzy numbers come down to the relevant operations with their boundaries.

Let us assume that the event A has happened. Let us find a posteriori probabilities of the hypotheses. According to the formula of total probability, we have:

$$P(H_i/A) = \frac{[c_i, d_i] \cdot [e_i, u_i]}{\sum\limits_{i=1}^{m} [c_i, d_i] \cdot [e_i, u_i]} = \frac{[c_i e_i, d_i u_i]}{\sum\limits_{i=1}^{m} [c_i e_i, d_i u_i]} = \frac{[c_i e_i, d_i u_i]}{\left[\sum\limits_{i=1}^{m} c_i e_i, \sum\limits_{i=1}^{m} d_i u_i\right]} = \left[\frac{c_i e_i}{\sum\limits_{i=1}^{m} c_i e_i}, \frac{d_i u_i}{\sum\limits_{i=1}^{m} d_i u_i}\right], \quad (2)$$

similarly,

$$P_1(H_i/A) = \left[\frac{e_i}{\sum\limits_{i=1}^{m} e_i}, \frac{u_i}{\sum\limits_{i=1}^{m} u_i}\right].$$

Let us assume that the test was repeated under the same conditions; let us calculate again a posteriori probabilities of the hypotheses, but the a posteriori probabilities that have been calculated in the first step will now be taken as a priori ones. For that, let us first find $P(A)$ and $P_1(A)$; at that, it will be kept in view that summing up goes from 1 to m.

$$P(A)=\sum_{i=1}^{m}\left[\frac{c_ie_i}{\sum\limits_{i=1}^{m}c_ie_i},\frac{d_iu_i}{\sum\limits_{i=1}^{m}d_iu_i}\right]\cdot[e_i,u_i]=\sum_{i=1}^{m}\left[\frac{c_ie_i^2}{\sum\limits_{i=1}^{m}c_ie_i},\frac{d_iu_i^2}{\sum\limits_{i=1}^{m}d_iu_i}\right]=\left[\sum_{i=1}^{m}\frac{c_ie_i^2}{\sum\limits_{i=1}^{m}c_ie_i},\sum_{i=1}^{m}\frac{d_iu_i^2}{\sum\limits_{i=1}^{m}d_iu_i}\right],$$

thus

$$P(A)=\left[\frac{\sum\limits_{i=1}^{m}c_ie_i^2}{\sum\limits_{i=1}^{m}c_ie_i},\frac{\sum\limits_{i=1}^{m}d_iu_i^2}{\sum\limits_{i=1}^{m}d_iu_i}\right];$$

similarly,

$$P_1(A)=\left[\frac{\sum\limits_{i=1}^{m}e_i^2}{\sum\limits_{i=1}^{m}e_i},\frac{\sum\limits_{i=1}^{m}u_i^2}{\sum\limits_{i=1}^{m}u_i}\right].$$

Hence,

$$P(H_i/A)=\left[\frac{c_ie_i^2}{\sum\limits_{i=1}^{m}c_ie_i},\frac{d_iu_i^2}{\sum\limits_{i=1}^{m}d_iu_i}\right]:\left[\frac{\sum\limits_{i=1}^{m}c_ie_i^2}{\sum\limits_{i=1}^{m}c_ie_i},\frac{\sum\limits_{i=1}^{m}d_iu_i^2}{\sum\limits_{i=1}^{m}d_iu_i}\right]=\left[\frac{c_ie_i^2}{\sum\limits_{i=1}^{m}c_ie_i^2},\frac{d_iu_i^2}{\sum\limits_{i=1}^{m}d_iu_i^2}\right],$$

$$P_1(H_i/A)=\left[\frac{e_i^2}{\sum\limits_{i=1}^{m}e_i},\frac{u_i^2}{\sum\limits_{i=1}^{m}u_i}\right]\left[\frac{\sum\limits_{i=1}^{m}e_i^2}{\sum\limits_{i=1}^{m}e_i},\frac{\sum\limits_{i=1}^{m}u_i^2}{\sum\limits_{i=1}^{m}u_i}\right]=\left[\frac{e_i^2}{\sum\limits_{i=1}^{m}e_i^2},\frac{u_i^2}{\sum\limits_{i=1}^{m}u_i^2}\right].$$

Let us now assume that the test is being repeated n times under the same conditions. Then, by analogy with the previous formulae (at $n=1$ and $n=2$), it is possible to get a posteriori probabilities at the n^{th} step:

$$P(H_i/A)=\left[\frac{c_ie_i^n}{\sum\limits_{i=1}^{m}c_ie_i^n},\frac{d_iu_i^n}{\sum\limits_{i=1}^{m}d_iu_i^n}\right],P_1(H_i/A)=\left[\frac{e_i^n}{\sum\limits_{i=1}^{m}e_i^n},\frac{u_i^n}{\sum\limits_{i=1}^{m}u_i^n}\right]. \quad (3)$$

Let us assume that $\max\limits_i c_i$ is achieved at $i=i0$. Let us simplify expressions found, divide the numerator and denominator of the first fraction at $P(H_i/A)$ by $c_{i0}e_{i0}^n$, of the first fraction at $P_1(H_i/A)$ - by e_{i0}^n . It should be mentioned that where n is sufficiently large, the expressions $(e_i/e_{i0})^n\to 0$ where $i\neq i0$ and $(e_{i0}/e_{i0})^n\to 1$.

Apart from that, the second coordinate of a fuzzy number cannot be smaller than the first one, and since that is a probability, then it cannot be larger than 1. Thus,

$$P(H_{i0}/A) \approx P_1(H_{i0}/A) \to [1,1]. \qquad (4)$$

And as far as the sum of a posteriori probabilities is equal to 1, then $P(H_i/A) \to [0,0]$, $P_1(H_i/A) \to [0,0]$ at all the $i \neq i0$, i.e. for these values of “i”

$$P(H_i/A) \approx P_1(H_i/A) \to [0,0]. \qquad (5)$$

Thus, when the number n of times the test was repeated is sufficiently large, a posteriori probability of the hypothesis with a larger initial conditional probability will tend to 1 and the one of the hypotheses with smaller initial conditional probabilities will tend to 0.

Hence, a fundamental conclusion follows: when the test is repeated multiple times (under the same conditions), the initial ratio of classes (hypotheses) it almost ceases to have an impact on the final result.

For example, let us assume that the test consists in the analysis of manufacturing activities of two companies, each of which is characterized by both positive and negative indices. Let us consider the event A - some kind of activity, which is characteristic of positive indices, is registered in the randomly selected company. Then, if the observation will be repeated multiple times the type of the index will no longer depend on the company.

Let us proceed to the formalization of the notion “multiple repeat of the test”. Let us set the accuracy $0 < \varepsilon < 1$. For the purposes of consistency, indices will not be specified, i.e. the number of iterations of the test will be defined in the general case - both for the left and for the right boundary of the fuzzy number. The final number will be defined as a maximum from the relevant numbers for the left and for the right boundary. Let us require the following condition to be met for any $i \neq i0$:

$$\frac{c_i}{c_{i0}}\left(\frac{e_i}{e_{i0}}\right)^n < \varepsilon, \text{i.e} \left(\frac{e_i}{e_{i0}}\right)^n < \frac{c_{i0}}{c_i}\varepsilon.$$

Let us take the logarithm of the latter inequality: $\ln\left(\frac{e_i}{e_{i0}}\right)^n < \ln\left(\frac{c_{i0}}{c_i}\varepsilon\right)$. Let us transform this inequality: $n \cdot \ln\left(\frac{e_i}{e_{i0}}\right) < \ln\left(\frac{c_{i0}}{c_i}\varepsilon\right)$ hence,

$$n > \frac{\ln\left(\frac{c_{i0}}{c_i}\varepsilon\right)}{\ln\left(\frac{e_i}{e_{i0}}\right)} = \frac{\ln\left(\frac{c_{i0}}{c_i}\right) + \ln\varepsilon}{\ln\left(\frac{e_i}{e_{i0}}\right)},$$

or

$$n > \log_{\frac{e_i}{e_{i0}}} \frac{c_{i0}}{c_i} + \log_{\frac{e_i}{e_{i0}}} \varepsilon. \qquad (6)$$

for any $i = 1, \ldots, n$ and $i \neq i0$. Let us consider the example. Let us assume that $n = 3$, $i0 = 1$, $e_{i0} = 0.6$, $e_2 = 0.3$, $e_3 = 0.1$, $c_{i0} = 0.3$, $c_2 = 0.5$, $c_3 = 0.2$, $\varepsilon = 1\%$. Let us find

$$k_1 = \log_{\frac{0.3}{0.6}} \frac{0.3}{0.5} = \log_{0.5} 0.6 = \frac{\ln 0.6}{\ln 0.5} = 0.74, k_2 = \log_{\frac{0.1}{0.6}} \frac{0.3}{0.2} = \log_{0.17} 1.5 = \frac{\ln 1.5}{\ln 0.17} = -0.23.$$

At the same time $\frac{e_2}{e_{i0}} = \frac{0.3}{0.6} = 0.5$, $\varepsilon = 1\%$ and $\log_{\frac{e_1}{e_{i0}}} \varepsilon = \log_{0.5} 0.01 \approx 6.64$; $\frac{e_3}{e_{i0}} = \frac{0.1}{0.6} = 0.17$; $\log_{\frac{e_2}{e_{i0}}} 0.01 = \log_{0.17} 0.01 \approx 2.6$. Hence, $n \geq \max\{0.74{+}6.54; -0.23{+}2.6\} = 7.28$, i.e. $n \geq 8$. Thus, at the 8_{th} step of repeating the test with the measure of inaccuracy $\varepsilon = 1\%$ it is fair to state that

$$P(H_{i0}/A) = P_1(H_{i0}/A) = 1.$$

In some instances, the formula (6) is transformed to a simpler form.

1) If $c_{i0} > c_i$ for any $i \neq i0$, then $\log_{\frac{e_i}{e_{i0}}} \frac{c_{i0}}{c_i}$;

2) if $\varepsilon > \frac{e_i}{e_{i0}}$ for any $i \neq i0$, then $\log_{\frac{e_i}{e_{i0}}} \varepsilon < 1$.

When both the conditions, 1) and 2), are met concurrently, n - is any natural number, inter alia, n can possess small values, e.g. 1 or 2, at which the relations (6), (5) and (4) are fulfilled, i.e. $n \geq 1$;

3) $\varepsilon > \max\limits_i \frac{e_i}{e_{i0}}$, then $n > \log_{\frac{e_i}{e_{i0}}} \frac{c_{i0}}{c_i} + 1$ for any $i \neq i0$, besides, two sub-cases are possible here:

3a) let us assume that $\log_{\frac{e_i}{e_{i0}}} \frac{c_{i0}}{c_i} + 1 \leq 0$, and since $\log_{\frac{e_i}{e_{i0}}} \frac{e_{i0}}{e_i} = 1$, then $\log_{\frac{e_i}{e_{i0}}} \frac{c_{i0}}{c_i} + \log_{\frac{e_i}{e_{i0}}} \frac{e_{i0}}{e_i} \leq 0$; then $\log_{\frac{e_i}{e_{i0}}} \frac{c_{i0}}{c_i} \leq \log_{\frac{e_i}{e_{i0}}} \frac{e_{i0}}{e_i}$ and $\frac{c_{i0}}{c_i} \geq \frac{e_i}{e_{i0}}$. The converse is valid as well, i.e. if $\frac{c_{i0}}{c_i} \geq \frac{e_i}{e_{i0}}$, then $\log_{\frac{e_i}{e_{i0}}} \frac{c_{i0}}{c_i} + 1 \leq 0$. Thus, in the case $\frac{c_{i0}}{c_i} \geq \frac{e_i}{e_{i0}}$ n has any value, i.e. $n \geq 1$;

3b) $\frac{c_{i0}}{c_i} < \frac{e_i}{e_{i0}}$ - then $n > k + 1$, where $k = \log_{\frac{e_i}{e_{i0}}} \frac{c_{i0}}{c_i}$.

It is not too difficult to see that the conclusions will be the same in the case of clear numbers a_i, b_i. Thus, the case of multiple (when $n \to \infty$) repeat of the test under the same conditions characterized by the probabilities c_i, d_i, e_i and u_i, $(i = 1, \ldots, m)$ is considered. Let us assume that, test by test, the probabilities e_i and u_i will change, i.e. we have the sequences: $\{e_i^{(j)}\}$, $\{u_i^{(j)}\}$, $(i = 1, \ldots, m)$, $(j = 1, \ldots, n)$. Then, in the formulae (3) e_i^n will be replaced by the product $e_i^{(1)} \cdot e_i^{(2)} \cdot \ldots \cdot e_i^{(n)}$, and u_i - by the product $u_i^{(1)} \cdot u_i^{(2)} \cdot \ldots \cdot u_i^{(n)}$.

It should be mentioned that the left boundary of the fuzzy number $P(H_i/A)$ will not be smaller than $c_i(\min\limits_j e_i^{(j)})^n / \sum\limits_i c_i(\max\limits_j e_i^{(j)})^n$, and the left boundary of the fuzzy number $P_1(H_i/A)$ will not be smaller than $(\min\limits_j e_i^{(j)})^n / \sum\limits_i (\max\limits_j e_i^{(j)})^n$. Similar relations are valid for the relevant right boundaries as well, but there is "d" instead of "c", and "u" instead of "e". Fuzzy numbers with such boundaries will be set by means of $P'(H_i/A)$ and $P'_1(H_i/A)$ correspondingly. Let us assume that $i = i0$ - is index, at which $\max\limits_i(\min\limits_j e_i^{(j)})^n$ is achieved. Having divided the numerator and denominator at the left boundary of each number $P'(H_i/A)$ by $c_{i0} \max\limits_i(\min\limits_j e_i^{(j)})^n$, correspondingly, at each number $P'_1(H_i/A)$ - by $\max\limits_i(\min\limits_j e_i^{(j)})^n$, we will get the same situation, as for the case of repeating the test under the same conditions, described by the formulae (4) and (5) for the numbers $P'(H_i/A)$ and $P'_1(H_i/A)$, considered above, and hence - for the numbers $P(H_i/A)$ and $P_1(H_i/A)$.

From a philosophical point of view, the considered method is descriptive of the principle of the absence of aftereffect, when the future of a system depends only on the present and does not depend on the background, i.e. the way through which the system found itself in the given state. However, it should be mentioned that the formulae (3) allows extrapolating the process to the future and to the past, i.e. solving not only the problem of prognostication, but the one of the retrospective journey into previous states as well.

Suggested method of accounting a posteriori probability as a priori one can become widely used in the systems of decision-making, artificial intelligence, e.g. in expert systems.

References

[1] Ganichev A. V. (2009). Optimality Conditions of Methods Based on Distance Func-tions when Making Decisions on Classes of the Objects. *The bulletin of Tver State Technical University*. Vol. **14**, pp. 59-62.

[2] Ganichev A. V. (2011). The Optimality of Pattern Classification through the Use of Distance Functions. *Scientific-technical bulletin of the Volga region*. Vol. **6**, pp. 133-136.

[3] Gorelik A. L., Skripkin V. A. (1984). *Recognition Methods*. Vysshaya Shkola, Moscow.

[4] Ganicheva A. V. (2008). Peculiarities of Decision-Making in the Social and Economic Sphere. *Interacademic scientific conference proceedings. - Tver: Tver Branch of the Moscow Humanitarian and Economic Institute.*

[5] Ganicheva A. V. (2014). Adaptive Bayesian Method of Decision-Making. *In the World of Scientific Discoveries*. Vol. **2.1 (50)**, pp. 618-633.

[6] Rutkovskaya D., Pilinskiy M., Rutkovskaya L. (2006). *Neural Networks, Genetic Algorithms and Fuzzy Systems*. Goryachaya Liniya - Telekom, Moscow.

**APPLIED METHODS OF STATISTICAL ANALYSIS.
NONPARAMETRIC APPROACH**

Proceedings
of the international workshop
Novosibirsk, 14-19 September 2015

Научное издание

**ПРИКЛАДНЫЕ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА.
НЕПАРАМЕТРИКА**

ТРУДЫ
Международного семинара
Новосибирск, 14-19 сентября 2015 г.

На английском языке

Подписано в печать 28.08.2015. Формат 60 × 84 1/8.
Усл. печ. л. 58,0. Уч.-изд. л. 107,88.
Тираж 60 экз. Заказ № 1219.

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20