APPLIED METHODS OF STATISTICAL ANALYSIS.

APPLICATIONS IN SURVIVAL ANALYSIS, RELIABILITY AND QUALITY CONTROL

PROCEEDINGS

of the International Workshop

25-27 September 2013

Novosibirsk

2013

UDC 519.22(063)
A 67

E d i t o r s:

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

A 67 **Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control** - AMSA'2013, Novosibirsk, Russia, 25-27 September, 2013: Proceedings of the International Workshop. - Novosibirsk: NSTU publisher, 2013. - 284 pp.
ISBN 978-5-7782-2311-0

ISBN 978-5-7782-2311-0 **UDC 519.22(063)**

Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control

C h a i r m e n:

Narayanaswamy Balakrishnan,	McMaster University, Canada
Mikhail Nikulin,	University of Bordeaux, France
Aleksey Vostretsov,	Novosibirsk State Technical University, Russia
Boris Lemeshko,	Novosibirsk State Technical University, Russia
Evgeny Tsoy,	Novosibirsk State Technical University, Russia

S c i e n t i f i c C o m m i t t e e:

A. Antonov,	Institute of Nuclear Power Engineering, Russia
N. Balakrishnan,	McMaster University, Canada
V. Denisov,	Novosibirsk State Technical University , Russia
K.-H. Eger,	Technische Univesitat Chemnitz, Germany
B. Lemeshko,	Novosibirsk State Technical University , Russia
N. Limnios,	Universite de Technologie de Compiegne, France
V. Melas,	St. Petersburg State University, Russia
M. Nikulin,	University of Bordeaux, France
B. Ryabko,	Siberian State University of Telecommunications and Information Sciences, Russia
V. Rykov,	Gubkin University of Oil and Gas, Russia
V. Solev,	St.Petersburg Department of Steklov Mathematical Institute RAS, Russia
A. Vostretsov,	Novosibirsk State Technical University , Russia
N. Zagoruiko,	Sobolev Institute of Mathematics of the Siberian Branch of the RAS, Russia

L o c a l C o m m i t t e e:

E. Chimitova, M. Semenova, V. Demin, N. Galanova

Preface

The Second International Workshop "Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control" -
AMSA'2013 is organized by Novosibirsk State Technical University. 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 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 First International Workshop "Applied Methods of Statistical Analysis. Simulations and Statistical Inference" - AMSA'2011 took place in Novosibirsk State Technical University, Novosibirsk, Russia. This city is very well known for its fundamental contributions to the development of theory of the probability, mathematical statistics, stochastical processes and statistical simulation. The meeting had focused on 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, simulation and research of probabilistic regularities, application of statistical methods.

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

Application of Nonparametric Goodness-of-Fit Tests for Composite Hypotheses in Case of Unknown Distributions of Statistics

Alisa A. Gorbunova, Boris Yu. Lemeshko, Stanislav B. Lemeshko, and Andrey P. Rogozhnikov
Department of Applied Mathematics, Novosibirsk State Technical University, K.Marx pr., 20, Novosibirsk, Russia
e-mail: gorbunova.alisa@gmail.com, lemeshko@fpm.ami.nstu.ru

Abstract

While testing composite hypotheses when a scalar or vector parameter of the probability distribution is calculated using the same sample, nonparametric Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling goodness-of-fit tests lose their distribution freedom. When testing composite hypotheses conditional distribution of the test statistic depends on several factors, even the specific values of the distribution shape parameters.

An interactive method for investigating distributions of nonparametric goodness-of-fit tests statistics, that allows us apply criteria for testing any composite hypotheses using a variety of estimation methods, is implemented.

Keywords: goodness-of-fit test, testing composite hypothesis, Kolmogorov test, Cramer-von Mises-Smirnov test, Anderson-Darling test, Kuiper test, Watson test, Zhang test.

Introduction

Classical nonparametric tests were constructed for testing simple hypotheses: $H_0 : F(x) = F(x, \theta)$, where θ is known scalar or vector parameter of the distribution function $F(x, \theta)$. When testing simple hypotheses nonparametric criteria are distribution free, i.e. the distribution $G(S|H_0)$, where S is the test statistic, does not depend on the $F(x, \theta)$ when the hypothesis H_0 is true.

When testing composite hypotheses of the form $H_0 : F(x) \in \{F(x, \theta), \theta \in \Theta\}$, where the estimate $\hat{\theta}$ of a scalar or vector parameter of the distribution $F(x, \theta)$ is calculated from the same sample, nonparametric tests lose the distribution freedom. Conditional distributions $G(S|H_0)$ of tests statistics for composite hypotheses depend on a number of factors: the type of the distribution $F(x, \theta)$, corresponding to the true hypothesis H_0; the type of the estimated parameter and the number of estimated parameters and, in some cases, the value of the parameter; the method of the parameter estimation.

1 Nonparametric goodness-of-fit criteria for testing simple hypotheses

In **Kolmogorov test** statistic the distance between the empirical and theoretical distribution is determined by

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x,\theta)|,$$

where $F_n(x)$ is the empirical distribution function, n is the sample size. When $n \to \infty$, distribution of statistic $\sqrt{n}D_n$ for true hypothesis under test uniformly converges to the Kolmogorov distribution [1]

$$K(S) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}.$$

While testing hypothesis using the Kolmogorov test it is advisable to use the statistic with Bolshev correction [2] given by [3]:

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}}, \tag{1}$$

where $D_n = \max(D_n^+, D_n^-)$,

$$D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F(x_i,\theta)\right\}, D_n^- = \max_{1\le i\le n}\left\{F(x_i,\theta) - \frac{i-1}{n}\right\}$$

n is the sample size, $x_1, x_2, \ldots, x_n$ are the sample values in an increasing order. When a simple hypothesis H_0 under test is true, the statistic () converges to the Kolmogorov distribution significantly faster than statistic $\sqrt{n}D_n$.

The statistic of **Cramer-von Mises-Smirnov test** has the following form [3]:

$$S_\omega = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(x_i,\theta) - \frac{2i-1}{2n}\right\}^2, \tag{2}$$

and **Anderson-Darling test** statistic [4], [5] is

$$S_\Omega = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i,\theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(x_i,\theta))\right\}. \tag{3}$$

When testing simple hypotheses, statistic (2) has the following distribution [3]

$$\begin{aligned} a1(s) &= \frac{1}{\sqrt{2s}}\sum_{j=0}^{\infty}\frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2}{16s}\right\}\times \\ &\times \left\{I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16s}\right] - I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16s}\right]\right\}, \end{aligned}$$

where $I_{-\frac{1}{4}}(\cdot)$ and $I_{\frac{1}{4}}(\cdot)$ are modified Bessel functions,

$$I_\nu(z) = \sum_{k=0}^{\infty} \frac{(z/2)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)}, \quad |z| < \infty, \quad |\arg z| < \pi$$

The statistic (3) has the distribution [3]

$$\begin{aligned} a2(s) &= \frac{\sqrt{2\pi}}{s} \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)} \exp\left\{-\frac{(4j+1)^2\pi^2}{8s}\right\} \times \\ &\times \int_0^{\infty} \exp\left\{\frac{s}{8(y^2+1)} - \frac{(4j+1)^2\pi^2 y^2}{8s}\right\} dy. \end{aligned}$$

The Kuiper test [6] is based on the statistic $V_n = D_n^+ + D_n^-$. The limit distribution of statistic $\sqrt{n}V_n$ while testing simple hypothesis is the following distribution function [7]:

$$G(s|H_0) = 1 - \sum_{m=1}^{\infty} 2(4m^2s^2 - 1)e^{-2m^2s^2}.$$

The following modification of statistics converges faster to the limit distribution [8]:

$$V = V_n\left(\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}}\right),$$

or the modification that we have chosen:

$$V_n^{mod} = \sqrt{n}(D_n^+ + D_n^-) + \frac{1}{3\sqrt{n}}. \tag{4}$$

Dependence of the distribution of statistic (4) on the sample size is practically negligible when $n \geq 30$.

As a model of limit distribution we can use the beta distribution of the third kind with the density

$$f(s) = \frac{\theta_2^{\theta_0}}{\theta_3 B(\theta_0, \theta_1)} \frac{\left(\frac{s-\theta_4}{\theta_3}\right)^{\theta_0-1} \left(1 - \frac{s-\theta_4}{\theta_3}\right)^{\theta_1-1}}{\left[1 + (\theta_2 - 1)\frac{s-\theta_4}{\theta_3}\right]^{\theta_0+\theta_1}},$$

and the vector of parameters $\theta = (7.8624, 7.6629, 2.6927, 0.495)^T$, obtained by the simulation of the distribution of the statistic (4).

Watson test [9], [10] is used in the following form

$$U_n^2 = \sum_{i=1}^{n} \left(F(x_i, \theta) - \frac{i - \frac{1}{2}}{n}\right)^2 - n\left(\frac{1}{n}\sum_{i=1}^{n} F(x_i, \theta) - \frac{1}{2}\right) + \frac{1}{12n} \tag{5}$$

The limit distribution of the statistic (5) while testing simple hypotheses is given by [9], [10]:

$$G(s|H_0) = 1 - 2\sum_{m=1}^{\infty}(-1)^{m-1}e^{-2m^2\pi^2 s}.$$

The good model for the limit distribution of the statistic (5) is the inverse Gaussian distribution with the density

$$f(s) = \frac{1}{\theta_2}\left(\frac{\theta_0}{2\pi\left(\frac{s-\theta_3}{\theta_2}\right)^2}\right)^{1/2} \exp\left(-\frac{\theta_0\left(\left(\frac{s-\theta_3}{\theta_2}\right)-\theta_1\right)}{2\theta_1^2\left(\frac{s-\theta_3}{\theta_2}\right)}\right)$$

and the vector of parameters $\theta = (0.2044, 0.08344, 1.0, 0.0)^T$, obtained by the simulation of the empirical distribution of the statistic (5). This distribution as well as the limit one could be used in testing simple hypotheses with Watson test to calculate the achieved significance level.

Zhang tests (Jin Zhang) were proposed in papers [11], [12], [13]. The statistics of these criteria are:

$$Z_K = \max_{1\le i\le n}\left(\left(i-\frac{1}{2}\right)\log\left\{\frac{i-\frac{1}{2}}{nF(x_i,\theta)}\right\} + \left(n-i+\frac{1}{2}\right)\log\left[\frac{n-1+\frac{1}{2}}{n\{1-F(x_i,\theta)\}}\right]\right), \tag{6}$$

$$Z_A = -\sum_{i=1} n\left[\frac{\log\{F(x_i,\theta)\}}{n-i+\frac{1}{2}} + \frac{\log\{1-F(x_i,\theta)\}}{i-\frac{1}{2}}\right], \tag{7}$$

$$Z_C = \sum_{i=1} n\left[\log\left\{\frac{[F(x_i,\theta)]^{-1}-1}{\left(n-\frac{1}{2}\right)/\left(i-\frac{3}{4}\right)-1}\right\}\right]^2. \tag{8}$$

The author gives the percentage points for statistics distributions for the case of testing simple hypotheses. The strong dependence of statistics distributions on the sample size n prevents one from wide use of the criteria with the statistics (6) — (8). For example, Figure 1 shows a dependence of the distribution of the statistics (7) on the sample size while testing simple hypotheses.

Of course, this dependence on the sample size n remains for the case of testing composite hypotheses.

2 Comparative analysis of the tests power

In papers [14], [15], [16] the power of Kolmogorov (K), Cramer-von Mises-Smirnov (KMS), Anderson-Darling (AD) tests, and also χ^2 criteria, was analyzed and compared for testing simple and composite hypotheses for a number of different pairs of competing distributions. In the case of testing simple hypotheses and using asymptotically optimal grouping [17] in χ^2 criterion, this test has the advantage in power

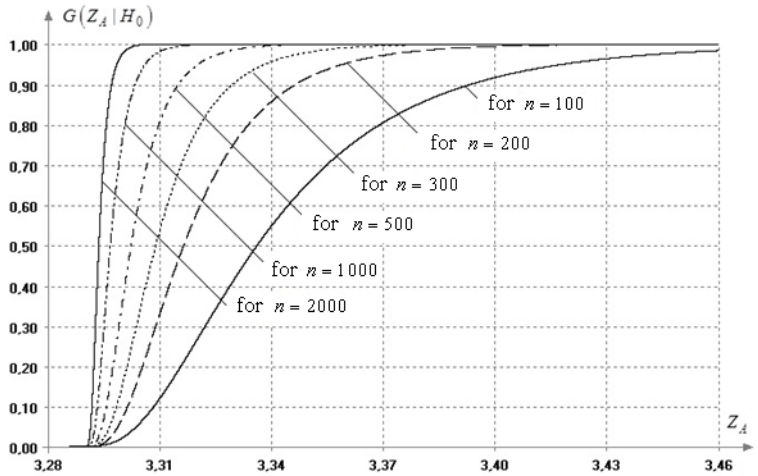

Figure 1: The distribution $G_n(Z_A|H_0)$ of statistic (7) depending on the sample size n for testing simple hypothesis

compared with nonparametric tests [14], [15]. When testing composite hypotheses, power of nonparametric tests increases significantly, and they become more powerful.

In order to be able to compare the power of Kuiper (V_n), Watson (U_n^2), and Zhang tests (Z_K, Z_A, Z_C) with the power of other goodness-of-fit tests, the power of these criteria was calculated for the same pairs of competing distributions in the paper [18] alike papers [14], [15], [16].

The first pair is the normal and logistics distribution: for the hypothesis H_0 — the normal distribution with the density:

$$f(x) = \frac{1}{\theta_0\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta_1)^2}{2\theta_0^2}\right\},$$

and for competing hypothesis H_1 — the logistic distribution with the density:

$$f(x) = \frac{\pi}{\theta_1\sqrt{3}} \exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\} \Bigg/ \left[1+\exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\right]^2,$$

and parameters $\theta_0 = 1$, $\theta_1 = 1$. For the simple hypothesis H_0 parameters of the normal distribution have the same values. These two distributions are close and difficult to distinguish with goodness-of-fit tests.

The second pair was the following: H_0 — Weibull distribution with the density

$$f(x) = \frac{\theta_0(x-\theta_2)^{\theta_0-1}}{\theta_1^{\theta_0}} \exp\left\{-\left(\frac{x-\theta_2}{\theta_1}\right)^{\theta_0}\right\}$$

and parameters $\theta_0 = 2$, $\theta_1 = 2$, $\theta_2 = 0$; H_1 corresponds to gamma distribution with the density

$$f(x) = \frac{1}{\theta_1 \Gamma(\theta_0)} \left(\frac{x - \theta_2}{\theta_1} \right)^{\theta_0 - 1} e^{-(x-\theta_2)/\theta_1}$$

and parameters $\theta_0 = 2.12154$, $\theta_1 = 0.557706$, $\theta_2 = 0$, when gamma distribution is the closest to the Weibull counterpart.

Comparing the estimates of the power for the Kuiper, Watson and Zhang tests [18] with results for Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests [14], [15], [16], the nonparametric tests can be ordered by decrease in power as follows:

- for testing simple hypotheses with a pair "normal — logistic": $Z_C \succ Z_A \succ Z_K \succ U_n^2 \succ V_n \succ AD \succ K \succ KMS$;
- for testing simple hypotheses with a pair "Weibull — gamma": $Z_C \succ Z_A \succ Z_K \succ U_n^2 \succ V_n \succ AD \succ KMS \succ K$;
- for testing composite hypotheses with a pair "normal — logistic": $Z_A \approx Z_C \succ Z_K \succ AD \succ KMS \succ U_n^2 \succ V_n \succ K$;
- for testing composite hypotheses with a pair "Weibull — gamma": $Z_A \succ Z_C \succ AD \succ Z_K \succ KMS \succ U_n^2 \succ V_n \succ K$.

3 The distribution of statistics for testing composite hypotheses

When testing composite hypotheses conditional distribution $G(S|H_0)$ of the statistic depends on several factors: the type of the observed distribution for true hypothesis H_0; the type of the estimated parameter and the number of parameters to be estimated, in some cases the parameter values (e.g., for the families of gamma and beta distributions), the method of parameter estimation. The differences between distributions of the one statistic for testing simple and composite hypotheses are very significant, so we could not neglect this fact. For example, Figure 2 shows the distribution of Kuiper statistic (4) for testing composite hypotheses for the different distributions using maximum likelihood estimates (MLE) of the two parameters.

Figure 3 illustrates the dependence of the distribution of the Watson test statistic (5) on the type and the number of estimated parameters having as an example the *Su*-Johnson distribution with a density:

$$f(x) = \frac{\theta_1}{\sqrt{2\pi}\sqrt{(x-\theta_3)^2 + \theta_2^2}} \exp\left\{ -\frac{1}{2} \left[\theta_0 + \theta_1 \ln\left\{ \frac{x-\theta_3}{\theta_2} + \sqrt{\left(\frac{x-\theta_3}{\theta_2} \right)^2 + 1} \right\} \right]^2 \right\}$$

Figure 4 shows the dependence of the distribution of Anderson-Darling test statistics (3) for testing composite hypotheses using MLEs of the 3 parameters of the generalized normal distribution depending on the value of the shape parameter θ_0.

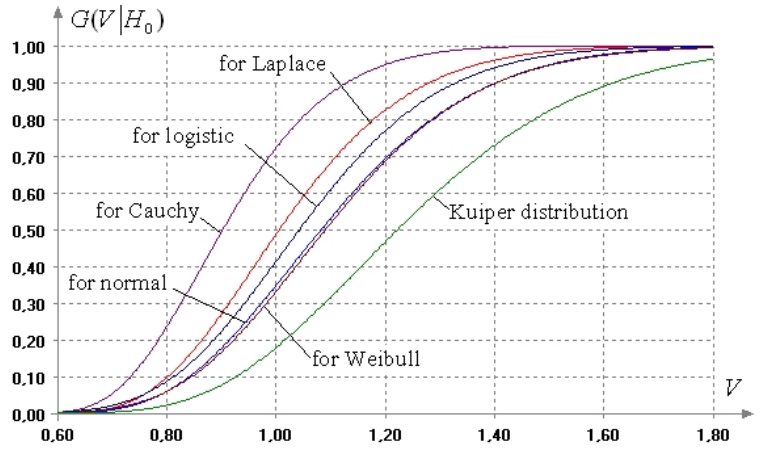

Figure 2: The distribution of Kuiper statistic (4) for testing composite hypotheses using MLEs of the two parameters

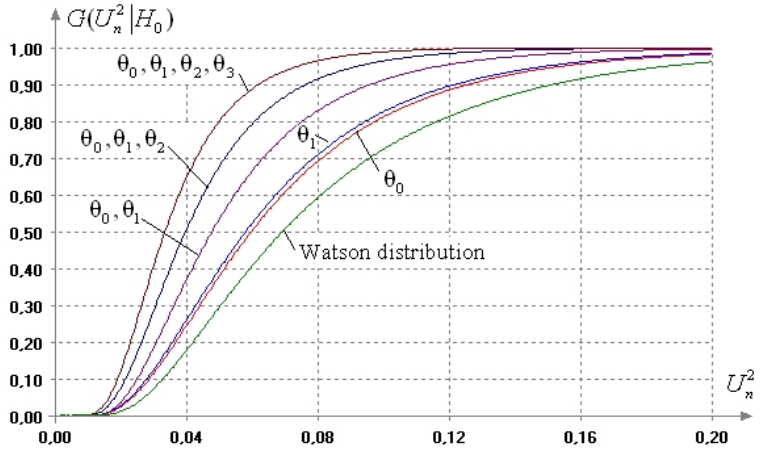

Figure 3: The distribution of Watson statistic (5) for testing composite hypotheses using MLEs of different number of parameters of the Su-Johnson distribution

The first work that initiates the study of limiting distributions of nonparametric goodness-of-fit statistics for composite hypotheses was [19]. Later, different approaches were used to solve this problem: the limit distribution was investigated by analytical methods [20]–[21], [22]–[23], [24]–[26], [27], [28]–[30], the percentage points were calculated using statistical modeling [31], [32]–[33], [34], the formulas were obtained to give a good approximation for small values of the probabilities [35], [36].

In our studies [37]–[49], the distribution of nonparametric Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests statistics were studied using statistical modeling.

Further, based on obtained empirical distribution of statistics, we construct an approximate analytical model of statistics distributions.

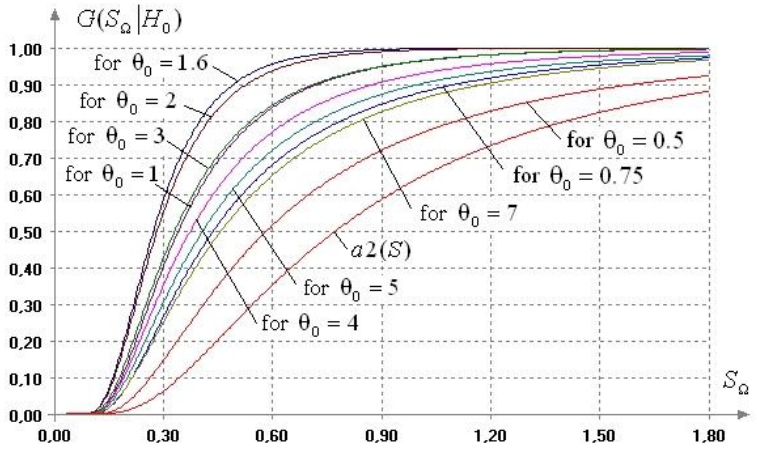

Figure 4: The distribution of Anderson-Darling statistics (3) for testing composite hypotheses using MLEs of 3 parameters of the generalized normal distribution, depending on the value of the shape parameter θ_0

The obtained models of limiting distributions and percentage points for Kuiper and Watson test statistics, which are required to test composite hypotheses (using MLEs) for the most often used in applications parametric distributions, listed in Table 1, could be found in the paper [50].

Previously obtained similar models (and percentage points) for distributions of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling test statistics (for distributions from Table 1) could be found in papers [43], [44], [45], [48], [49].

The tables of percentage points and models of test statistics distributions were based on simulated samples of the statistics with the size $N = 10^6$. Such N makes the difference between the actual distribution $G(S|H_0)$ and empirical counterpart $G_N(S|H_0)$ that does not exceed 10^{-3}. The values of the test statistic were calculated using samples of pseudorandom values simulated for the observed distribution $F(x, \theta)$ with the size $n = 10^3$. In such a case the distribution $G(S_n|H_0)$ practically equal to the limit one $G(S|H_0)$. The given models could be used for statistical analysis if the sample sizes $n > 25$.

Table 1: Random variable distributions

Random variable distribution	Density function $f(x,\theta)$	Random variable distribution	Density function $f(x,\theta)$
Exponential	$\frac{1}{\theta_0}e^{-x/\theta_0}$	Laplace	$\frac{1}{2\theta_0}e^{-\lvert x-\theta_1\rvert/\theta_0}$
Seminormal	$\frac{2}{\theta_0\sqrt{2\pi}}e^{-x^2/2\theta_0^2}$	Normal	$\frac{1}{\theta_0\sqrt{2\pi}}e^{-\frac{(x-\theta_1)^2}{2\theta_0^2}}$
Rayleigh	$\frac{x}{\theta_0^2}e^{-x^2/2\theta_0^2}$	Log-normal	$\frac{1}{x\theta_0\sqrt{2\pi}}e^{-(\ln x-\theta_1)^2/2\theta_0^2}$
Maxwell	$\frac{2x^2}{\theta_0^3\sqrt{2\pi}}e^{-x^2/2\theta_0^2}$	Cauchy	$\frac{\theta_0}{\pi[\theta_0^2+(x-\theta_1)^2]}$
Random variable distribution	Density function $f(x,\theta)$		
Logistic	$\frac{\pi}{\theta_0\sqrt{3}}\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\Big/\left[1+\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\right]^2$		
Extreme-value (maximum)	$\frac{1}{\theta_0}\exp\left\{-\frac{x-\theta_1}{\theta_0}-\exp\left(-\frac{x-\theta_1}{\theta_0}\right)\right\}$		
Extreme-value (minimum)	$\frac{1}{\theta_0}\exp\left\{\frac{x-\theta_1}{\theta_0}-\exp\left(\frac{x-\theta_1}{\theta_0}\right)\right\}$		
Weibull	$\frac{\theta_0 x^{\theta_0-1}}{\theta_1^{\theta_0}}\exp\left\{-\left(\frac{x}{\theta_1}\right)^{\theta_0}\right\}$		
Sb- Johnson $Sb(\theta_0,\theta_1,\theta_2,\theta_3)$	$\frac{\theta_1\theta_2}{(x-\theta_3)(\theta_2+\theta_3-x)}\exp\left\{-\frac{1}{2}\left[\theta_0-\theta_1\ln\frac{x-\theta_3}{\theta_2+\theta_3-x}\right]^2\right\}$		
Sl-Johnson $Sl(\theta_0,\theta_1,\theta_2,\theta_3)$	$\frac{\theta_1}{(x-\theta_3)\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\theta_0+\theta_1\ln\frac{x-\theta_3}{\theta_2}\right]^2\right\}$		
Su-Johnson $Su(\theta_0,\theta_1,\theta_2,\theta_3)$	$\frac{\theta_1}{\sqrt{2\pi}\sqrt{(x-\theta_3)^2+\theta_2^2}}\exp\left\{-\frac{1}{2}\left[\theta_0+\theta_1\ln\left\{\frac{x-\theta_3}{\theta_2}+\sqrt{\left(\frac{x-\theta_3}{\theta_2}\right)^2+1}\right\}\right]^2\right\}$		

Unfortunately, the dependence of the nonparametric goodness-of-fit tests statistics distributions for testing composite hypotheses on the values of the shape parameter (or parameters) (see Fig. 4) appears to be for many parametric distributions implemented in the most interesting applications, particularly in problems of survival and reliability. This is true for families of gamma, beta distributions of the 1st, 2nd and 3rd kind, generalized normal, generalized Weibull, inverse Gaussian distributions, and many others.

The limit distributions and percentage points for Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for testing composite hypotheses with the family of gamma distributions were obtained in paper [44], with the inverse Gaussian distribution — in papers [46], [47], with families of beta distributions — in paper [51], with generalized normal distribution — in paper [52], with the generalized Weibull distribution — in paper [53]. It should be noted that the data in these papers were obtained only for a limited number of, generally, integer values of the shape parameter (or parameters).

4 An interactive method to study distributions of statistics

The dependence of the test statistics distributions on the values of the shape parameter or parameters is the most serious difficulty that is faced while applying nonparametric goodness-of-fit criteria to test composite hypotheses in different applications.

Since estimates of the parameters are only known during the analysis, so the statistic distribution required to test the hypothesis could not be obtained in advance (before calculating estimates for the analyzed sample!). For criteria with statistics (6) — (8), the problem is harder as statistics distributions depend on the samples sizes. Therefore, statistics distributions of applied criteria should be obtained interactively during statistical analysis [54], and then should be used to make conclusions about composite hypothesis under test.

The implementation of such an interactive mode requires developed software that allows parallelizing the simulation process and taking available computing resources. While using parallel computing the time to obtain the required test statistic distribution $G_N(S_n|H_0)$ (with the required accuracy) and use it to calculate the achieved significance level $P\{S_n \geq S^*\}$, where S^* is the value of the statistic calculated using an original sample, is not very noticeable compared to a process of statistical analysis.

In the paper [55], an interactive method to research statistics distributions is implemented for the following nonparametric goodness-of-fit tests: Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling, Kuiper, Watson, and three Zhang tests. Moreover, the different methods of parameter estimation could be used there.

The following example demonstrates the accuracy of calculating the achieved significance level depending on sample size N of simulated interactively empirical statistics distributions [55].

Example. You should check the composite hypothesis that the following sample

with the size $n = 100$ has the inverse Gaussian distribution with the density (9):

0.945	1.040	0.239	0.382	0.398	0.946	1.248	1.437	0.286	0.987
2.009	0.319	0.498	0.694	0.340	1.289	0.316	1.839	0.432	0.705
0.371	0.668	0.421	1.267	0.466	0.311	0.466	0.967	1.031	0.477
0.322	1.656	1.745	0.786	0.253	1.260	0.145	3.032	0.329	0.645
0.374	0.236	2.081	1.198	0.692	0.599	0.811	0.274	1.311	0.534
1.048	1.411	1.052	1.051	4.682	0.111	1.201	0.375	0.373	3.694
0.426	0.675	3.150	0.424	1.422	3.058	1.579	0.436	1.167	0.445
0.463	0.759	1.598	2.270	0.884	0.448	0.858	0.310	0.431	0.919
0.796	0.415	0.143	0.805	0.827	0.161	8.028	0.149	2.396	2.514
1.027	0.775	0.240	2.745	0.885	0.672	0.810	0.144	0.125	1.621

$$f(x) = \frac{1}{\theta_2}\left(\frac{\theta_0}{2\pi\left(\frac{x-\theta_3}{\theta_2}\right)^3}\right)^{1/2} \exp\left(-\frac{\theta_0\left(\left(\frac{x-\theta_3}{\theta_2}\right)-\theta_1\right)^2}{2\theta_1^2\left(\frac{x-\theta_3}{\theta_2}\right)}\right). \quad (9)$$

The shift parameter θ_3 is assumed to be known and equal to 0.

The shape parameters θ_0, θ_1, and the scale parameter θ_2 are estimated using the sample. The MLEs calculated using the sample above are the following: $\hat{\theta}_0 = 0.7481$, $\hat{\theta}_1 = 0.7808$, $\hat{\theta}_2 = 1.3202$. Statistics distributions of nonparametric goodness-of-fit tests depend on the values of the shape parameters θ_0 and θ_1 [46, 47], do not depend on the value of the scale parameter θ_2 and can to be calculated using values $\theta_0 = 0.7481$, $\theta_1 = 0.7808$.

The calculated values of the statistics S_i^* for Kuiper, Watson, Zhang, Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling tests and achieved significance levels for these values $P\{S \geq S_i^*|H_0\}$ (p-values), obtained with different accuracy of simulation (with different sizes N of simulated samples of statistics) are given in Table 2.

Table 2: The achieved significance levels for different sizes N when testing goodness-of-fit with the inverse Gaussian distribution

The values of test statistics	$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 10^6$
$V_n^{mod} = 1.1113$	0.479	0.492	0.493	0.492
$U_n^2 = 0.05200$	0.467	0.479	0.483	0.482
$Z_A = 3.3043$	0.661	0.681	0.679	0.678
$Z_C = 4.7975$	0.751	0.776	0.777	0.776
$Z_K = 1.4164$	0.263	0.278	0.272	0.270
$K = 0.5919$	0.643	0.659	0.662	0.662
$KMS = 0.05387$	0.540	0.557	0.560	0.561
$AD = 0.3514$	0.529	0.549	0.548	0.547

The similar results for testing goodness-of-fit of a given sample with Γ-distribution with the density:

$$f(x) = \frac{\theta_1}{\theta_3 \Gamma(\theta_0)} \left(\frac{x - \theta_4}{\theta_3} \right)^{\theta_0 \theta_1 - 1} e^{-\left(\frac{x-\theta_4}{\theta_3}\right)^{\theta_1}}$$

are given in Table 3. The MLEs of the parameters are $\theta_0 = 2.4933$, $\theta_1 = 0.6065$, $\theta_2 = 0.1697$, $\theta_4 = 0.10308$. In this case the distribution of the test statistic depends on the values of the shape parameters θ_0 and θ_1.

Figure 5 presents the empirical distribution and the two theoretical ones (IG-distribution and Γ-distribution), obtained by the sample above while testing composite hypotheses.

Table 3: The achieved significance levels for different sizes N when testing goodness-of-fit with the Γ-distribution

The values of test statistics	$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 10^6$
$V_n^{mod} = 1.14855$	0.321	0.321	0.323	0.322
$U_n^2 = 0.057777$	0.271	0.265	0.267	0.269
$Z_A = 3.30999$	0.235	0.245	0.240	0.240
$Z_C = 4.26688$	0.512	0.557	0.559	0.559
$Z_K = 1.01942$	0.336	0.347	0.345	0.344
$K = 0.60265$	0.425	0.423	0.423	0.424
$KMS = 0.05831$	0.278	0.272	0.276	0.277
$AD = 0.39234$	0.234	0.238	0.238	0.237

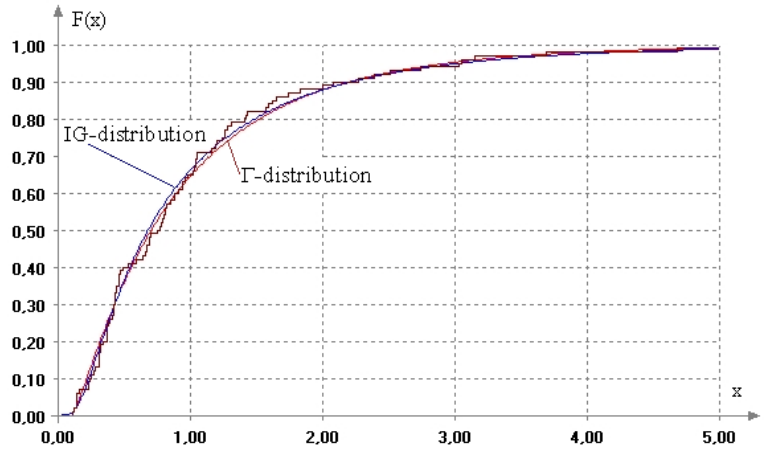

Figure 5: Empirical and theoretical distributions (IG-distribution and Γ-distribution), calculated using the given sample

Conclusion

The implemented interactive mode to study statistics distributions enables to correctly apply goodness-of-fit Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling,

Kuiper, Watson, Zhang (with statistics Z_C, Z_A, Z_K) tests with calculating the achieved significance level (p-value) even in those cases when the statistic distribution for true hypothesis H_0 is unknown while testing composite hypothesis. For Zhang tests, this method allows us to test a simple hypothesis for every sample size.

Acknowledgements

This research has been supported by the Russian Ministry of Education and Science as part of the state task (project 8.1274.2011) and the Federal Target Program "Research and scientific-pedagogical personnel of innovative Russia" (contract number 14.B37.21.0860).

References

[1] Kolmogoroff A.N. Sulla determinazione empirica di una legge di distribuzione // Giornale dell' Istituto Italiano degly Attuari. 1933. — Vol. 4. — No. 1. — P. 83–91.

[2] Bolshev L.N. Asymptotic Pearson transformations // Teor. Veroyatn. Ee Primen., 1963. — Vol. 8, — No. 2. — P. 129–155. (in Russian)

[3] Bolshev L.N., and Smirnov N.V. Tables for Mathematical Statistics [in Russian], Nauka, Moscow, 1983. — 416 p. (in Russian)

[4] Anderson T.W., and Darling D.A. Asymptotic theory of certain goodness of fit criteria based on stochastic processes // Ann. Math. Statist., 1952. V.23. — P.193–212.

[5] Anderson T.W., and Darling D.A. A test of goodness of fit // J. Amer. Stist. Assoc., 1954. V.29. — P.765–769.

[6] Kuiper N.H. Tests concerning random points on a circle // Proc. Konikl. Nederl. Akad. Van Wettenschappen. 1960. Series A, V. 63. P. 38–47.

[7] Stephens M.A. The goodness-of-fit statistic VN: distribution and significance points // Biometrika. 1965. V. 52. No. 3–4. P. 309–321.

[8] Stephens M.A. EDF Statistics for Goodness of Fit and Some Comparisons // Journal of the American Statistical Association, 1974. V. 69. No. 347. P. 730–737.

[9] Watson G.S. Goodness-of-fit tests on a circle. I. // Biometrika. 1961. V. 48. No. 1–2. P. 109–114.

[10] Watson G.S. Goodness-of-fit tests on a circle. II. // Biometrika. 1962. V. 49. No. 1–2. P. 57 — 63.

[11] Zhang J. Powerful goodness-of-fit tests based on the likelihood ratio // Journal of the Royal Statistical Society: Series B. 2002. V. 64. Part 2. P. 281–294.

[12] Zhang J., Wub Yu. Likelihood-ratio tests for normality // Computational Statistics & Data Analysis. 2005. V. 49. No. 3. P. 709–721.

[13] Zhang J. Powerful Two-Sample Tests Based on the Likelihood Ratio // Technometrics. 2006. V. 48. No. 1. P. 95–103.

[14] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N. The power of goodness of fit tests for close alternatives // Measurement Techniques. 2007. V. 50. No. 2. P. 132–141.

[15] Lemeshko B.Yu., Lemeshko S.B. and Postovalov S.N. Comparative Analysis of the Power of Goodness-of-Fit Tests for Near Competing Hypotheses. I. The Verification of Simple Hypotheses // Journal of Applied and Industrial Mathematics. 2009, V. 3. No. 4. P. 462–475.

[16] Lemeshko B.Yu., Lemeshko S.B. and Postovalov S.N. Comparative analysis of the power of goodness-of-fit tests for near competing hypotheses. II. Verification of complex hypotheses // Journal of Applied and Industrial Mathematics. 2010 V. 4. No. 1. P. 79–93.

[17] Lemeshko B.Yu. Asymptotically optimum grouping of observations in goodness-of-fit tests // Industrial laboratory (Ind. lab.). 1998, vol. 64, No. 1, pp. 59–67. (Consultants Bureau, New York)

[18] Lemeshko B.Yu., Gorbunova A.A. About application and power of nonparametric goodness-of-fit tests of Kuiper, Watson and Zhang // Izmeritel'naya Tekhnika. 2013. (in print)

[19] Kac M., Kiefer J., Wolfowitz J. On Tests of Normality and Other Tests of Goodness of Fit Based on Distance Methods // Ann. Math. Stat., 1955. V.26. — P.189–211.

[20] Darling D.A. The Cramer-Smirnov test in the parametric case // Ann. Math. Statist., 1955. V.26. — P.1–20.

[21] Darling D.A. The Cramer-Smirnov test in the parametric case // Ann. Math. Statist., 1957. V.28. — P.823–838.

[22] Lilliefors H.W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown // J. Am. Statist. Assoc., 1967. V.62. — P.399–402.

[23] Lilliefors H.W. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown // J. Am. Statist. Assoc., 1969. V.64. — P.387–389.

[24] Durbin J. Weak convergence of the sample distribution function when parameters are estimated. Ann. Statist., 1973. 1: 279–290.

[25] Durbin J. Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests of spacings // Biometrika, 1975. V.62. — P.5–22.

[26] Durbin J. Kolmogorov—Smirnov Test when Parameters are Estimated // Lect. Notes Math., 1976. V.566. — P.33—44.

[27] Martynov G.V. The Omega Squared Test. — Nauka, Moscow. 1978. — 80 p. (in Russian)

[28] Dzhaparidze K.O., Nikulin M.S. Probability distribution of the Kolmogorov and omega-square statistics for continuous distributions with shift and scale parameters // J. Soviet Math., 1982. V.20. — P.2147–2163.

[29] Nikulin M.S. Gihman and goodness-of-fit tests for grouped data. Mathematical Reports of the academy of Science of the Royal Society of Canada, 1992. 14 (4): 151–156.

[30] Nikulin M.S. A variant of the generalized omega-square statistic // J. Soviet Math., 1992. V.61 (4). — P.1896–1900.

[31] Pearson E.S., Hartley H.O. Biometrica Tables for Statistics. 1972. V.2. Cambridge: University Press.

[32] Stephens M.A. Use of Kolmogorov—Smirnov, Cramer — von Mises and Related Statistics — Without Extensive Table // J. R. Stat. Soc., 1970. V.32. — P.115–122.

[33] Stephens M.A. EDF Statistics for Goodness of Fit and Some Comparisons // J. Am. Statist. Assoc., 1974. V.69. — P.730–737.

[34] Chandra M., Singpurwalla N.D., Stephens M.A. Statistics for Test of Fit for the Extrem—Value and Weibull Distribution // J. Am. Statist. Assoc., 1981. V.76(375). — P.729–731.

[35] Tyurin Yu. N. On the limiting Kolmogorov—Smirnov statistic distribution for composite hypothesis // News of the AS USSR Series in Mathematics, 1984. Vol. 48. No 6. — P.1314—1343. (in Russian)

[36] Tyurin Yu.N., Savvushkina N.E. Goodness-of-fit test for Weibull—Gnedenko distribution // News of the AS USSR. Series in Technical Cybernetics, 1984. No. 3. — P.109–112. (in Russian)

[37] Lemeshko B.Yu., Postovalov S.N. Statistical distributions of nonparametric goodness-of-fit tests as estimated by the sample parameters of experimentally observed laws // Industrial laboratory (Ind. lab.). 1998, vol. 64, No. 3, pp. 197–208 (Consultants Bureau, New York)

[38] Lemeshko B.Yu., Postovalov S.N. Application of the nonparametric goodness-of-fit Tests in testing composite hypotheses // Optoelectronics, Instrumentation and Data Processing. 2001. — No. 2. — P. 76–88.

[39] Lemeshko B.Yu., Postovalov S.N., Frantsuzov A.V. Application of the nonparametric goodness-of-fit tests to testing nonparametric model adequacy // Optoelectronics, Instrumentation and Data Processing. 2002. — No. 2. — P.3–12.

[40] Lemeshko B.Yu., Maklakov A.A. Nonparametric Test in Testing Composite Hypotheses on Goodness of Fit Exponential Family Distributions // Optoelectronics, Instrumentation and Data Processing, 2004. Vol. 40, No. 3. — P.3–18.

[41] Lemeshko B. Yu. Errors when using nonparametric fitting criteria // Measurement Techniques, 2004. Vol. 47, No. 2. — P.134–142.

[42] Denisov V.I. Design of experiments and statistical analysis for grouped observations: Monograph / V.I. Denisov, K.-H. Eger, B.Yu. Lemeshko, E.B. Tsoy. — Novosibirsk: NSTU Publishing house, 2004. — 464 p.

[43] 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.

[44] 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.

[45] Lemeshko B.Yu., Lemeshko S.B. and Postovalov S.N. Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses // Communications in Statistics — Theory and Methods, 2010. Vol. 39, No. 3. — P. 460–471.

[46] Lemeshko B.Yu. Lemeshko S.B., Nikulin M.S., Saaidia N. Modeling statistic distributions for nonparametric goodness-of-fit criteria for testing complex hypotheses with respect to the inverse Gaussian law // Automation and Remote Control, 2010. Vol. 71, No. 7. — P. 1358–1373.

[47] Lemeshko B.Yu. Inverse Gaussian Model and Its Applications in Reliability and Survival Analysis / B.Yu. Lemeshko, S.B. Lemeshko, K.A. Akushkina, M.S. Nikulin, Noureddine Saaidia // In: Mathematical and Statistical Models and Methods in Reliability. Applications to Medicine, Finance, and Quality Control / Editors: V. Rykov, N. Balakrishnan, M. Nikulin / Series Statistics for Industry and Technology / Birkhuser, Boston. 2011. — P. 433–453.

[48] Lemeshko B.Yu., Lemeshko S.B. Construction of Statistic Distribution Models for Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses: The Computer Approach // Quality Technology & Quantitative Management, 2011. Vol. 8, No. 4. — P. 359–373.

[49] Lemeshko B.Yu., Statistical Data Analysis, Simulation and Study of Probability Regularities. Computer Approach : monograph / B.Yu. Lemeshko, S.B. Lemeshko, S.N. Postovalov, E.V. Chimitova. — Novosibirsk : NSTU Publisher, 2011. — 888 p. ("NSTU Monographs" series). (in Russian)

[50] Lemeshko B.Yu. About the solution of problems of application and power of some nonparametric goodness-of-fit tests / B.Yu. Lemeshko, A.A. Gorbunova, S.B. Lemeshko, A.P. Rogozhnikov // Avtometriya. 2013. No. 3. (in print)

[51] Lemeshko B.Yu., Lemeshko S.B. Statistic distributions of the nonparametric goodness-of-fit tests in testing hypotheses relative to beta-distributions // DAN VSh Rossii. 2007. No. 2(9). — P. 6–16. (in Russian)

[52] Lemeshko B.Yu., Lemeshko S.B. Models of Statistic Distributions of Nonparametric Goodness-of-Fit Tests in Composite Hypotheses Testing for Double Exponential Law Cases // Communications in Statistics — Theory and Methods, 2011. Vol. 40, No. 16. — P. 2879–2892.

[53] Akushkina K.A., Lemeshko S.B., Lemeshko B.Yu. Models of statistical distributions of nonparametric goodness-of-fit tests in testing composite hypotheses of the generalized Weibull distribution // Proceedings Third International Conference on Accelerated Life Testing, Reliability-based Analysis and Design. 19-21 May 2010, Clermont-Ferrand, France. — P.125–132.

[54] Lemeshko B.Yu., Lemeshko S.B., Rogozhnikov A.P. Real-Time Studying of Statistic Distributions of Non-Parametric Goodness-of-Fit Tests when Testing Complex Hypotheses // Proceedings of the International Workshop Applied Methods of Statistical Analysis. Simulations and Statistical Inference — AMSA'2011, Novosibirsk, Russia, 20-22 September, 2011. — P. 19–27.

[55] ISW — Program system of the statistical analysis of one-dimensional random variables. URL: http://www.ami.nstu.ru/~headrd/ISW_exe-r1006-2012.12.28-14_03.zip (address date 28.01.2013).

Reliability Approach in Statistical Modelling of Credit Risk

Vilijandas Bagdonavicius[1], Mikhail Nikulin[2] and Jia Shen[3]
[1] *Department of Statistics, University of Vilnius, Lithuania,*
[2] *Université Victor Segalen Bordeaux 2, Bordeaux, France,*
[3] *School of Management, Fudan University, Shanghai, China*

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 Horowitz (1998), Mosler (2002), Duffie and Singleton (2003), Kiefer (1988), Gouriéroux and Josiak (2005), etc... Dynamic regression models are applied often in reliability and survival analysis, see, for example, Bedford and Cooke (2001), Bagdonavicius and Nikulin (2002), Martinussen and Scheike (2006), Nikulin and Wu (2007), Nikulin, Gerville-Reache, Couallier (2007), Bagdonavicius, Kruopis and Nikulin (2011), 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, 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 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 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. 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) , Wu (2007), Nikulin and Wu (2007), Bagdonavicius and Nikulin (2002), Bagdonavicius, Kruopis and Nikulin (2011) . 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 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 $x(\cdot), x(\cdot) \in E$.

1 The Cox or the proportional default rate model

Under the *proportional default rate* model (traditionally PH model or Cox 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:

$$\mathbf{P}(T \le t + s \mid T > t) = 1 - e^{-\int_t^{t+s} e^{\beta^T x(u)} \lambda_0(u) du},$$

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.

2 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.

The AFT model is a good choice when the lifetime distribution class is supposed to be known. Nevertheless, 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.

3 Generalized proportional hazards model

a 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.

b 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.

c 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.

c.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*, $0 < \alpha < 1$.

c.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

$$p_Z(z) = \left(\frac{\sigma}{\pi}\right)^{1/2} e^{\sqrt{4\sigma\theta}} z^{-3/2} e^{-\theta z - \frac{\sigma}{z}}, \quad z > 0.$$

c.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.

d 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.

e 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.

f 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.

g 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 GHSS 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.

4 Models with time-dependent regression coefficients

a 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 PS model with time dependent regression coefficients of above given form can be written in the form

$$\lambda_{x(\cdot)}(t) = e^{\theta^T z(t)} \lambda_0(t). \tag{23}$$

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.

b 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

$$S_{x(\cdot)}(t) = S_0\left\{\int_0^t e^{-\beta^T(u)x(u)}du\right\}, \tag{24}$$

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

$$S_{x(\cdot)} = G\left\{\int_0^t e^{-\theta^T z(u)}du\right\}. \tag{24}$$

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.

5 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.

6 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

$$X_i = T_i \wedge C_i, \quad \delta_i = \mathbf{1}_{\{T_i \le C_i\}} \quad (i = 1, \cdots, n),$$

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

$$(N_1(t), Y_1(t), x_1(t), t \ge 0), \cdots, (N_n(t), Y_n(t), x_n(\cdot), t \ge 0),$$

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 $\theta \in \Theta \subset \mathbf{R}^{\mathbf{s}}$.

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:

$$U(\theta) = \frac{\partial}{\partial \theta} \ln L(\theta) = \sum_{i=1}^{n} \int_0^{\infty} \frac{\partial}{\partial \theta} \log \lambda_i(u, \theta) \{ dN_i(u) - Y_i(u) \lambda_i(u, \theta) du. \tag{32}$$

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

$$N_i(t) - \int_0^t Y_i(u) \lambda_i(u, \theta) du \tag{33}$$

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

$$U(\eta) = \sum_{i=1}^n \int_0^\infty \frac{\varphi^T(t) x^{(i)}(t)}{\lambda_i(t)} (dN_i(t) - Y_i(t)(x^{(i)}(t))^T d\Lambda_0(t)),$$

and the score functions for the parameters η and β (PPAR model) are:

$$U_1(\eta,\beta) = \sum_{i=1}^n \int_0^\infty \frac{\varphi^T(t) x_1^{(i)}}{\lambda_i(t)} (dN_i(t) - Y_i(t)(x_1^{(i)})^T d\Lambda_0(t) - \beta^T x_2 Y_i(t) dt) = 0,$$

$$U_2(\eta,\beta)=\sum_{i=1}^{n}\int_0^{\infty}\frac{x_2^{(i)}}{\lambda_i(t)}(dN_i(t)-Y_i(t)(x_1^{(i)})^T d\Lambda_0(t)-\beta^T x_2^{(i)}Y_i(t)dt)=0. \qquad (34)$$

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

$$g(x_i(s), \Lambda_0(s), 0 \le s \le t, \theta) = \left(1 + \gamma\left(\int_0^t e^{\beta^T x(u)} d\Lambda_0(u) + \delta^T \int_0^t x(u)du\right)\right)^{\frac{1}{\gamma}-1}.$$

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).

References

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.

Andersen, P.K. (1991). Survival analysis 1981-1991: The second decade of the proportional hazards regression model. *Statistics in Medicine*, **10**, # 12, 1931-1941.

Andersen, P.K., Borgan, O., Gill, R.D.& Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer.

Bagdonavičius, V. (1978) Testing the hyphothesis of the additive accumulation of damages. *Probab. Theory and its Appl.*, **23**, No. 2, 403-408.

Bagdonavičius, V. and Nikulin, M. (1994). Stochastic models of accelerated life , Advanced Topics in Stochastic Modelling (ed. J.Gutienez, M.Valderrama), *World Scient.*, Singapore.

Bagdonavičius, V. and Nikulin, M. (1999). Generalized Proportional Hazards Model Based on Modified Partial Likelihood, *Lifetime Data Analysis*, **5**, 329-350.

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.

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).

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).

Bagdonavičius V. and Nikulin M. (2002). *Accelerated Life Models: Modeling and Statistical Analysis.* Boca Raton: Chapman and Hall/CRC.

Bagdonavičius V., Kruops J., and Nikulin M. (2011). *Non-parametric tests for censored data.* London: ISTE & J.WILEY.

Bedford,T. and Cooke,R. (2001). Probability Risk Analysis. Fondation and Methods. Cambridge: Cambridge University Press.

Cox, D.R. (1972). Regression models and life tables, *J.R.Statist.Soc.*, B, **34**, 187-220.

Cox, D.R. (1975) Partial likelihood. *Biometrika*, **62**, 269-276.

Cox, D.R., and Oakes, D. (1984). Analysis of Suvival Data, *Methuen (Chapman and Hall)*, New York.

Dabrowska, D.M., Doksum, K.A. (1988). Partial likelihood in Transformations Models with Censored Data, *Scand. J. Statist.* **15**, 1-23.

Duffie,D. and Singleton,K.J. (2003). *Credit Risk. Pricing, Measurement, and Management.* Prinseton University Press: Prinseton and Oxford.

Gouriéroux, Ch. and Jasiak, J. (2005). Duration. In: A Companion to Theoretical Econometrics, (Ed. Badji H. Baltagi), Beijing: Peking University Press

Hougaard, P. (1986) Survival models for heterogeneous populations derived from stable distributions, *Biometrika*, **73**, 3, 387-396.
Hsieh, F. (2001). On heteroscedastic hazards regression models: theory and application. *Journal of the Royal Statistical Society,* Series B **63**, 63-79.
Kiefer, N. (1988). Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, Vol. 26, No.2. pp. 646-679.
Kleinbaum, D, (1996). *Survival Analysis: A Self-Learning text.* New York: Springer-Verlag.
Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis*, New York: Springer.
Lawless, J.F. (1982). Statistical Models and Methods for Lifetime Data, *Wiley*, New York.
Lin, D.Y., Geyer, C.J. (1992) Computational methods for semiparametric linear regression with censored data. *Journal Comput. and Graph. Statist.,*, **1**, 77-90.
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.
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.
McKeague, I.W., Sasieni, P.D.(1994) A partly parametric additive risk model. *Biometrika*, **81**,#3, 501-514.
Sedyakin, N.M. (1966) On one physical principle in reliability theory.(in russian). *Techn. Cybernetics*, **3**, 80-87.
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.

Whole Genome Sequence Coverage Estimation Re-examined

Alexey Antonik[1], Stephen J. O'Brien[1] and Sergey V. Malov[1,2]
[1] *Theodosius Dobzhansky Center for Genome Informatics, St.-Petersburg State University, St.-Petersburg, Russia*
[2] *St.-Petersburg Electrotechnical University, St.-Petersburg, Russia*
e-mail: alexey.antonik@gmail.com, lgdchief@gmail.com, malovs@sm14820.spb.edu

Abstract

Novel methods for estimating the depth of whole genome sequencing required for adequate coverage are presented. Employing the notion of r-spacing existent results for the uniform coverage were extended for the large scope of non-uniform distributions.

Keywords: coverage problem, spacings, uniform spacings.

Introduction

With next generation random shotgun sequencing and assembly of large eukaryotic genomes, it is important to develop a robust and accurate estimation *a priori* of the extent of fold coverage required to confidently sample the complete representation of a sequenced genome.

Next generation sequencers (NGS) make use of the short reads ($\sim$200 bp). To assemble the entire genome, available sequence read should include all or nearly all contiguous sequence of reads with overlaps. The last are needed for the correct overlapping of reads and also for the error detection, which may be efficient only if coverage reaches sufficient multiplicity, i. e. the every base pair is sequenced by not less than a sufficient number of reads.

Thus one needs to obtain a full genome coverage with an adequate multiplicity by reads of known minimum lengths (it also may be random, so its distribution is assumed to be known in that case) with overlaps, The question posed here is **how many reads of minimum length would be required to achieve coverage specified above with sufficiently large probability?**

The partial answer was estimated in the pioneer paper [2], where an efficient heuristic method was introduced. This result was extended in [4] employing the notion of uniform spacings, which was employed in multiple subsequent studies. The nonuniform coverage problem for the large scope of distributions was resolved in [1]. In this study we perform an extended analysis of the "Coverage Problem", expanding the notations and methods of the mentioned paper.

For the sake of consistency, we will remind the basic notations of [1]. The entire genome is represented by an interval $[0, 1]$. The reads are the subintervals of small fixed length l, $0 < l < 1$, and assumed to be random.

The problem of r-times covering of an interval $[0,1]$ can be expressed in the terms of maximal r-spacings as follows. Let $X_1, \ldots, X_n$ be i.i.d. random variables and n the number of reads. Denote by $X_{1,n} \leq \ldots \leq X_{n,n}$ the order statistics associated to $X_1, \ldots, X_n$ and introduce spacings $S_{0,n}^{(r)} = X_{r,n}$, $S_{i,n}^{(r)} = X_{i+r,n} - X_{r,n}$ for $i = 1, \ldots, n-r$, and $S_{n-r+1,n}^{(r)} = 1 - X_{n-r+1,n}$. Let $M_{n-r+2,n}^{(r)} \leq \ldots \leq M_{1,n}^{(r)}$ be their associated order statistics. Then the probability of r-times coverage of the interval $[0,1]$ by random segments of length l is $\mathbb{P}(M_{1,n}^{(r)} < l)$.

The last statement can be rewritten in the "scale" of genome. Let L be a length of single read in bp (base of pairs), N be a total length of genome, I be a minimal required overlap of two random reads assembled in the sequence. Then obviously $l = (L-I)/N$ and $\mathbb{P}(Q_r) = \mathbb{P}(M_{1,n}^{(r)} < (L-I)/N)$, where Q_r is an event in which the existing reads containing at least r bases from all positions, taking into account only reads with minimal intersection I bp with at least r neighbors.

1 Theoretical results

To summarize of results of [1], which are necessary for our further considerations, we stipulate:

Uniform distribution In this case we assume, that $X_i = U_i$, $i = 1, \ldots, n$, where $U_1, \ldots, U_n$ are i.i.d. standard uniform random variables. Desired probability can be efficiently estimated in the following way:

$$\mathbb{P}(nM_{1,n}^{(r)} - b_{n,x}^{(r)} \leq x) \geq e^{-e^{-x}}, \tag{1}$$

where $b_{n,x}^{(r)}$ can be obtained from the equation

$$b_{n,x}^{(r)} - \log n - \log\left(\sum_{i=1}^{r} \frac{(b_{n,x}^{(r)} + x)^{i-1}}{\Gamma(i)}\right) = 0.$$

Distributions with bounded support In this case we deal with a sample $X_1, \ldots, X_n$ from an absolutely continuous distributions with bounder support $[A,B]$, $-\infty < A < B < \infty$. For the sake of clarity we redefine $S_{0,n}^{(r)} = X_{r,n} - A$ and $S_{n-r+1,n} = B - X_{n-r+1,n}$. Fix $r \geq 1$. Let $\{f_{n,r}\}_{n\in\mathbb{N}}$ be a sequence of uniformly equicontinuous and nonnegative functions such that

$$\sup_{x>0} |\mathbb{P}(nM_{1,n}^{(r)} < x) - \exp(-nf_{n,r}(x))| \to 0 \quad \text{as} \quad n \to \infty.$$

Assume additionally that $X_1, \ldots, X_n$ have a PDF $p(x)$, satisfying $0 < \delta \leq p(x) \leq M < \infty$, $x \in [A,B]$ and the Hölder's condition piece-wise

$$|p(x) - p(y)| \leq C|x-y|^{\alpha} \quad \text{for all} \quad x, y \in T_i,\ i = 1, \ldots, s$$

with some C and $\alpha > 0$, and some intreval partition $\{T_i\}_{i=1}^s$ of the interval $[A, B]$. Then

$$\sup_{x>0} \left| \mathbb{P}(nM_{1,n}^{(r)} < x) - \exp\Big(-n \int_A^B f_{n,r}(xp(u))\, p(u)\, du\Big) \right| \to 0 \quad \text{as} \quad n \to \infty. \qquad (2)$$

Extended class of distributions Some limit results, see theorem 4.1 from [1], can be obtained for the classes of distributions related to the so called three extremal types: *Gumbel type*, *Fréchet type* and *Weibull type*. Not going into a meticulous formulation of the corresponding theorems and definitions, we only investigate their particular cases.

This results can be applied for some symmetric distributions with $p(1_-) = p(0_+) = 0$. For instance, the trapezoidal distribution having PDF

$$p(x) = \begin{cases} (\kappa(1-\kappa))^{-1} x, & x \in [0, \kappa); \\ (1-\kappa)^{-1}, & x \in [\kappa, (1-\kappa)); \\ (\kappa(1-\kappa))^{-1}(1-x), & x \in [1-\kappa, 1]; \\ 0, & x \notin [0, 1], \end{cases}$$

with some $\kappa \in [0, 1/2]$ belongs to Weibull's extremal type with $a = 2$ (see [3], Theorem 1.6.1). Then, by [1], Corollary 4.1 (iii),

$$\mathbb{P}(M_{k,n}^{(r)} \le l) \approx \big(H_{1,r}^W\big(l\sqrt{n}/\sqrt{2\kappa(1-\kappa)}\,\big)\big)^2,$$

where $H_{1,r}^W$ is a CDF of $\max_{j\ge 0}\big\{(\sum_{s=1}^{r+j} E_s)^{1/2} - (\sum_{s=1}^{j} E_s)^{1/2}\big\}$, and E_1, $E_2, \ldots$ is a sequence of i.i.d. random variables having the standard exponential $E(1)$ distribution. Despite the fact this CDF can't be easily obtained exactly, it can be efficiently estimated via empiric simulations.

By [1], Lemma 6.3, the same approximation is valid for any PDF having the same behavior near bounds 0 and 1 and separated from zero in other points of the interval $[0, 1]$.

Random read lengths The important extension to the coverage problem gives a permission for reads to have random lengths.

Introduce a sample $Y_1, \ldots, Y_n$ from a positive distribution with CDF F_Y and independent of $X_1, \ldots, X_n$. Denote by $N_{n,r}(x)$ the number of of r-spacings greater than $x\, Y_i$, *i.e.* $N_{n,r}(x) = \#\{i \in \{1, \ldots, n^*\} : S_{i,n}^{(r)} \ge x\, Y_i\}$.

Under the independence assumption of the original sample of covering segments left ends $X_1, \ldots, X_n$ and the corresponding sample of their length $Y_1, \ldots, Y_n$ the r-times coverage probability can be expressed as $\mathbb{P}(M_{1,n}^{(r)} < Y)$, where Y has the distribution of random length of the small segments concentrated on $[0, 1]$ and independent of $M_{1,n}^{(r)}$. As a result the probability of r-times coverage of whole interval $[0, 1]$ is

$$\mathbb{P}(M_{1,n}^{(r)} < Y) = \int_0^1 \mathbb{P}(M_{1,n}^{(r)} < s)\, dF_Y(s). \qquad (3)$$

The desired probability can be obtained utilizing estimates and approximative formulas for $M_{1,n}^{(r)}$. Similarly, the probability to have less then k regions without r-times coverage is

$$\mathbb{P}(N_{n,r} < k) = \mathbb{P}(M_{k,n}^{(r)} < Y) = \int_0^1 \mathbb{P}(M_{k,n}^{(r)} < s)\, dF_Y(s).$$

2 Simulations and Applications

In this section we address the question stated in the Introduction: "how many reads do we need to obtain coverage with sufficiently large probability?". To assess numeric results of the methods described above we performed simulations via R Project software [5].

In the tables 1–2 we give a ratios between a total (expected) length of reads and whole genome length required for full r-coverage of Human's genome with 95% probability. In our computations we assume $N = 3.2 \cdot 10^9$ and $I = 50$.

Table 1: Total (expected) length of reads divided by whole genome length: uniform distribution

Distribution	Uniform					
L	100	150	200	250	300	Random
$r = 1$	48	35	31	28	27	27
$r = 2$	55	40	35	33	31	31
$r = 5$	72	52	46	43	41	41
$r = 10$	94	69	61	56	54	54
$r = 25$	148	110	97	90	86	86
$r = 50$	227	168	149	139	133	133

Table 2: Total (expected) length of reads divided by whole genome length: truncated normal distributions

Distribution	Truncated $N(1/2, 1)$				Truncated $N(1/2, 1/4)$			
L	100	200	300	Random	100	200	300	Random
$r = 1$	49	31	27	27	173	109	95	95
$r = 2$	56	36	32	32	201	127	112	112
$r = 5$	73	47	42	42	268	171	151	151
$r = 10$	96	62	55	55	359	231	204	204
$r = 25$	154	100	89	89	586	381	338	338
$r = 50$	237	155	138	138	916	599	534	534

To fulfill table for the uniform case we used estimator (1). Truncated normal distribution was estimated throughout formula (2). Formula (3) was applied in the case of random length with $Y \sim N(300, 50^2)$.

The results in the table demonstrate how much the total length of reads should be greater then the length of genome. It is important to note that 'random' results are equal to the corresponding 'non-random' ones due to the considerable length of the entire genome.

Acknowledgements

This research was supported in part by Russian Ministry of Education and Science Mega-grant no.11.G34.31.0068 (Dr. Stephen J. O'Brien, Principal Investigator).

References

[1] Antonik, A., Berred, A., Malov, S. V. (2013). Coverage Problem revisited, *The Annals of Applied Probability*, Submitted.

[2] Lander, E.S. & Waterman, M.S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231 – 239.

[3] Leadbetter, M.R., Lindgren, G. & Rotzen, H. (1883) *Extremes and related properties of random sequences and processes*. Springer-Verlag, New York Heidelberg Berlin, 1983.

[4] Roach, D.C. (1995). Random subcloning. *Genome Research* **5**, 464 – 473.

[5] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Mathematical Model of the Residual Lifetime of NPP Equipment Calculation based on Operational Information Specific Type

Alexander Antonov, Sergey Sokolov, Valeriy Chepurko
Institute of Nuclear Power Engineering
Obninsk, Russia
e-mail: antonov@iate.obninsk.ru

Abstract

Probabilistic estimation method of the average straight residual lifetime for nuclear power plants (NPPs) systems and their constituent elements is considered. The mathematical model for calculating of this reliability characteristic for the objects to be recovered from the initial data on failures censored interval is presented. Besides, the issue of its accuracy estimating using the bootstrap method is considered.

Keywords: residual lifetime, system, element, reliability characteristic, operational data.

Introduction

Currently, increased attention of researchers in the reliability theory is given to the analysis of technical object operation subject to the aging. This problem is particularly relevant in the nuclear power industry. The most of power capacities in nuclear power engineering were put into operation in the 70ies-80ies. Today operating organizations are oriented to extend the assigned lifetime of the NPPs, but for the reasonable prolongation of the lifetime of power units as a whole and their individual components, systems, parts, etc. reliability analysis of all the constitu-ent elements, assemblies and systems is required.

In this paper the problem of estimating the residual operating time between failures of renewal objects is solved. It is assumed that the strategy of maintenance facilities include monitoring for proper operation of functioning, as well as routine preventive and emergency repairs. The proposed method for estimating the residual operating time is based on building a stochastic model, which is mathematically described by the Voltaire integral equation. One of the problems that appear during the calculation of systems reliability characteristics is the problem of determining the reliability of elements included in the structure of the system using operational failure information. During the NPPs reliability characteristics calculations the reliability characteristics is assessed as well as confidence estimation is required. It means that the problem becomes to the task of assessing the accuracy of the calculated parameters.

1 Problem Definition

One of the conditions for the extension of equipment assigned lifetime is the substantiation of its reliability. The present article is concerned with computational methods for determining the reliability of renewal facilities, when repairs are possible and provided by regulatory, technical, repair and design documentation. In the capacity of determinants of reliability a residual operating time of the object was selected by the authors. A residual operating time is an operating time of the object from the beginning of the operation or installation into the system to the recovery to date of failure.

The general practice of calculating the reliability characteristics of renewal systems is based on the application of the mathematical methods of the renewal theory under the assumption that the system renewal time is small in comparison with the normal operating time value and it can be neglected. As a result, in the capacity of the computational model the model of regenerative processes is used, which suggests that during the repairs carried out a complete restoration of all the original properties of the system. This model describes well the practical situation when the renewal of the system in operative condition after failure means the replacement of the failed element by the same type element from repair kits or spare parts.

However, even with a simple model calculation of the residual operating time is a very complicated mathematical problem, an analytic solution of which can be obtained only in special cases with a parametric specification of the original data. In addition, nuclear power has a large number of serviced systems and equipment, the renewal time of which has the same order of magnitude that the operating time to failure. During the operation diagnostic tests organized in a special way are carried out, spare equipment sets are created that is installed into the system in case of failure. Repair system including the current plan, secondary and capital repairs is organized and planned. All these facts lead to the conclusion that the renewal time can not be neglected. In this case, it is necessary to use the theory of alternating processes for the description of models of equipment functioning.

Thus, the task of this paper is the construction of adequate and reliable models of the residual lifetime estimating taking into account the different operation strategies, service activities.

2 The Estimation of Straight Residual Time for Renewal Objects

a The Strategy of the System Operation with a Built-in Monitoring Efficiency and Low Renewal Time

In modern technical systems different devices of the equipment efficiency monitoring are used. Early detection of failures and defects allows carrying out procedures for their elimination and, consequently, exploit the technical facilities more effectively.

The system which is characterized by the presence of elements with faultiness control is going to be considered in this paper. In case of failure of a subsystem the operation staff immediately become aware of the failure element (e.g., the alarm indication goes off). Monitoring of performance is carried out constantly.

Lets consider the model of the objects operation, which has a built-in test system. In the case of failure the system becomes inoperable and the emergency maintenance work to renew functionality shall be started.

Lets suppose that in initial time $t_0 = 0$ the object is in working condition. The system operates until failure τ_i. Built-in monitoring system instantly and unambiguously provides infor-mation about the place of failure to service staff, and system renewals for a negligible time. After restoring the system continues to operate until the next failure. The cycle of such states changing is repeated until a certain time t. Denote time to failure at the i-th operating cycle ξ_i. Described strategy of functioning is shown in Figure 1.

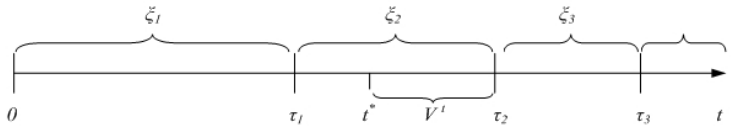

Figure 1: The strategy of the system with failure indication and a small recovery time

In [1] a process $\{V_i^t, t{\geq}0, i = 1, 2, ...\}$ called the straight residual time process is described, where

$$V_i^t = \tau_{i+1} - t. \tag{1}$$

It should be noted that V^t is the straight residual time, or the residual operating time of system at time t.

Also in [1] showed that the average straight residual time can be defined as

$$MV^t(t) = (H(t) + 1) \int_0^\infty u f_\xi(u) du - t. \tag{2}$$

where $f_\xi(t)$ is a failure density function, $H(t)$ renewal function, which is determined by solving of the equation

$$H(t) = F_\xi(t) + \int_0^t H(t - u) dF_\xi(u).$$

However, in practice this analytical solution for the straight residual time is quite difficult to be use even in special cases, because it is not always possible to calc an estimation of the renewal function.

Its possible to find the average straight residual lifetime $MV^t(t)$ using the definition of the mathematical expectation of the time remaining until the next system

failure, starting at time t in which the system was operable. According to this definition

$$MV^t(t) = M\sum_{i=0}^{\infty}(\tau_i - t)\cdot I\{\tau_i \leq t < \tau_{i+1}\}, \tag{3}$$

where τ_i - the failure time. Then

$$MV^t(t) = \sum_{i=0}^{\infty} M(\tau_i - t)\cdot I\{\tau_i \leq t < \tau_{i+1}\} = \sum_{i=0}^{\infty}\psi_i(t). \tag{4}$$

Write down the expression under the summation sign

$$\psi_i(t) = \int_0^{\infty}\int_0^{\infty}(s + x - t)\cdot I\{s \leq t < s + x\} f_{\tau_i}(s) f_{\xi}(x) dx ds = \int_0^{\infty} x\varphi_i(t; x) dx \tag{5}$$

Lets make the Laplace transform of the inner integral $\varphi_i(t; x)$ and obtain:

$$\overline{\varphi}_i(p; x) = \overline{f}_{\tau_i}(p)\cdot\overline{g}(p; x) = (\overline{f}_{\xi}(p))^i\cdot\overline{g}(p; x), \tag{6}$$

where $\overline{g}(p; x)$ the image of the function $g(t; x) = f_{\xi}(t+x)$. Then the Laplace function of average straight residual lifetime $\overline{MV^t(p)}$ will be determined by the expression

$$\overline{MV^t}(p) = \frac{1}{1 - \overline{f}_{\xi}(p)}\cdot\int_0^{\infty} x\overline{g}(p; x) dx.$$

Turning to the originals, it obtains the Voltaire integral equation

$$MV^t(t) = \int_0^{\infty} x f_{\xi}(t + x) dx + \int_0^t MV^t(u) f_{\xi}(t - u) du. \tag{7}$$

Its solving allows estimating the value of the average straight residual lifetime.

b Calculation of the Characteristics Used in the Equation for Average Straight Residual Time

In order to make calculations of reliability characteristics including average straight residual time it is necessary to know the density function of operating time to the i-th failure and renewal time. It should be note that information obtained from operating experience should be used to estimate the density functions. It is important for the described method that failure times of devices are unknown when the collection of data on nuclear power systems (NPS) equipment failures is performed. There is only data about the number of failures of the same-type elements, distributed at intervals of efficiency. As the range of efficiency a calendar year is considered. In other words, failures are grouped by the operating year and only the facts of failures are known. Based on the analysis of such statistics it is quite difficult to determine the distribution of failure time. In order to renew the density of failure time the method of kernel estimates is used.

Consider the observation period for the object operation as an array of observation time intervals

$$\overrightarrow{LR} = [(l_1, r_1); (l_2, r_2); ...; (l_s, r_s)]$$

, where the random number of failures

$$\overrightarrow{\nu} = [\nu_1, \nu_2, ..., \nu_s]$$

has taken place. Note that the intervals are disjoint and the right border of the considered interval is equal to the left border of the subsequent interval $r_i = l_{i+1}$.

Lets consider that n is the total number of failures, m the number of similar objects forming this failure flow. Suppose that in case of failure the failed element is replaced by another analog with the same characteristics. In this case, there is a complete renewal of the system. Consequently, the failure flow parameter $\omega(t)$ can be determined. If there are data about failures censored intervals for the failure flow parameter following kernel estimation is obtained

$$\widehat{\omega}(t) = \sum_{i=1}^{s} \frac{\nu_i}{m(r_i - l_i)} \left(G(\frac{t - l_i}{h}) - G(\frac{t - r_i}{h} \right) + \varepsilon(t), \tag{8}$$

where

$$G(x) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{x} exp\left(-\frac{u^2}{2}\right) du$$

- Gaussian kernel; h is the locality parameter (the measure that depends on the standard deviation of the failure time); $\varepsilon(t)$ is the estimated systematic errors of the failure flow parameter which should be obtain as

$$\varepsilon(t) \approx \frac{1}{2a} \left[erfc\left(\frac{an/m - t}{\sqrt{2n\sigma^2/m}}\right) + exp\left(\frac{2at}{\sigma^2}\right) \cdot erfc\left(\frac{an/m + t}{\sqrt{2n\sigma^2/m}}\right) \right].$$

As it is known from renewal theory, the failure flow parameter is related with the density distribution of failure time through the Voltaire integral equation

$$f_\xi(t) = \omega(t) - \int_0^t f(\tau)\omega(t - \tau)d\tau. \tag{9}$$

Thus, having sufficient statistical data, it is possible to estimate the density distribution of the failure time, solving the equation (9), and then estimate the mathematical expectation of straight residual lifetime (7). Lets consider the example of calculation. Suppose it is known that the system consists of $m = 4$ the same elements. The vector of failures is

$$\nu = (1, 9, 3, 4, 3, 2, 1, 0, 3, 0, 3, 0, 0, 0, 0, 1, 0, 6, 1, 0, 1, 2, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0). \tag{10}$$

Data are grouped by operating years, the total number of failures amounted to $n = 42$ for $s = 34$ years. Using formula (8) for these elements lets estimate the failure flow

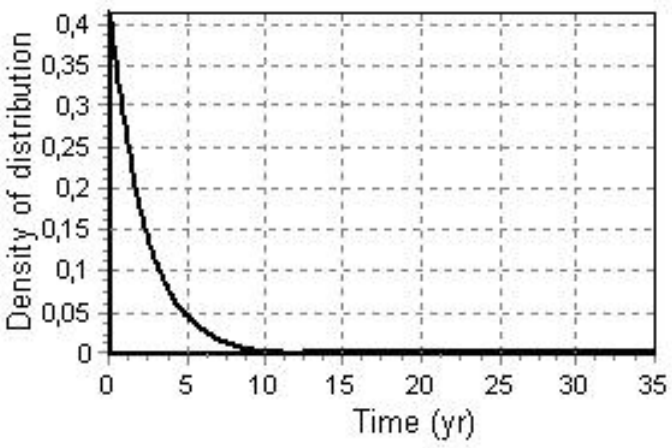

Figure 2: The kernel estimation of failure time density

parameter and then solving the equation (9), define the density of failure. The result of the density distribution calculation is shown in Figure 2.

Now the mathematical expectation of straight residual lifetime can be estimated using formula (7). Solving the equation (7) and taking into account these initial data, the following estimate of the average straight residual lifetime for the element can be obtained (Figure 3).

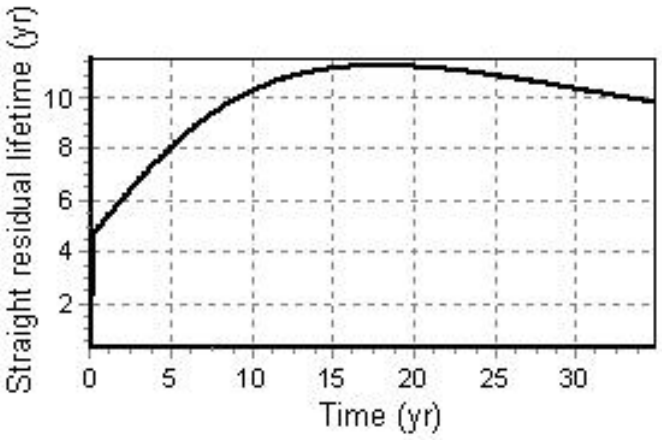

Figure 3: The mathematical expectation of average straight residual time

Thus, the presented methodology allows to obtain the estimation of the straight residual lifetime and to predict the residual lifetime of the technical objects.

3 The Estimations Accuracy Determination for the Average Straight Residual Lifetime

Determining the accuracy estimation of the mathematical expectation of straight residual lifetime is a quite difficult task in the set of initial conditions. In the case of nonparametric estimation for the moment it is not yet decided how to obtain estimation for the variance of the straight residual time. Therefore, the authors suggest carrying out estimation of the accuracy using bootstrap method.

Bootstrap method was described in [2], and its essence is that one-sample statistics of observations is transformed into many samples with the same sample size. The transformation is carried out on basis of the primary sample and taking into account its distribution law.

Thus, the main idea of the bootstrap method is in the multiplication of the available data. The task is to simulate random samples with the same size as the primary sample. In addition, each simulated sample is generated by random selection with the returning of one of the events from the primary sample. This procedure allows to build the sampling distribution of the estimated feature without any additional assumptions and to make nonparametric confidence intervals. Lets explain the essence of the accuracy estimating of the bootstrap method applied to the sample paragraph 2.2, where a vector of element failures grouped by operating years is defined.

In the contrast to the classical application of the bootstrap method, in our case, the available data are grouped by operating year and failure times are unknown. Therefore, if in case of the classical bootstrap method implementation a random uniformly distributed variable is played on the axis of the probability (interval $[0, 1]$) and then it is projected to the axis of failure times, but in our case it is necessary to map the simulated random uniformly distributed variable on the axis of failures events implementation.

Lets consider the sequence of action in determining the accuracy of the estimated feature with bootstrap method.

Step 1. Each failure event is assigned to an ordinal index i and determine to the observation time interval when happened. There are n failures distributed in k observation time intervals. For our example ($n = 42, k = 34$), there are 1st failure in 1st observation time interval, 2nd-10th failures in 2nd interval, 11th-13th failures in 3rd interval, etc.

Step 2. The axis of the probability is divided to n equal disjoint intervals $[0, y_1)$, $[y_1, y_2)$, ..., $[y_{n-1}, y_n]$. Simulate a random uniformly distributed variable $U[0, 1]$ on the axis of the probability. Determine which of the n observation time intervals contains this variable. If $[y_{i-1}, y_i)$, then it means that the event with index i is realized. Repeat the operation of modeling n times, thereby a sample of the event numbers of the failures is built. Finding events should be assigned to observation time intervals according to the partition, in step 1. Using the obtained bootstrap sample as input data, lets calculate the expectation of straight residual lifetime MV_1^t.

Step 3. Repeat step 2 many times independently. Thus a lot of ratings MV_i^t are got.

Step 4. For nonparametric estimation performance limits of the confidence intervals are defined as follows. First, set the significance level α in accordance with the confidence level of $1-2\alpha$. Secondly, define the boundaries of intervals that satisfy the following relations for the given α.

$$\alpha = \frac{d(MV_i^t \leq MV_{low}^t)}{r}; \tag{11}$$

$$1-\alpha = \frac{d(MV_i^t \leq MV_{high}^t)}{r}; \tag{12}$$

where r the amount of bootstrap repetitions; $d(MV_i^t \leq MV_{high}^t)$ the number of parameter MV^t bootstrap repetitions which took values less than MV_{high}^t. In this case, evaluation MV_{low}^t and MV_{high}^t defined by expressions (10) and (11) will characterize the approximate confidence interval, corresponding to a confidence probability $1-2\alpha$. The results of MV_{low}^t and MV_{high}^t calculations according to initial data of the represented example and formulas (10) and (11) are shown in figure 4.

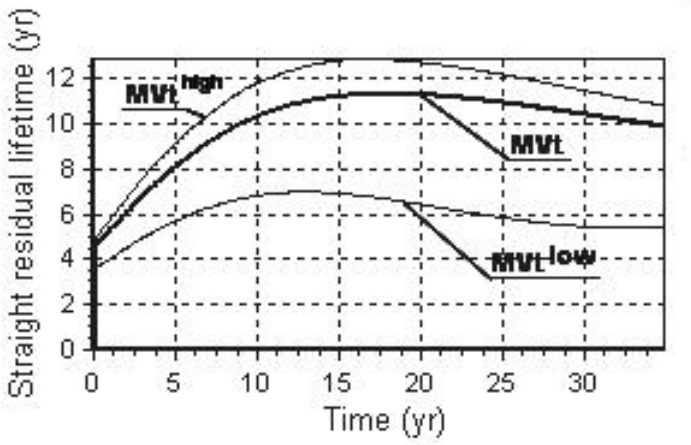

Figure 4: The construction of confidence intervals for estimating of average straight residual lifetime

The advantage of the represented method is the possibility to build the confidence interval for estimation besides the estimation of the reliability features on the basis of initial censored sample of small size.

Conclusions

In this paper the method of estimating the average straight residual lifetime is considered and the algorithm for estimating the calculations accuracy is described. The distinctive feature of the presented method is the possibility of using non-parametric methods of estimation. The considered method allows carrying out practical research, taking into account the quality of available basic statistical data. This method can

be used to estimate the reliability characteristics of systems with complex service strategies.

References

[1] Beichelt E., Franken P. (1983). *Zuverlassigkeit und Instandhaltung. Mathematische Methoden.*

[2] Efron B. (1976). Censored data and bootstrap. *J.A.S.A.*. pp. 312-319.

Pattern Recognition by the Monte Carlo Method

Antyufeev V.S.[1]
[1] *Institute of Computational Mathematics and Mathematical Geophysics SD RAS, Novosibirsk State University*
Novosibirsk, Russia
e-mail: ant@osmf.sscc.ru

Abstract

A new method for solving the pattern recognition problems is proposed. Unlike artificial neural networks it does not utilize concepts of an the artificial neuron and the neural network. The Monte Carlo method is applied for modeling the training signals. A numerical example is considered.

Keywords: artificial neural networks, fuzzy algorithm, Monte Carlo method.

Introduction

The pattern recognition theory studies methods of classification/identification of objects with a finite number of numerical parameters. To solve these problems artificial neural networks (ANN) were proposed. ANN's replaced humans in many tasks. The simplest networks (perceptrons) were able to solve simple problems; the training theorem was proved for perceptrons [2]. But perceptrons were unable to solve some important problems. Then powerful and complicated multilayer ANN's were developed. But theoretical study of these ANN's is too difficult; the training theorem has not been proved for them and there are still theoretical/practical difficulties and unsolved issues related to them.

A new method is introduced in this paper which can replace ANN. It does not use concepts of the neuron, the network, the layer. Its work is based on another principle. Methods of mathematical analysis and probability theory can be applied to study it. As a result, the training theorem was formulated and proved.

1 Brief description of the work of ANN

Here we briefly consider the main phases in the work of ANN.
1) First, the ANN works in the training mode. A long sequence of training signals $z = (z_1, \ldots, z_n)$ is supplied to the input of the ANN. Coordinates z_k are properties of the studied object. For example, these may be results of a medical test of a patient: the blood pressure, body temperature and so on. The so-called ≪teacher≫ gives correct answers to a stated question (the main question) for each training signal. Using these answers ANN adjusts its weights w_k.
2) When the training process is over, the ANN resumes the work in an operating mode. Working signals z with unknown answers are supplied to the input, and the ANN performs two successive operations:
2.1) $z \Longrightarrow net = w \cdot z$ (w is the weight vector),

2.2) $net \Longrightarrow out = F(net)$ ($F : R \to R$ is the activation function [3]),
Finally, ANN produces the answer:
2.3) If $out > 1/2$ we assume that the ANN puts out 0 and the object under examination is related to the first group, otherwise it is related to the second group (the patient can be attributed to the sick group or to the healthy one).

2 Geometric description of the ANN

Let ANN be in the operating mode and $z = (z_1, \ldots, z_n)$ be a signal. ANN checks whether the components z_i satisfy the system of inequalities (1) (each inequality has the form (2)) [3]:

$$\{\text{inequality 1}\} \;\&\; \{\text{inequality 2}\} \;\&\; \{\text{inequality 3}\} \vee \{\text{inequality 4}\} \ldots, \qquad (1)$$

$$\sum_j w_{mj} z_j < 1 \qquad \sum_j w_{mj} z_j > 1. \qquad (2)$$

(the order of the logical signs &, $\vee$ is determined by the complex topology of the graph of the ANN). From geometric viewpoint, each inequality (2) determines a half-

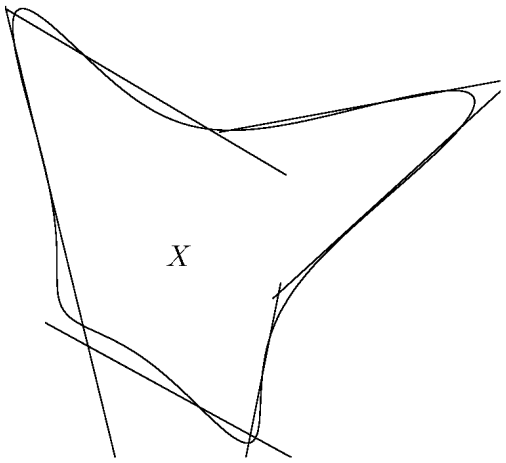

Figure 1: Approximation of the required set X by the polyhedron S_k.

space in R^n bounded by a hyperplane. The whole system (1) determines a polyhedron S_k (possibly nonconvex), or a union of separated polyhedra. The signal z satisfies system of inequalities (1) if and only if the point $z \in S_k$. Thus, the geometric event $\{\ z \in S_k\ \}$ is equivalent to the logical event $\{$ the numbers $z_i,\ i = 1, \ldots, n$ satisfy the system of inequalities (1) $\}$.

Denote by X ($X \subset R^n$) the set of signal points corresponding to the positive answer 1. Solving problem for ANN means finding the set X. Therefore, we may call

the sets S_k and X approximate and ideal correspondingly. At each step, the ANN determines the approximate set S_k corresponding to the weights, and thus forms the sequence of sets $S_1, 2, \ldots$. In terms of the set theory, the learning capability of the ANN means that the sequence $S_1, S_2, \ldots$ converges in a way to the ideal set X (see. Fig.1). We give an exact definition of this convergence below. From the geometric viewpoint, the process of ANN training may be reduced to approximating the ideal set X by the polyhedra S_k (each S_k is a composition – unions, intersections, or supplements – of half-spaces).

If the set X is not a union of polyhedra, then certainly $S_k \neq X$ for any k and a finite number of training steps cannot produce the exact solution to the problem.

The perceptrons can form only simply connected convex sets. Therefore, many problems cannot be solved using perceptrons. Multilayer networks only are able to construct complex sets for solving complicated problems. Note that the nonconvexity and disconnectedness of the ideal set X do not make any obstacle to solving problems by the proposed method.

3 Description of ANN in terms of functions

Describe the work of an ANN in the language of functions. ANN can be identified with a function ANN : $R^n \to \{0;\ 1\}$ which is defined by the logical algorithm of the ANN and the weights w_k are its parameters: $\mathrm{ANN}(z) \equiv \mathrm{ANN}_w(z)$. In the same way we can define a fuzzy neural network [3]. In this case the value domain of the function f is the whole segment $[0, 1]$ (not the two numbers 0, 1): ANN : $R^n \to [0,\ 1]$. There is one-to-one correspondence between the sets $S \subset R^n$ and their characteristic functions $\chi_S(z),\ z \in R^n$:

$$\chi_S(z) = \begin{cases} 1, & z \in S, \\ 0, & z \notin S \end{cases}$$

We have defined X, S_k as the ideal and approximate sets. Now define the ideal decision function $f(z) \equiv \chi_X(z)$ and approximate decision functions $f_k(z) \equiv \chi_{S_k}(z)$. In terms of 01, the correct answer to the signal z is equal to $\chi_X(z)$. From the viewpoint of of functions, the ANN training process is the process of approximating the ideal decision function $f(z)$ by approximate decision functions $f_k(z)$. Now we can give a rigorous definition of the convergence $S_k \to X$. This limiting relation is equivalent to $\lim\limits_{k\to\infty} \chi_{S_k} = \chi_X$

4 The new decision algorithm of competition

In this section we introduce some auxiliary functions and finally the new algorithm.

a Requirements for approximate decision functions f_k

First, We list some natural conditions for the functions f_k and then construct functions satisfying those conditions.

(1) The new algorithm is fuzzy and its action in the working mode can be identified with the mapping $f: \ R^n \to [0;\ 1]$. The value domain of f is the segment $[0;\ 1]$, the number $f(z)$ gives the empirical probability of the event $z \in X$ after the whole training course. Similarly, $f_k(z)$ gives the same probability after the k-th training step. So, the value domain of the function f_k is the segment $[0; 1]$.
(2) Consider the new algorithm in the training mode. Let $z^1 = (z_1^1, \dots z_n^1)$ be the first input training signal. Let the point $z^1 \in X$. Then $f_1(z^1) = 1$. Now, if a signal z falls into a small neighbourhood of the point z^1, then the probability of the event $z \in X$ should be close to 1. Similar rules with the corresponding corrections should hold in the case $z^1 \notin X$ as well. This gives the requirement for the continuity of f_k.
(3) Each function f_k must be constructed according to all previous training signals $z^1, z^2, \dots, z^k$. Its values at the points $z^1, z^2, \dots, z^k$ must remain equal to 1 or 0 when subsequent training signals come in.
(4) The same training signals $z^1, z^2, \dots, z^N$ may come to the input in a different order. Obviously, the result of this training , i.e. the function f_N, should not depend on the order of the incoming signals.
Now we briefly formulate these requirements:

- $0 \le f_k(z) \le 1$.
- The functions f_k must be continuous;
- $f_k(z^i) = \begin{cases} 1, & z^i \in X \\ 0, & z^i \notin X \end{cases}, \quad i = 1, \dots, k$
- the functions f_k do not depend on the incoming order of training signals.

Constructing functions of a simple form satisfying thess conditions is the main goal of the paper.

b Influence function

Let Z be a set where all possible signals lie. Divide the set Z into the subset X, Y of the positive and the negative signals: $X \cup Y = Z, \ X \cap Y = \emptyset$.
Let $z^1 = (z_1^1, \dots, z_n^1)$ be the first training signal and, as the teacher told, $z^1 \in X$. We are absolutely sure (our confidence is infinite) that $z^1 \in X$. If the point z lies in a small neighbourhood of z^1, then it most likely falls into the set X too. This time our confidence is not infinite but still is big. As the point z moves away from z_1, this confidence decreases to zero since the system forgets what has occurred at z_1. In the same way, we assign the value $-\infty$ to our confidence if the point z^1 falls into the negative set Y and so on. We can imagine that each training signal creates a scalar influence field (positive or negative) around itself.

Specify an analytic expression for the intensity of the scalar influence field of a signal. First, determine the the indicator $\varepsilon(z)$ of the signal z:

$$\varepsilon(z) = \begin{cases} +1, & z \in X \\ -1, & z \notin X \end{cases}$$

The function $h(z)$

$$h(z) = \begin{cases} \dfrac{1}{|z|^m}, & z \neq \mathbf{0} \\ +\infty, & z = \mathbf{0} \end{cases}$$

defines the intensity of the elementary influence field. Here n is the dimension of the space Z, $m \geq n$ is an arbitrary number.

Let the next training signal z^k be supplied to the input. The number

$$\varepsilon(z^k) \cdot h(z - z^k) = \frac{\varepsilon(z^k)}{|z - z^k|^m}, \quad z \neq z^k \tag{3}$$

may be called a heuristic value of our confidence in the fact that the z-sign coincides with the sign of the training signal z_k (recall that the z_k-sign has been prompted by the teacher). Really, values of the function (3) are large in a small neighbourhood of z_k, which corresponds to a large confidence in the sign of z in this neighbourhood. If the point z goes away from z_k, absolute value of the function (4.1) decreases to zero which corresponds to the decrease of the confidence in the z-sign.

Expression (4.1) resembles the formula of Coulombs law for the interaction of point charges. The difference is that the influence field is scalar here. Besides, the exponent m in the formula can exceed 2. Each training signal creates the influence field around itself, which spreads onto surrounding test points. We call the function $\varepsilon(z^k) \cdot h(z - z^k)$ (the argument z and parameter z_k) the elementary influence function of the signal z_k. This function determines the influence field of a particular signal z_k taking into account the sign of the signal and the point of its location. Further we use the notation $h_k(z) \equiv \varepsilon(z^k) \cdot h(z - z^k)$. Thus, each training signal creates the scalar field of positive or negative influence on all the other points around it. The elementary influence function determines the intensity magnitude of this field. The set $\{z^1, z^2, \ldots, z^N\}$ of several training signals also induces the influence field. This field is a superposition of elementary fields of the signals z^k, $k = 1, 2, \ldots, N$. We define the intensity magnitude of this field at the point z as the sum $h_1(z) + \ldots + h_N(z)$ of the values of the elementary influence functions of the training signals z^k.

All the training signals influence the signs of other point signals. So the positive and the negative groups of training signals compete with each other in their «struggle» for the influence on a particular signal $z \neq z^1, z^2, \ldots$. So, the new algorithm can be called competition algorithm (CA). The idea of the competition lies at its basis, as well as the idea of an artificial neuron lies at the basis of the ANN algorithm.

Fig. 2 shows the effect of the influence functions of negative (black) and positive (white) training signals. As the color of the points of the space continuously changes from pure black to pure white, the values of the influence function grow from $-\infty$ to $+\infty$. The points distant from the training signals are colored neutral grey and the values of the influence function are close to zero at this points.

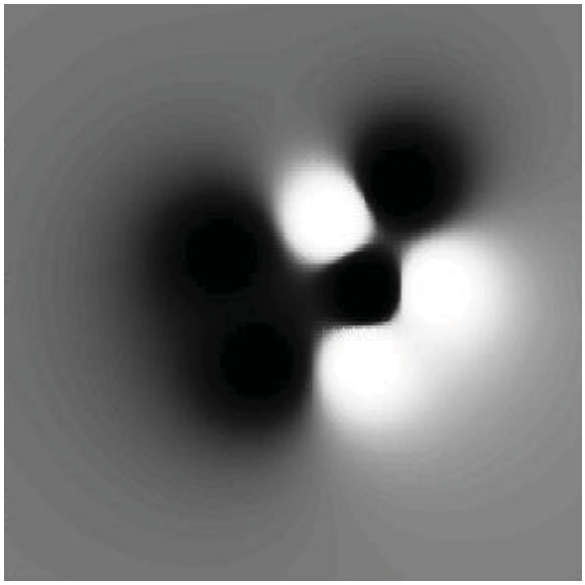

Figure 2: Influence function diagram.

c The approximate decision functions and the new algorithm

Introduce a few auxiliary functions.
(1) $F(t):\ (-\infty,\ +\infty) \to (0,\ 1)$ is a function variable with a typical s-shaped graph. This function is intended for transforming a function $R^1 \to R^1$ with an unbounded value domain into a function with the required value domain $[0, 1]$ and can be defined by the formula

$$F(t) = \frac{1}{2}\left[\frac{t}{\sqrt{t^2+1}} + 1\right] \qquad \lim_{t\to-\infty} F(t) = 0, \quad \lim_{t\to+\infty} F(t) = 1$$

(2) $H_N(z) = \sum_{k=1}^{N} h_k(z)$ is the influence function for the set $\{z^1, z^2, \ldots,\ z^N\}$.
(3) Define the decision functions f_k. Let $f_0(z) \equiv 1/2$ and

$$f_N(z) = F(H_N(z)), \quad z \neq z^k, \qquad f_N(z^k) = \lim_{z\to z^k} F(H_N(z)) = \begin{cases} 1, & z^k \in X \\ 0, & z^k \notin X \end{cases}$$

The function H_k is not defined at the points $z = z^k,\ k = 1, \ldots N$ and we define f_k at these points. Figure 5 presents the graph of the decision functions for n = 2. It is easy to see that the constructed functions f_k satisfy the predesigned properties.

Write down a sequence of operations for solving the problem which make the competition algorithm.

- Specify the influence function $H_0(z) \equiv 0, z \in Z$.
- Simulate the training signal z^k uniformly in Z $(X \subset Z \subset R^n)$.
- Change the influence function: $H_{k-1}(z) \to H_k(z) = H_{k-1}(z) + h_k(z)$.

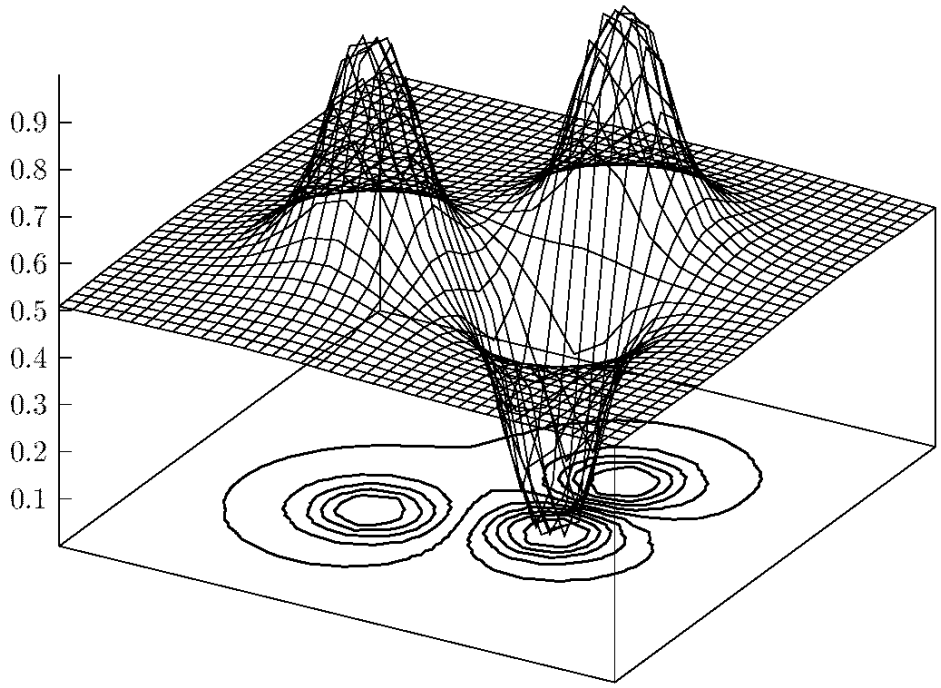

Figure 3: Graph of the decision function, $n = 2$.

- Construct the decision function f_k.

- Choose the step number N for finishing, stop the training process and pass to its working mode. In the working mode, the value $f_N(z)$ is calculated for each point $z \in Z$. If $f_N(z) \leq 1/2$, we hold that the event does not occur; otherwise, we assume the event takes place.

5 Numerical experiment

Let A be a 3×3 matrix with known coefficients and unknown determinant $det(A)$. We want to separate the matrices $A : det(A) < 0$ from those $A : det(A) > 0$.

Write down the matrix A as a set of its columns: $A = (a_{\cdot 1}, a_{\cdot 2}, a_{\cdot 3})$. The set of all 3×3 matrices is unbounded. Therefore, solving the problem, we reduce consideration to the bounded subset $\{\overline{A}\}$ of matrices $\overline{A}$ composed of the normed column vectors $\overline{a}_{\cdot k}$. Just consider association $A \to \overline{A}$ and notice that the sign of $det(\overline{A})$ coincides with the sign of $det(A)$.

Each normed column vector is a point on the two-dimensional sphere S_2, so the matrix $\overline{A} \in S_2 \times S_2 \times S_2$. The training process consists of modeling uniformly distributed training signals, which should be densely distributed in the set $Z = \{\overline{A}\}$. We use the following Monte Carlo formulas for modeling the vector coordinates [1]: $x_1 = 1 - 2\alpha_1$, $x_2 = \sqrt{1 - (2\alpha_2 - 1)^2}\ \cos(2\pi\alpha_3)$, $x_3 = \sqrt{1 - (2\alpha_4 - 1)^2}\ \cos(2\pi\alpha_5)$,. After finding the fraction of correct answers in the group of working signals, we use

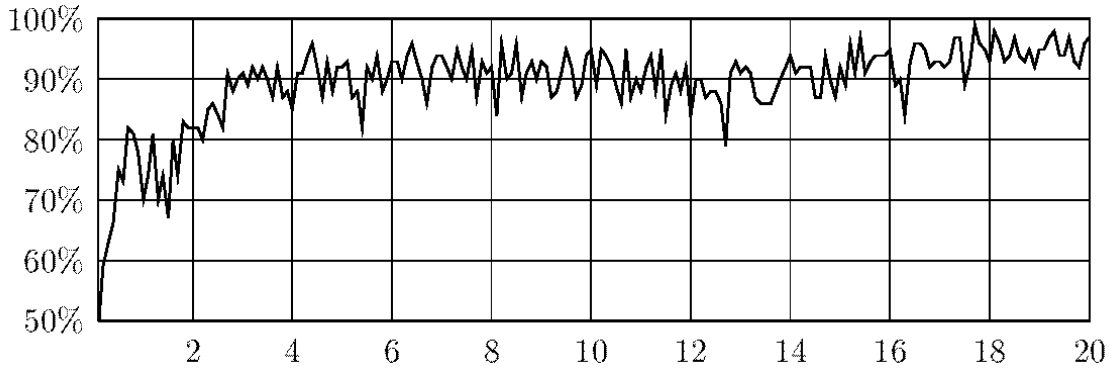

Figure 4: Graph of the frequency p (the number of training samples in thousands).

these signals for further training. We can track how the frequency of the correct answers depends on the number of training samples. Divide the matrices sequence $A_1, A_2, \ldots$ into groups: $G_1 = \{A_1, \ldots, A_{100}\}$, $G_2 = \{A_{101}, \ldots, A_{200}\}, \ldots$, and calculate the number n_k of the correct answers for each group G_k: $p(k) = (n_k/100) \cdot 100\%$. The graph in Fig. 6 shows: at the start $p = 1/2$ (the absence of information), the value of p grows quickly, then the growth rate decreases, at the end the value of p is close to one.

References

[1] Ermakov S.M., Mikhailov G.A. (1976) *The Course of Statistical Modelling.* Nauka, Moscow.

[2] Minsky M.L., Papert S.A. (1988) *Perceptrons: An Introduction to Computational Geometry.* MIT Press, Cambridge.

[3] Wasserman P. (1990) *Neurocomputing: Theory and Practice.* Van Nostrand Reinhold, New York.

Parametric Analysis of the Solution to Smoluchowski Equation by Weight Simulation of Multi-Particle System

Aleksandr Burmistrov[1,2] and Mariya Korotchenko[1]
[1] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS, prospect Akademika Lavrentjeva, 6, Novosibirsk, Russia, 630090*
[2] *Novosibirsk State University, Pirogova str., 2, Novosibirsk, Russia, 630090*
e-mail: burm@osmf.sscc.ru, kmaria@osmf.sscc.ru

Abstract

The Smoluchowski equation with linear coagulation coefficients depending on two parameters is considered. We construct weight algorithm for estimating various linear functionals in ensemble, which is governed by the equation under study. Numerical results show that constructed algorithm simultaneously estimate both the functionals for various parameters and parametric derivatives on the same trajectories of the simulated Markov chain.

Keywords: Monte Carlo method, coagulation, linear functional, weight estimator.

Introduction

In this paper we consider pure coagulation Smoluchowski equation, which describes a wide class of coalescence processes in physical systems consisting of particles with positive integer sizes. For given coagulation coefficients K_{ij} let the probability of interaction (or *collision*) for particles with sizes i and j during a time interval Δt be equal to $K_{ij}\Delta t$. Call a particle of size l an l-mer. Under these notations, a concentration of l-mers $n_l(t)$ at the instant t in spatially homogeneous case satisfies the following kinetic equation:

$$\frac{\partial n_l(t)}{\partial t} = \frac{1}{2}\sum_{i+j=l} K_{ij} n_i(t) n_j(t) - \sum_{i\geq 1} K_{il} n_i(t) n_l(t), \;\; l \geq 1. \tag{1}$$

This equation gives the rate of change of the l-mer concentration with respect to time as the sum of two terms: the first one is the rate at which l-mers form from the coagulation of smaller particles (the factor of $1/2$ ensures that each of such coagulation is counted once); the second one is the rate of l-mer coagulation with other particles, causing l-mer disappearance. Adding the initial data

$$n_l(0) = n_0(l), \;\; l > 0,$$

to the equation (1), we obtain a Cauchy problem for the nonlinear Smoluchowski equation. Further we will develop weight modifications of statistical simulation for estimating the linear functionals of the function $n_l(t)$.

For numerical estimation of linear functionals we are going to consider evolution of a many-particle system [4]. For this purpose we will simulate a homogeneous Markov chain, which transitions are due to elementary pair interactions. Further we will use the following notations:

- N_0 is the initial number of particles in the system, be given at time $t = 0$;
- l_i is the size of the particle with number i;
- $N \leq N_0$ is the current number of particles in the system;
- number $\varpi = (i, j)$ is the interacting pair;
- $X = (N, L_N) = (N, l_1, \cdots, l_N)$ describes the phase state of the system;
- $A(X) = \sum\limits_{\varpi} a(N, l_i, l_j)$, where for $N > 1$ we have
  $a(\varpi) \equiv a(N, l_i, l_j) = \sum\limits_{l=1}^{\infty} N_0^{-1} K_{l_i,l_j} \delta_{l_i+l_j,l}$, here $\delta_{m,n}$ is a Kronecker delta (and $a(1, l_i, l_j) \equiv 0$);
- $P(X, t)$ is the set of probabilities, which determines the state distribution of the system at the time t;
- $Z = (X, \varpi)$, $\mathtt{d}Z = \mathtt{d}X\mathtt{d}\mu_0(\varpi)$. Integration with respect to the measure μ_0 implies summation over all possible pairs ϖ, and integration over $\mathtt{d}X$ means summation over all values of N and L_N.

Under molecular chaos assumption one can obtain in the limit (see [7] for details)

$$\frac{1}{N_0} \sum_{N=1}^{\infty} \sum_{l_2=1}^{\infty} \cdots \sum_{l_N=1}^{\infty} NP(N, l, l_2, \ldots, l_N, t) \to n_l(t), \text{ when } N_0 \to \infty.$$

This limit allows us to estimate solution to the equation (1) with the help of linear integral equation in the way described as follows.

We construct weight modifications of the Monte Carlo algorithms on the basis of the technique suggested in [6], which introduces the pair number ϖ responsible for a collision in the system to the set of phase coordinates. This approach allowed in [6] to derive a special integral equation for the function $F(Z, t) = F(X, \varpi, t) = a(\varpi)P(X, t)$ in the transformed phase space $\mathbf{Z} \times [0, T]$:

$$F(Z, t) = \int_0^t \int_{\mathbf{Z}} F(Z', t') K(Z', t' \to Z, t)\, \mathtt{d}Z'\, \mathtt{d}t' + F_0(Z)\delta(t).$$

Here $\delta(\cdot)$ is a Dirac delta function. The latter equation can be used to construct standard weight modifications of the Markov chain simulation for a many-particle system due to multiplicative structure of its kernel:

$$K(Z', t' \to Z, t) = K_1(t' \to t|X') \cdot K_2(\varpi|X') \cdot K_3(X' \to X|\varpi).$$

The distribution density of the time between elementary interactions is exponential:

$$K_1(t' \to t|X') = A(X') \exp\{-A(X')(t-t')\}.$$

The probability that a pair of particles $\varpi = (i,j)$ interacts in the N'-particle system is

$$K_2(\varpi|X') \equiv K_2(i,j|X') = \frac{a(\varpi)}{A(X')} = \frac{a(N', l_i, l_j)}{A(X')}.$$

Finally, the function $K_3(X' \to X|\varpi)$ defines the transformation of the system after an interaction of the pair ϖ, which results in replacement of two interacting particles i and j by a single particle of the size $l = l_i + l_j$, so $N = N' - 1$.

Thus, the simulation process of the next interaction in the Markov chain includes two successive elementary transitions: first we choose the time interval between interactions, and then we choose two particles for interaction.

Usually the following functionals are of interest:

$$J_H(T) = \int\limits_{\mathbf{X}} H(X) P(X,T)\, \mathtt{d}X.$$

For the function

$$\tilde{H}(X,t) = H(X) \exp\{-A(X)t\}, \quad H(X) \in L_\infty, \tag{2}$$

the following equality was derived in [6]:

$$J_H(T) = \int\limits_0^T \int\limits_{\mathbf{Z}} \tilde{H}(X, T-t') F(Z,t')\, \mathtt{d}Z\, \mathtt{d}t' \equiv (F, \tilde{H}),$$

which we will make use of later.

1 Problem statement

In this work we consider the case of linear coefficients depending on two parameters **a** and **b**

$$K_{ij} = \mathbf{a} + \mathbf{b}\frac{(i+j)}{2}. \tag{3}$$

These coefficients can be found, for example, in the classical polymer model **A**—R—$\mathbf{B}_{f-1}$ (see [1]). In this model the molecules with $(f-1)$ chemically active units of one kind (**B**), and a single unit of another kind (**A**), are regarded as monomers. Chemical bonds could be formed between **A** and **B** units, regardless of ring formation (i. e. *cyclization*), and reactions between units of the same kind are forbidden. This leads to branched molecules if $f \geq 3$ (see Figure 1). As the number of unreacted **A**'s per j-meric molecule is one and the number of unreacted **B**'s is $(f-2)j + 1$, for this model the coagulation (or polymerization) rate is proportional to $(i+j)(f-2)+2$, i. e. has a form (3).

In this paper we are interested in construction of new algorithms for two problems:

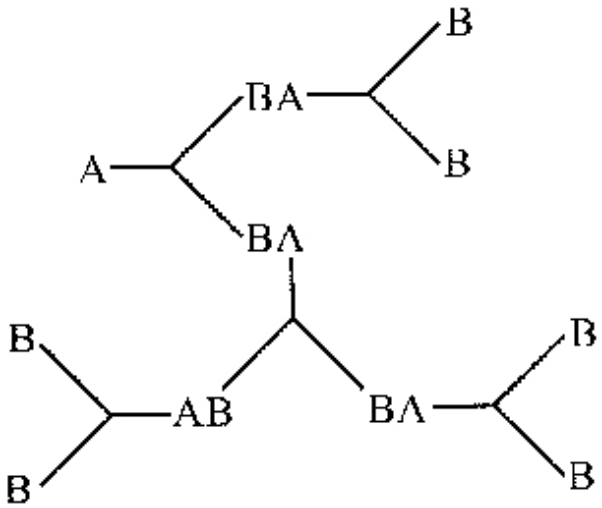

Figure 1: An example of 5-mer for $f = 3$

1. estimation of functionals $J_H(T) \equiv J_H(\mathbf{a}, \mathbf{b}, T)$ for various parameters $\mathbf{a}$ and $\mathbf{b}$ using simulation of the many-particles ensemble for a given $\mathbf{a}^*$ and $\mathbf{b}^*$;

2. estimation of parametric derivatives $\dfrac{\partial J_H}{\partial \mathbf{a}}(\mathbf{a}^*, \mathbf{b}^*, T)$ and $\dfrac{\partial J_H}{\partial \mathbf{b}}(\mathbf{a}^*, \mathbf{b}^*, T)$. with respect to $\mathbf{a}$ and $\mathbf{b}$.

These problems were stated in our previous work [3], in which we suggested value algorithms for reduction in computational cost. For the considered case of linear coefficients K_{ij} we have

$$a(\mathbf{a}, \mathbf{b}, \varpi) = \frac{2\mathbf{a} + \mathbf{b}(l_i + l_j)}{2N_0}, A(\mathbf{a}, \mathbf{b}, X) = \frac{(N-1)}{2}\left[\mathbf{a}\frac{N}{N_0} + \mathbf{b}\right],$$

$$K_1(\mathbf{a}, \mathbf{b}, t' \to t|X') = A(\mathbf{a}, \mathbf{b}, X')\mathrm{e}^{\{-A(\mathbf{a},\mathbf{b},X')(t-t')\}},$$

$$K_2(\mathbf{a}, \mathbf{b}, i, j|X') = \frac{2\mathbf{a} + \mathbf{b}(l_i + l_j)}{\mathbf{a}N(N-1) + \mathbf{b}N_0(N-1)}.$$

2 Weight simulation of Markov chain for integral equation

To solve the first problem stated in the previous section we suggest to simulate ensemble evolution for parameters $\mathbf{a}^*$ and $\mathbf{b}^*$ and estimate the functionals for another parameters with the help of the weight simulation. For this purpose we define the simulated Markov chain $\{Z_n, t_n\}_{n=0}^{\kappa}$; $\kappa = \max\limits_n\{n : t_n < T\}$ with a transition density

$$P^*(Z', t' \to Z, t) = P_1(t' \to t|X') \cdot P_2(\varpi|X') \cdot K_3(X' \to X|\varpi)$$

and a distribution density $F_0(Z)\delta(t)$ of the initial state (Z_0, t_0). Then we can define random weights by the formulas:

$$Q_0 = 1,\ \ Q_n = Q_{n-1}Q(Z_{n-1}, t_{n-1}; Z_n, t_n);\ \ Q(Z', t'; Z, t) = \frac{K_1(t' \to t|X')}{P_1(t' \to t|X')} \cdot \frac{K_2(\varpi|X')}{P_2(\varpi|X')}.$$

We propose to use for simulation the following probability density function P_1 and probabilities P_2:

$$P_1(t' \to t|X') = K_1(\mathbf{a}^*, \mathbf{b}^*, t' \to t|X');\ \ P_2(\varpi|X') = K_2(\mathbf{a}^*, \mathbf{b}^*, i, j|X').$$

Using the forms of K_1 and K_2 we obtain the set of weights $Q(\mathbf{a}, \mathbf{b})$ in the following form:

$$Q(\mathbf{a}, \mathbf{b}, Z_{n-1}, t_{n-1}; Z_n, t_n) = \left.\frac{2\mathbf{a} + \mathbf{b}(l_i + l_j)}{2\mathbf{a}^* + \mathbf{b}^*(l_i + l_j)}\right|_{t=t_n} \times \mathrm{e}^{\left\{-\frac{N_{n-1}-1}{2}\left[\frac{N_{n-1}}{N_0}(\mathbf{a}-\mathbf{a}^*)+(\mathbf{b}-\mathbf{b}^*)\right](t_n - t_{n-1})\right\}}.$$

In order to estimate the functional $J_H(\mathbf{a}, \mathbf{b}, T) = (F(\mathbf{a}, \mathbf{b}), \tilde{H})$, the "weight" collision estimator ξ and absorption estimator η could be used (see [8]):

$$\xi(\mathbf{a}, \mathbf{b}) = \sum_{n=0}^{\kappa} Q_n \tilde{H}(X_n, T - t_n),\ \ \eta(\mathbf{a}, \mathbf{b}) = \frac{Q_\kappa \tilde{H}(X_\kappa, T - t_\kappa)}{q(X_\kappa, t_\kappa)},$$

$$q(X', t') = 1 - \int_{t'}^{T} P_1(t' \to t|X')\, \mathrm{d}t.$$

Taking into account the representation (2), we can show that $\eta = \tilde{Q}_\kappa H(X_\kappa)$ with

$$\tilde{Q}_\kappa(\mathbf{a}, \mathbf{b}) = \prod_{k=1}^{\kappa} \left[\left.\frac{2\mathbf{a} + \mathbf{b}(l_i + l_j)}{2\mathbf{a}^* + \mathbf{b}^*(l_i + l_j)}\right|_{t=t_k}\right] \times \prod_{k=1}^{\kappa+1} \mathrm{e}^{\left\{-\frac{N_{k-1}-1}{2}\left[\frac{N_{k-1}}{N_0}(\mathbf{a}-\mathbf{a}^*)+(\mathbf{b}-\mathbf{b}^*)\right](t_k - t_{k-1})\right\}},$$

where $t_0 \equiv 0$ and $t_{\kappa+1} \equiv T$. Using the results of [5], we can obtain the following theorem.

Theorem 1. *Let $Q(Z', t'; Z, t) < +\infty$, for $Z', Z \in \mathbf{Z}$, and $t', t < T$, then $\mathbf{E}\xi = J_H(T)$. If $q(X, t') > 0$, for $t' < T$, then also $\mathbf{E}\eta = J_H(T)$. Moreover, if the weights are uniformly bounded and $H \in L_\infty$, then there exists T^*, that for $T < T^*$ the variances of the estimators are finite.* □

Taking into consideration the form of weights $Q_n(\mathbf{a}, \mathbf{b})$, we can show that Theorem 1 is valid for our problem for some intervals $\mathbf{a}^* - \varepsilon_a \le \mathbf{a} \le \mathbf{a}^* + \varepsilon_a$ and $\mathbf{b}^* - \varepsilon_b \le \mathbf{b} \le \mathbf{b}^* + \varepsilon_b$.

For solving the second stated problem we suggest to use the following theorem, which is valid for our problem due to the forms of the integral operator $\mathbf{K}$ with the kernel K and the integral operator $\mathbf{K}_P$ with the kernel K^2/P^*.

Theorem 2. *Under conditions of the Theorem 1 let also the spectral radii* $\rho(\mathbf{K}) < 1$, $\rho(\mathbf{K}_P) < 1$, *and the value of* $||K_c'||$ *be uniformly bounded in some interval* $c^* - \varepsilon_c \le c \le c^* + \varepsilon_c$ *for* $c = \mathbf{a}$ *or* $c = \mathbf{b}$. *Then for* $\zeta = \xi$ *or* $\zeta = \eta$ *we have:*

$$\mathbf{E}\left(\frac{\partial \zeta}{\partial c}\right) = \frac{\partial J_H}{\partial c}(\mathbf{a}^*, \mathbf{b}^*, T), \quad \mathbf{Var}\left(\frac{\partial \zeta}{\partial c}\right) < +\infty. \qquad \square$$

For example, for the absorption estimator we have

$$\frac{\partial J_H}{\partial \mathbf{a}}(\mathbf{a}^*, \mathbf{b}^*, T) = \mathbf{E}\left[\tilde{Q}^a_\kappa H(X_\kappa)\right], \;\; \frac{\partial J_H}{\partial \mathbf{b}}(\mathbf{a}^*, \mathbf{b}^*, T) = \mathbf{E}\left[\tilde{Q}^b_\kappa H(X_\kappa)\right],$$

where the weights are the following:

$$\begin{aligned}
\tilde{Q}^b_\kappa &= \sum_{k=1}^{\kappa} \frac{(l_i + l_j)}{2}\left[\mathbf{a}^* + \frac{\mathbf{b}^*}{2}(l_i + l_j)\right]^{-1}_{t=t_k} - \sum_{k=1}^{\kappa+1}\left[\frac{(N_{k-1} - 1)}{2}\right](t_k - t_{k-1}), \\
\tilde{Q}^a_\kappa &= \sum_{k=1}^{\kappa} \left[\mathbf{a}^* + \frac{\mathbf{b}^*}{2}(l_i + l_j)\right]^{-1}_{t=t_k} - \sum_{k=1}^{\kappa+1}\left[\frac{N_{k-1}(N_{k-1} - 1)}{2N_0}\right](t_k - t_{k-1}).
\end{aligned}$$

3 Results of the numerical experiments

Without loss of generality we present numerical results for estimation of two functionals: the monomer concentration $J_{H_1}(T)$, and the total polymer concentration $J_{H_\mu}(T)$ with

$$H_1(X) = \frac{1}{N_0}\sum_{i=1}^{N} \delta(l_i - 1); \;\; H_\mu(X) = \frac{1}{N_0}\sum_{i=1}^{N} 1 \equiv \frac{N}{N_0}.$$

These functionals estimate the solution to the initial equation (1): $J_{H_1}(T) = n_1(T) + \mathcal{O}(N_0^{-1})$, $J_{H_\mu}(T) = \mu(T) + \mathcal{O}(N_0^{-1})$. Here $\mu(T) = \sum\limits_{i=1}^{\infty} n_i(T)$ is the total polymer concentration. The deterministic error of order $\mathcal{O}(N_0^{-1})$ occurs due to the finiteness of N_0 (see [7] for details).

Further in this section the simulation results according to the suggested algorithms are presented and compared to the analytic solution of the test problem. As a test problem for implementation of the algorithms described above, we take the problem (1) with the coagulation coefficients (3) and the initial data $n_0(l) = \delta_{l,1}$ (monodisperse equation). This problem has an exact solution in the form (see [9]):

$$n_l(t) = \mu(t)(1 - \mu(t))^{l-1}\left(\frac{\mathbf{a}\mu(t) + \mathbf{b}}{\mathbf{a} + \mathbf{b}}\right)^{1 + l\mathbf{b}/\mathbf{a}}, \;\; \mu(t) = \frac{\mathbf{b}}{(\mathbf{a} + \mathbf{b})\exp\{\mathbf{b}t/2\} - \mathbf{a}}.$$

We used the following data in the algorithm: $N_0 = 200$, $T = 1.2$, $\mathbf{a}^* = 1.3$, $\mathbf{b}^* = 1.1$, $\mathbf{a} = \mathbf{a}^* \pm 10\%$, $\mathbf{b} = \mathbf{b}^* \pm 10\%$, and $M = 10^7$ is the number of simulated trajectories. Note that the statistical error is of order $\mathcal{O}(M^{-1/2})$ (see, e. g., [8]). We also used the following notations in the tables: $\bar{\sigma}$ is the mean square error (square root of the estimate variance); and PE% is the percent error. You can find some results in the Tables 1-3.

Table 1: Estimation of $J_{H_1}(T)$.

a	**b**	exact solution	estimator $\pm$ $\bar{\sigma}$	PE%
1.3	1.1	$1.429 \cdot 10^{-1}$	$1.436 \cdot 10^{-1} \pm 8.1 \cdot 10^{-6}$	0.46
1.43	1.1	$1.340 \cdot 10^{-1}$	$1.347 \cdot 10^{-1} \pm 1.8 \cdot 10^{-5}$	0.50
1.43	1.21	$1.247 \cdot 10^{-1}$	$1.253 \cdot 10^{-1} \pm 5.1 \cdot 10^{-5}$	0.45
1.3	1.21	$1.328 \cdot 10^{-1}$	$1.333 \cdot 10^{-1} \pm 2.6 \cdot 10^{-5}$	0.40
1.17	1.21	$1.416 \cdot 10^{-1}$	$1.422 \cdot 10^{-1} \pm 1.6 \cdot 10^{-5}$	0.37
1.17	1.1	$1.527 \cdot 10^{-1}$	$1.534 \cdot 10^{-1} \pm 3.7 \cdot 10^{-5}$	0.43
1.17	0.99	$1.648 \cdot 10^{-1}$	$1.656 \cdot 10^{-1} \pm 1.2 \cdot 10^{-4}$	0.49
1.3	0.99	$1.539 \cdot 10^{-1}$	$1.547 \cdot 10^{-1} \pm 4.6 \cdot 10^{-5}$	0.49
1.43	0.99	$1.441 \cdot 10^{-1}$	$1.449 \cdot 10^{-1} \pm 1.8 \cdot 10^{-5}$	0.53

Table 2: Estimation of $J_{H_\mu}(T)$.

a	**b**	exact solution	estimator $\pm$ $\bar{\sigma}$	PE%
1.3	1.1	$3.290 \cdot 10^{-1}$	$3.315 \cdot 10^{-1} \pm 9.1 \cdot 10^{-6}$	0.76
1.43	1.1	$3.175 \cdot 10^{-1}$	$3.201 \cdot 10^{-1} \pm 4.6 \cdot 10^{-5}$	0.83
1.43	1.21	$3.005 \cdot 10^{-1}$	$3.031 \cdot 10^{-1} \pm 1.3 \cdot 10^{-4}$	0.85
1.3	1.21	$3.112 \cdot 10^{-1}$	$3.137 \cdot 10^{-1} \pm 6.2 \cdot 10^{-5}$	0.78
1.17	1.21	$3.228 \cdot 10^{-1}$	$3.251 \cdot 10^{-1} \pm 2.9 \cdot 10^{-5}$	0.71
1.17	1.1	$3.414 \cdot 10^{-1}$	$3.438 \cdot 10^{-1} \pm 7.3 \cdot 10^{-5}$	0.70
1.17	0.99	$3.610 \cdot 10^{-1}$	$3.636 \cdot 10^{-1} \pm 2.4 \cdot 10^{-4}$	0.70
1.3	0.99	$3.477 \cdot 10^{-1}$	$3.502 \cdot 10^{-1} \pm 9.8 \cdot 10^{-5}$	0.73
1.43	0.99	$3.352 \cdot 10^{-1}$	$3.379 \cdot 10^{-1} \pm 4.2 \cdot 10^{-5}$	0.80

Table 3: Estimation of $\dfrac{\partial J_H}{\partial \mathbf{a}}(T)$ and $\dfrac{\partial J_H}{\partial \mathbf{b}}(T)$.

functional	exact solution	estimator $\pm$ $\bar{\sigma}$	PE%
$\dfrac{\partial J_{H_1}}{\partial \mathbf{a}}(T)$	$-7.170 \cdot 10^{-2}$	$-7.156 \cdot 10^{-2} \pm 1.9 \cdot 10^{-4}$	0.18
$\dfrac{\partial J_{H_1}}{\partial \mathbf{b}}(T)$	$-9.589 \cdot 10^{-2}$	$-9.687 \cdot 10^{-2} \pm 2.9 \cdot 10^{-4}$	1.02
$\dfrac{\partial J_{H_\mu}}{\partial \mathbf{a}}(T)$	$-9.198 \cdot 10^{-2}$	$-9.093 \cdot 10^{-2} \pm 4.2 \cdot 10^{-4}$	1.14
$\dfrac{\partial J_{H_\mu}}{\partial \mathbf{b}}(T)$	$-1.654 \cdot 10^{-1}$	$-1.659 \cdot 10^{-2} \pm 6.5 \cdot 10^{-4}$	0.28

Conclusions

We have constructed algorithm to estimate the functionals for various parameters ($\mathbf{a}$, $\mathbf{b}$) as well as parametric derivatives using the same set of trajectories. This algorithm could be useful for solving the interpolation problem. We would like to point out that

we used the monodisperse pure coagulation equation as test problem due to the fact that there exists an analytic solution only for this problem. The algorithm works for polydisperse case (initial particles of varying size) as well. The reduction of the computational cost for estimation of various functionals for large values of T is a challenging problem. It could be solved by the combination of the weight parametric simulation (suggested in this work) and the value simulation (see [2, 3]).

Acknowledgements

The authors acknowledge the kind hospitality of the hotel "Lesnaya Skazka", NSTU and the AMSA'2013 conference organizers. This work was partly supported by SB RAS (Interdisciplinary Integration Grant No. 47) and the RFBR (grants 11-01-00252, 12-01-31134, 13-01-00746).

References

[1] Flory P.J. (1953). *Principles of polymer chemistry.* Cornell University Press, New York.

[2] Korotchenko M.A. (2012). Value Monte Carlo algorithms for estimating the solution to the coagulation equation. In: *L. Plaskota and H. Wozniakowski (eds.), Monte Carlo and Quasi-Monte Carlo Methods 2010*, Springer Proceedings in Mathematics & Statistics. Vol. **23**, pp. 511-522.

[3] Korotchenko M.A., Burmistrov A.V. (2012). Value simulation of the interacting pair number for solution of the monodisperse coagulation equation. In: *Monte Carlo Methods and Applications. Proceedings of the 8th IMACS Seminar on Monte Carlo Methods, August 29 – September 2, 2011, Borovets, Bulgaria, edited by Karl K. Sabelfeld, Ivan Dimov*, De Gruyter Proceedings in Mathematics, pp. 135-143.

[4] Lushnikov A.A. (1978) Some new aspects of coagulation theory. *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosfer. i Okeana.* Vol. **14**:10, pp. 738-743. (In Russian)

[5] Mikhailov G.A. (1999). *Parametric Estimates by the Monte Carlo Method.* VSP, Utrecht.

[6] Mikhailov G.A., Rogasinsky S.V. (2002). Weighted Monte Carlo methods for approximate solution of the nonlinear Boltzmann equation. *Sib. Math. J.* Vol. **43**:3, pp. 496-503.

[7] Mikhailov G.A., Rogasinsky S.V., Ureva N.M. (2006). Weighted Monte Carlo methods for an approximate solution of the nonlinear coagulation equation. *Comp. Math. Math. Phys.* Vol. **46**:4, pp. 680-690.

[8] Mikhailov G.A., Voitishek A.V. (2006). *Numerical Statistical Modelling (Monte Carlo Method).* Akademia, Moscow. (In Russian)

[9] Spouge J.L. (1983). Solutions and critical times for the monodisperse coagulation equation when $a(i, j) = A + B(i + j) + Cij$. *J. Phys. A: Math. Gen.* Vol. **16**, pp. 767-773.

Alternatives for Wiener and Gamma Degradation Models: Method of Selection

Ekaterina V. Chimitova and Evgeniya S. Chetvertakova
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: ekaterina.chimitova@gmail.com, evgenia.chetvertakova@gmail.com

Abstract

In this paper, degradation models are considered. We suppose, that the degradation process is a stochastic process with independent increments (in this case increments of the degradation index). We propose the algorithm for identification of the distribution of degradation increments, as well as the algorithm for estimation of the reliability at some given moment of time. We have investigated the operation of these algorithms using computer simulations. In this paper, we also give an example of the application of proposed algorithms for the GaAs lasers data.

Keywords: degradation process, stochastic process with independent increments, reliability, distribution of degradation increments, GaAs lasers data.

Introduction

There are many scientists, who carry out the research in area of degradation. The topic of degradation processes is considered, for example, in papers [1-5]. In theory, the degradation process is often supposed to be a stochastic process with independent increments. Many researchers use gamma-distribution (for example, Nikulin and Bagdonavicius (2001)) or normal distribution (for example, Bordes (2010) and Tang, Yang and Xie (2004)) as the distribution of degradation increments. And it can be explained by repeatability of these distributions (the sum of random variates has the same distribution as each variate). But in practice, there are many cases, when the goodness-of-fit hypothesis with the gamma and normal distributions is not confirmed.

The aim of this research work is to develop the identification algorithm for distribution of degradation increments and estimation algorithm for reliability basing on the degradation data.

1 Identification of the degradation model

Suppose that an increasing stochastic process $Z(t)$ describes the degradation level of an item. The failure occurs, when the degradation level reaches the critical value z_0 [4]:

$$T = \sup\{t : Z < z_0\} = \inf\{t : Z \geq z_0\},.$$

In this paper, the mean degradation $m(t) = \mathbf{E}\left(Z(t)\right)$ is refered to as the trend function. Let we know values of the degradation $Z_j^i, i = \overline{1,n}, j = \overline{1,k}$ for n items at moments t_j. We suppose that the initial value of the degradation level is zero.

We introduce some assumptions.

Assumption 1. The degradation process is a stochastic process with independent increments. We denote the increment as

$$X_j^i = Z_j^i - Z_{j-1}^i, i = 1, n, j = 1, k \tag{1}$$

Assumption 2. Let the distribution of the degradation increments has the following form:

$$F\left(\frac{x}{m(t;\gamma) - m(s;\gamma)}; \theta\right) \tag{2}$$

where $m(t;\gamma)$ is the known trend function of the degradation level, θ is the vector parameter of the distribution (shift and shape parameters). The main assumption is that the difference between the trend functions at time $t = t_j$ and time $s = t_{j-1}$ is the scale parameter of the distribution of increments.

Assumption 3. Let the distribution 2 of increments X_j^i belongs to the collection of distributions $\mathbf{F} = \{F_1, F_2, ..., F_L\}$.

The identification algorithm for the distribution of the degradation increments can be written as following:

1. calculate the degradation increments 1 basing on degradation data;
2. set $i = 1$;
3. select the distribution F_i from the collection of distributions $\mathbf{F}$;
4. estimate the distribution parameters by the sample of increments with maximum likelihood method: trend function parameters are always estimated, shift and shape parameters are estimated if it is necessary;
5. standardize the sample of increments: $\tilde{X}_j^i = \frac{X_j^i}{m(t_j;\hat{\gamma}) - m(t_{j-1};\hat{\gamma})}$, because degradation increments X_j^i can be not identically distributed;
6. calculate the p-value by one of the goodness-of-fit tests, for example, Kolmogorov, Cramer-von Mises-Smirnov or Anderson-Darling tests, basing on the standardized sample of increments $\tilde{X}_j^i$ and the distribution F_i with the scale parameter equal to 1;
7. if $i = L$, go to step 8, otherwise set $i = i + 1$ and go to step 3;
8. choose the distribution, which has the biggest p-value $\alpha_n > \alpha$, where α is the significance level.

The most difficult stage of the algorithm is the calculation of p-values of the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests, which is based on the distribution of the test statistics under true null hypothesis. Here, the values of test statistics are calculated after estimation of unknown parameters of the model. In this case the distribution of the test statistic is affected by a number of factors, such

as the form of the distribution F_i, the type and the number of estimated parameters, the method of parameter estimation used, and so on. Approximations for the limiting distributions of the test statistics for testing various composite hypotheses have been discussed in works of Lemeshko, for example, [6].

In this paper, we have investigated the distributions of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling statistics by means of Monte-Carlo simulations. It has been shown, that on step 6 of the algorithm, it is possible to use approximations of limiting distributions of considered statistics, obtained in [6], for calculation of p-values.

To check the assumption, that the degradation process is a stochastic process with independent increments, it is possible to test the lack of trend hypothesis by the sample of increments, ordered according to the time of measurements. In [6], it is recommended to use the Foster-Stuart and Cox-Stuart tests as they have rather high power comparing to the Wald-Wolfowitz, Bartlett and Hsu tests. The assumption of independence of increments for the constructed degradation model can not be held, if the trend function is not appropriate. In this case, the lack of trend hypothesis will be rejected. It has been shown, that the Foster-Stuart and Cox-Stuart tests have rather high power in such situations.

The aim of degradation data analysis is to calculate the probability of no-failure lifetime for a given period of time, that is the value of reliability function, which is defined as

$$S(t_s) = P(T > t_s) = P\left(Z(t_s) < z_0\right), \tag{3}$$

where z_0 is the critical value of the degradation level.

The value of reliability function at the given moment t_s can be estimated on the basis of the obtained distribution of degradation increments $F(t; \hat{\gamma}, \hat{\theta})$ by means of computer simulations using the following algorithm:

1. calculate the moments of time from the initial point till the moment t_s with the time step equal to the arithmetic mean of original time steps;

2. generate the sample of degradation increments for N objects in accordance with the distribution $F(t; \hat{\gamma}, \hat{\theta})$;

3. summing obtained increments for each item, get the sample of values of degradation level at the moment t_s;

4. estimate the reliability 3 basing on the empirical distribution function F_N obtained at step 3:
$$S(t_s) = F_N(z_0).$$

The amount of simulations N should be chosen according to the desired deviation of the empirical distribution F_N from the distribution of degradation level at the moment t_s. For example, if we want to get the deviation not more than 0.01 with the probability 0.99, then we need to set N not less than 16590.

2 The analysis of the degradation processes based on the data of the GaAs lasers testing

To analyze the work of identification and estimation algorithms, we have considered the data on degradation of gallium arsenide (GaAs) lasers [3]. Gallium arsenide (GaAs) lasers are used in telecommunication systems. During the operation, they consume more and more current to maintain a fixed level of light output. They are provided with feedback devices that support the consistency of luminous flux. A device fails, when it consumes a current on 10% higher than the nominal value. 15 lasers have been tested under ambient temperature, increased to 80 degrees of Celsium. During the test 3 lasers had failed. Failures had occurred at 3374, 3521 and 3781 hours.

In accordance with the requirements, these lasers must operate at least 200 thousand hours under ambient temperature equal to $20C$. Basing on the previous experience, the engineers supposed that the increase of ambient temperatures up to 80 degrees accelerates the failure in 40 times (conservative estimate). In other words, we have to estimate the reliability at the moment of $200000/40 = 5000$ hours (that is equivalent to two decades of operation).

There are a lot of papers devoted to the analysis of these data. However, in most of them, authors made an assumption on the distribution of increments, and all following conclusions are based on this assumption. For example, in paper [2], authors considered Wiener degradation process for these data, in papers [5] authors compared Wiener and gamma degradation models. In this paper, we question of the normal and gamma distributions for the degradation increments. In accordance with the steps of the identification algorithm of the distribution of increments, we included normal and gamma distributions, as well as the Weibull, Maxwell, exponential and inverse Gaussian distributions into the collection of distributions $\mathbf{F}$. Then, we estimated parameters of the degradation models for each distribution from $\mathbf{F}$. Following the steps of the proposed algorithm, we used the Kolmogorov test to calculate p-values for all considered distributions. The results are given in Table 1. As it is seen from this table, the best distribution for these data is the inverse Gaussian distribution. Moreover, the hypothesis of goodness-of-fit with this distribution is not rejected for $\alpha = 0.05$. Thus, it is seen that the normal and gamma distributions are not appropriate for increments of GaAs lasers degradation.

Table 1: The estimates of parameters of considered degradation models for the GaAs lasers data

Name of the distribution	Shift parameter μ	Form parameter θ	Trend parameter γ	Value of the statistics	p-value α_k
Normal	0.5135	—	0.0008	1.6651	$5.4086e-6$
Inverse Gaussian	—	0.7276; 0.1122	0.0183	0.6774	0.383
Maxwell	—	—	0.0013	1.2558	0.0107
Weibull	—	2.6525	0.0023	1.4127	0.0001
Gamma	—	7.0911	0.0002	1.0696	0.0082

Then, in accordance with the algorithm for estimation of reliability and the selected inverse Gaussian distribution we simulated the distribution of degradation level at moment of time 5000 hours. The fixed value of the critical level is 10. So, the estimate of reliability at moment 5000 hours is $S(5000) = 0.6252$. We can say, that engineers suppositions on the reliability of lasers were incorrect: about 40% of lasers will fail during the 5000 hours of operation under temperature 80 degrees of Celsium (engineers predicted no-failure operation during this period).

Conclusions

In this paper, we proposed the algorithm for identification of the distribution of degradation increments, as well as the algorithm for estimation of the reliability at some given moment of time under assumptions. On the example of the analysis of the GaAs lasers data, it was shown that Wiener and gamma degradation processes, which are commonly used in the analysis of degradation, can be inadequate for the data and, hence, they will provide the wrong prediction of reliability. The most appropriate model among considered is the inverse Gaussian distribution of degradation increments. It was shown, that about 40% of lasers would fail during the 5000 hours of accelerated tests although engineers have predicted no-failure operation during this period.

This research has been supported by the Russian Ministry of Education and Science as part of the state task (project 8.1274.2011) and the Federal Target Program "Research and scientific-pedagogical personnel of innovative Russia" (contract number 14.B37.21.0860).

References

[1] Bordes L., Paroissin C., Salami A. (2010). Parametric inference in a perturbed gamma degradation process. *Preprint/Statistics and Probability Letters*, Vol. **13**.

[2] Liao C.M., Tseng S.-T. (2006). Optimal design for step-stress accelerated degradation test. *IEEE Trans. Reliab.*, Vol. **55**, pp. 59-66.

[3] Meeker W.Q., Escobar L.A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.

[4] Nikulin M.,Bagdonavicius V. (2001). *Accelerated Life Models: Modeling and Statistical Analisys*, Chapman & Hall/CRC, Boca Raton.

[5] Tsai C.-C., Tseng S.-T., Balakrishnan N. (2011). Mis-specification analyses of gamma and Wiener degradation processes. *Journal of Statistical Planning and Inference*, Vol. **141**, pp. 25-35.

[6] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N., Chimitova E.V. (2011). *Statistical data analysis, simulation and study of probability regularities. Computer approach: monograph* , NSTU Publisher, Novosibirsk.

Selection of the Optimal Smoothing Parameter for the Nonparametric Estimation of the Regression Reliability Model

Victor A. Demin[1] and Ekaterina V. Chimitova[1]
[1] *Novosibirsk State Technical University, Novosibirsk, Russia*
e-mail: vicdemin@gmail.com, ekaterina.chimitova@gmail.com

Abstract

The most popular approach for nonparametric estimation of a regression reliability model, proposed by Beran, is considered in this paper. In this paper, we give the detailed analysis of the selection method for the bandwidth parameter, which is based on minimization of the distance of failure times from kernel estimate of the inverse reliability function. The accuracy of the Beran estimator is studied depending on the plan of experiment (the sample size and the number of values of the explanatory variable) and the way of calculating kernel estimates of the inverse reliability function. We formulate some conclusions on the choice of smoothing parameter and kernel function for the kernel estimates of the inverse reliability function, which give the best accuracy of Beran's estimator.

Keywords: reliability function, the regression model, nonparametric Beran estimator, smoothing parameter, bandwidth parameter.

Introduction

In problems of the statistical analysis of lifetime data, such as failure time of technical devices in reliability theory or time of death in medical studies, the most common task is the analysis of the dependence of the reliability (survival) function on the observed explanatory variables. In reliability theory, such factors as temperature, pressure, voltage, mechanical and other are usually taken as explanatory variables or, as they are usually called in the lifetime data analysis, covariates. The most popular parametric regression models in reliability are the AFT (Accelerated Failure Time) model and the proportional hazards model. The construction of any parametric model requires knowledge of the lifetime distribution and the kind of dependence of reliability function on the observed covariates. In practice, however, this information is usually absent. In such a situation it is advisable to use nonparametric methods, which enable not only to estimate the reliability function for different values of the covariate, but also can be used to construct a goodness-of-fit test for some parametric reliability model.

One of the most popular approaches to nonparametric estimation of the regression reliability model is the estimator, proposed by Beran [1]. The investigation of statistical properties of this estimator in the case of random plans, when the value of covariates are not fixed, are presented in [2-5]. In [6], the properties of Beran's estimator are studied, when the values of covariate are defined in advance.

Today, a great number of publications is devoted to the problem of the kernel smoothing; the main attention is usually paid on the problem of selecting the optimal smoothing parameter. In the context of this problem, it is important to understand, that such methods as reference heuristic methods, substitution methods and cross-validation are not applicable for the nonparametric Beran estimator, as in this case the kernel function determines only the weight of each observation according to the value of the covariate.

However, it is known that the quality of the Beran estimator essentially depends on the chosen value of the bandwidth parameter. In [6], a theoretical method of selection of the optimal bandwidth parameter is suggested, however, it is extremely difficult to implement this method in practice, as it uses several functions, which are usually unknown. In [7], the method of selection of the optimal bandwidth parameter, based on the bootstrap procedure is offered, however, this approach is applicable only to the case of the random plan. Thus, it is necessary to develop the method of calculation of the optimal value of the bandwidth parameter for the Beran estimator. In [8], we have proposed the idea of selecting the optimal bandwidth parameter, which is based on the minimization of the distance of failure times from kernel estimate of the inverse reliability function. So, the purpose of this paper is to investigate the statistical properties of the Beran estimator and to give some recommendations on the way of application of the proposed method.

1 Nonparametric Beran estimator

Denote by T_x the lifetime of considered technical product, which depends on scalar covariate. The reliability function is denoted by

$$S(t|x) = P(T_x \geq t) = 1 - F(t|x), \tag{1}$$

where $F(t|x)$ is the conditional distribution function of a random variable T_x.

The main feature of the 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 Beran estimator is defined as follows [1]

$$\tilde{S}_{h_n}(t|x) = \prod_{Y_{(i)} \leq t} \left\{ 1 - \frac{W_n^i(x; h_n)}{1 - \sum_{j=1}^{i-1} W_n^j(x; h_n)} \right\}^{\delta_i}, \tag{2}$$

where x is the value of the covariate, for which reliability function is estimated, $W_n^i(x; h_n)$, $i = 1, \ldots, n$ are the Nadaraya-Watson weights, which are defined as fol-

lows [5]:

$$W_n^i(x;h_n) = K\left(\frac{x-x_i}{h_n}\right)\Big/\sum_{j=1}^{n} K\left(\frac{x-x_j}{h_n}\right),$$

where $K\left(\frac{x-x_i}{h_n}\right)$ is the kernel function, satisfying to the regularity conditions: $K(y) = K(-y)$, $0 \le K(y) < \infty$, $\int_{-\infty}^{\infty} K(y)dy = 1$; $h_n > 0$ is the bandwidth parameter, which satisfies to the conditions: $\lim\limits_{n\to\infty} h_n = 0$, $\lim\limits_{n\to\infty} nh_n = \infty$.

It should be noted, that in the case of the values of the Nadaraya-Watson weights $W_n^i(x;h_n) = n^{-1}$, the Beran estimator is led to the Kaplan-Meier estimator [5]. As it was shown in [9], the accuracy of the Beran estimates essentially depends on the values of the bandwidth parameter and is almost independent of the type of kernel function. The optimal bandwidth parameter depends primarily on the degree of influence of the covariate on the reliability function, while the effect of sample size is not so significant.

2 The choice of bandwidth parameter

The choice of the bandwidth parameter determines the values of the weights $W_n^i(x;h_n)$, which in turn determine, which observations will participate in the construction of the estimate of the conditional reliability function (1). Thus, varying the bandwidth parameter, in a certain way, it is possible to drop "bad" observations.

In this paper, we consider the method for selecting an optimal parameter, which is based on the minimization of the mean deviation failure times $Y_1, Y_2, ..., Y_n$ from nonparametric estimation of the inverse reliability function $S_x^{-1}(p)$ [8]. We denote the inverse reliability function through $g(p|x)$. Then, the model (1) can be rewritten in the form:

$$T_x = g(p|x) + \varepsilon, \tag{3}$$

where $p \in (0,1)$, ε is the error of observation, which, in general, may depend on p and x.

Kernel estimator for the model (3) can be written as

$$\hat{g}(\hat{p}_i|x_i) = \frac{1}{n}\sum_{j=1}^{n} \omega_n^j(\hat{p}_i)\cdot Y_j, \tag{4}$$

where ω_n^j is a certain weight, which can be calculated using various weighting functions. In particular, we consider the Nadaraya-Watson weights of the first order

$$\omega_n^j(\hat{p}_i) = K\left(\frac{\hat{p}_i-\hat{p}_j}{b_n}\right)\Big/\sum_{k=1}^{n} K\left(\frac{\hat{p}_i-\hat{p}_k}{b_n}\right)$$

and the Priestley-Chao weights of the second order [1]:

$$\omega_n^j(\hat{p}_i) = \left\{\hat{p}_{(i)} - \hat{p}_{(i-1)}\right\} K\left(\frac{\hat{p}_i-\hat{p}_j}{b_n}\right),$$

where the smoothing parameter b_n can be selected using one of the methods proposed for kernel smoothing [1,10]. Probabilities $\hat{p}_i$ are calculated using the Beran estimates: $\hat{p}_i = \tilde{S}_{h_n}(Y_i|x_i)$.

Thus, the optimal value of the bandwidth parameter can be obtained by solving the following optimization problem:

$$h_n^{opt} = \arg\min_{h_n} \frac{1}{n} \sum_{i=1}^{n} \delta_i \cdot |\hat{g}(\hat{p}_i|x_i) - Y_i|. \tag{5}$$

3 Choice of weights and smoothing parameter

As we consider the problem, involving the use of kernel smoothing, we can use predeveloped approaches for the optimal bandwidth parameter for the kernel estimator of regression. Let us consider the following methods:

1. The method of cross-validation, which is often regarded as the most accurate; however, it requires significant computational resources.

2. Method of minimal mean of integrated error according to the smoothing parameter, which is calculated as:

$$b_{NS} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2 n}\right]^{1/5} \hat{\sigma},$$

where $\mu_2(K) = \int x^2 K(x)dx$, $R(K) = \int K^2(x)dx$, $\hat{\sigma}$ is the estimate of the variance, which can be calculated in various ways, most often used for this purpose, for example, the sample variance:

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{p}_i - \bar{\hat{p}}\right)^2.$$

However, firstly, this estimate is not robust, and secondly, has "good" properties only if the distribution is close to normal. Therefore, in this paper we shall also consider the robust estimate of the variance:

$$\hat{\sigma} = S_{rob} = \underset{i=1..n}{med} \left| \hat{p}_i - \underset{j=1..n,\, k=j..n}{med} \left(\frac{\hat{p}_j + \hat{p}_k}{2} \right) \right|.$$

This estimate is a combination of the well-known robust estimate of Hodges-Lehmann for the shift parameter and the robust estimate of Rousseeuw, and it is called the median absolute deviation for the scale parameter.

Let us investigate the statistical properties of the Beran estimator using the optimal bandwidth parameter (5). The investigation of the properties of the Beran estimates is carried out by the Monte Carlo simulations. The following statistic is

used as the distance between the Beran estimates and the true conditional reliability function:

$$D_{h_n} = \sup_{j=1..k,\ t<\infty} \left| \tilde{S}_h(t|x_j) - S_{x_j}(t) \right|. \tag{6}$$

It is obvious, that the quality of estimates (4) directly influences on that, how well the bandwidth parameter will be chosen. So, let us compare different weights ω_n^j for the kernel estimator $\hat{g}(\hat{p}_i|x_i)$, as well as different methods of choosing smoothing parameter from the point of view of the accuracy of the Beran estimation.

As the true reliability model we consider the parametric Cox proportional hazards model [9]:

$$S_x(t) = (S_0(t))^{r(x;\beta)}, \tag{7}$$

with the covariate function $r(x;\beta) = \ln(1 + e^{\beta x})$ and the lognormal baseline distribution with the density function:

$$f_0(t) = \frac{1}{\sqrt{2\pi}\theta_1 t} \exp\left(-\frac{1}{2\theta_1^2} \ln^2\left(\frac{t}{\theta_2}\right)\right).$$

We consider the case, when the covariate takes the values from the set
$\{0, 0.11, 0.22, 0.33, 0.44,$
$0.56, 0.67, 0.78, 0.89, 1\}$, the sample size $n = 100$, 200, 300 and the number of observations corresponding to different values of the covariate is equal to each other. The samples were generated according to the model (7) with parameters: $\theta_1 = 21.5, \theta_2 = 1.6$, $\beta = 2$ or $\beta = 5$. The values of the distance (6) are given in Figure 1; the average values of chosen bandwidth parameter $h_n^o pt$ and smoothing parameter b_{NS} are presented in Figures 2 and 3, correspondingly.

In all figures the notation "PCh" indicates the Priestley-Chao weights and "NW" indicates the Nadaraya-Watson weights.

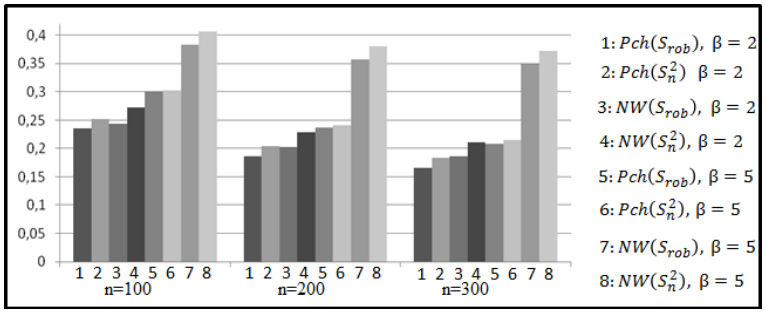

Figure 1: The distance D_n for different sample sizes

It is seen from Figure 1, that the Priestley-Chao weight function allows to get more accurate Beran estimates. Thus, when the sample size is equal to 100, the value of distance (6) in the case of using Prestly-Chao weights is less by 3% in comparison with the case of using Nadaraya-Watson weights; if $n = 200$ the winning is 8% and when $n = 300$ the winning is 11%. Moreover, the usage of robust estimator S_{rob} in

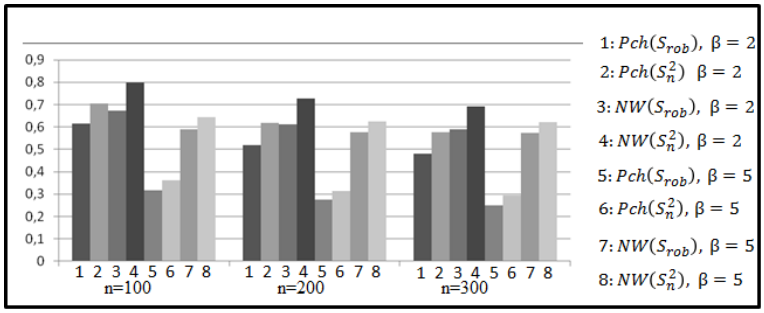

Figure 2: Average values of the bandwidth parameter h_n^{opt} for different sample sizes

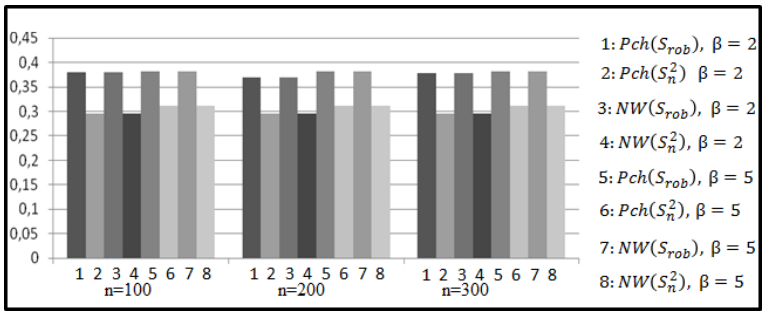

Figure 3: Average values of the smoothing parameter b_n for different sample sizes

calculation of the smoothing parameter b_n gives better accuracy, and accuracy of the Beran estimates increases with the sample size growth.

Figure 2 shows the average values of the chosen bandwidth parameter h_n^{opt}. It is seen, that when the sample size increases, the value of optimal bandwidth parameter reduces; it is quite natural, since the number of observations in groups increases, and hence the number of "bad" observations increases.

Figure 3 illustrates the average values of smoothing parameter b_n. It is curious, that the value of the smoothing parameter practically does not depend on the sample size and the weight function.

It is necessary to note, that the results obtained by the cross-validation are not presented in these figures by two reasons: firstly, the procedure of cross-validation requires extremely large computational resources, for example, when $n = 100$ the time of calculation increases in about 6 times; and secondly, the method of cross-validation minimizes the function (6) according to parameter b_n, but not to the bandwidth parameter h_n.

Similar results have been obtained in experiments for the parameter value $beta = 5$ (i.e. with a stronger covariate effect). As in the considered case, the application of the robust method in conjunction with the usage of Priestley-Chao weights result in better accuracy of Beran estimates. It is interesting to consider apart the behavior of optimal bandwidth parameter h_n^{opt}: when the influence of the covariate on the reliability function increased, the average value of h_n^{opt} decreased almost twice in

the case of using Priestley-Chao weights; however, in the case of Nadaraya-Watson weights such a change is not observed.

Now, let us consider, what happens with the Beran estimates, when different number of groups (the number of different values of the covariate) is taken.

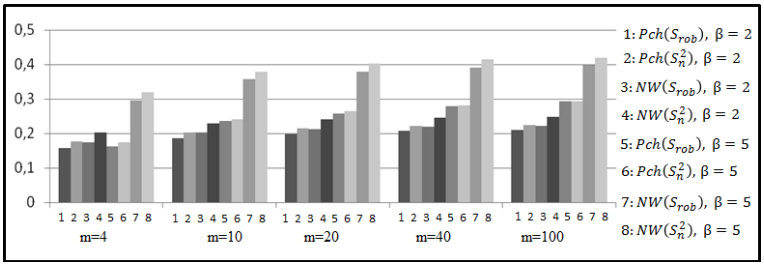

Figure 4: The distance D_n for different numbers of groups

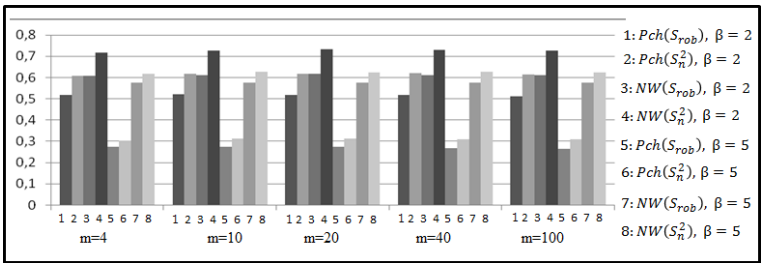

Figure 5: Average values of the bandwidth parameter h_n^{opt} for different numbers of groups

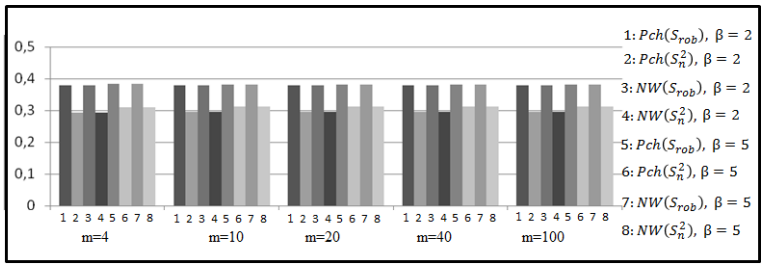

Figure 6: Average values of the smoothing parameter b_n for different numbers of groups

As can be seen from Figure 4, when the number of groups increases for the fixed sample size, the accuracy of the Beran estimator decreases, but this fall is not significant. This result can be explained as follows: the number of observations in a group decreases, therefore, the amount of information for each covariate value also becomes less, what leads to the loss of accuracy. However, the average values of bandwidth

and smoothing parameters (see Figures 5 and 6, correspondingly) practically do not change. Thus, we can assume, that the value of smoothing parameter b_{NS} doesn't depend on the sample size and number of groups. Similar result has been obtained in the case of $\beta = 5$, when the degree of influence of covariate on the reliability function was increased.

Similar investigation has been carried out for the Cox proportional hazards model with exponential baseline distribution. The revealed regularities were almost the same, so specific numerical results are not given here.

Conclusions

In this paper, we have investigated the selection method of the bandwidth parameter for the Beran estimator, which is based on minimization of the distance between failure times and the kernel estimator of the inverse reliability function. We have examined different ways of calculation of the kernel estimator from the position of accuracy of the Beran estimator. It has been shown, that it is preferable to use the Priestley-Chao weight function and to calculate the value of smoothing parameter by the method of minimal mean of integrated error with the robust estimator of variance, when calculating the kernel estimator of the inverse reliability function.

During the investigation it has been found, that the parameter b_{NS} almost does not depend on the experimental design, so it can be calculated only once for one experimental design. This behavior of smoothing parameter can be explained by the fact, that it is necessary to optimize the accuracy of the Beran estimator rather than the kernel estimator of the inverse reliability function, so the accuracy of kernel estimation is not so important.

Acknowledgements

This research has been supported by the Russian Ministry of Education and Science as part of the state task (project 8.1274.2011) and the Federal Target Program "Research and scientific-pedagogical personnel of innovative Russia" (contract number 14.B37.21.0860).

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] Van Keilegom I., Veraverbeke N. (1998). Nonparametric Estimation of the Conditional Distribution in Regression with Censored Data. *Dissertation.* pp. 7-51.

[7] Li G., Datta S. (1999). A bootstrap approach to nonparametric regression for right censored data. *Technical Report.* pp. 1-10.

[8] 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.

Nonparametric Algorithms of Identification, Forecasting and Control

Irina L. Foox, Irina Yu. Glukhova and Gennady M. Koshkin
Tomsk State University,
Tomsk, Russia
e-mail: fooxil@sibmail.com, win32_86@mail.ru, kgm@mail.tsu.ru

Abstract

The principal parts of mean square errors for kernel plug-in estimators of the functions defining ARX-process are found. We use simulation to compare parametric and nonparametric identification algorithms and to study also nonparametric control algorithms. To investigate the dependence of Russian Federation's Industrial Production Index on the dollar exchange rate, direct investments, and export for the period from September 1994 to January 2013, the proposed algorithms of identification and forecasting are applied.

Keywords: Kernel plug-in estimator, conditional mean, mean square error (MSE), ARX-process, nonparametric identification, forecasting algorithm, control.

Introduction

Suppose that a sequence $(Y_t)_{t=\ldots,-1,0,1,\ldots}$ is generated by ARX(m, p, d)-process

$$Y_t = \Psi\left(Y_{t,m}, X_{t,s}\right) + \xi_t, \tag{1}$$

where $Y_{t,m} = \left(Y_{t-i_1}, ..., Y_{t-i_m}\right)$, $X_{t,s} = \left(X^1_{t-j_1}, ..., X^1_{t-j_r}, ..., X^p_{t-j_1}, ..., X^p_{t-j_k}\right)$, $s = r + \ldots + k$, $d = \max(r, \ldots, k)$, $1 \le i_1 < \ldots < i_m \ll n$, $0 \le j_1 < \ldots < j_r \ll n, \ldots, 0 \le j_1 < \ldots < j_k \ll n$ are known subsequences of natural numbers, (ξ_t) is a sequence of i.i.d. random variables with zero mean, finite variance, zero third, and finite fourth moments, $\Psi\left(Y_{t,m}, X_{t,s}\right)$ is an unknown non-periodic function bounded on compact.

Models (1) are used on identification of economic systems and financial time series analysis. By identifying model (1) we mean the problem of parametric or nonparametric estimation of the function Ψ. In this paper, we assume that the process $(Y_t)_{t=\ldots,-1,0,1,\ldots}$ is a strictly stationary process and satisfies the strong mixing (s.m.) condition with s.m. coefficient ([10], [11], [3], [4])

$$\alpha(\tau) \approx e^{-\delta\tau}, \quad \delta > 0, \quad \tau \to \infty. \tag{2}$$

Let $Y_1, Y_2, \ldots, Y_n$ be observations generated by the process (1). As a model of the structure of Ψ in (1), we take the conditional expectation

$$b(y, x) = E(Y_t | Y_{t,m} = y, X_{t,s} = x) = E(Y | y, x), \quad (y, x) \in R^{m+s}.$$

According to [1] the integrals $a_g(y,x) = \int q^g f(q,y,x)dq, \;\; g = 0,1$, are basic functionals, where $f(q,y,x)$ is an unknown probability density function (p.d.f.) of a random vector $(Y_t, Y_{t,m}, Y_{t,s})$ in stationary conditions. Since $a_0(y,x) = \int f(q,y,x)dq = p(y,x)$, where $p(y,x)$ is p.d.f. of $(Y_{t,m}, Y_{t,s})$, then the conditional expectation can be written as

$$b(y,x) = \frac{a_1(y,x)}{a_0(y,x)} = \frac{a_1(y,x)}{p(y,x)} = \int Y_t f(Y_t|y,x)dY_t.$$

We take the kernel estimators of basic functionals $a_g(y,x)$ at the point (y,x) in the form

$$a_{gn}(y,x) = \frac{1}{n-Q}\sum_{i=Q+1}^{n} \frac{Y_i^g}{\prod\limits_{j=1}^{m} h_j} K_m\left(\frac{y-Y_{i,m}}{h^y}\right) \frac{K_s\left(\frac{x-X_{i,s}}{h^x}\right)}{\prod\limits_{j=1}^{r} h_{1j}\cdots\prod\limits_{j=1}^{k} h_{pj}},$$

where $Q = \max(i_m, \max(j_r,\ldots,j_k))$, $h^y = (h_1,\ldots,h_m)$, $h^x = (h_1^x,...,h_p^x)$, $h_1^x = (h_{11},...,h_{1r}),\ldots,$ $h_p^x = (h_{p1},...,h_{pk})$ are suitable bandwidths (positive numbers), K_m and K_s are m- and s-dimensional kernels. Thus, the kernel plug-in estimator of conditional functional $b(y,x)$ at the point (y,x) and, hence, the function $\Psi(y,x)$ in (1) is the ratio

$$b_n(y,x) = \Psi_n(y,x) = \frac{\sum\limits_{i=Q+1}^{n} Y_i K_m\left(\frac{y-Y_{i,m}}{h^y}\right) K_s\left(\frac{x-X_{i,s}}{h^x}\right)}{\sum\limits_{i=Q+1}^{n} K_m\left(\frac{y-Y_{i,m}}{h^y}\right) K_s\left(\frac{x-X_{i,s}}{h^x}\right)}. \tag{3}$$

The problem of identifying model (1) is a problem of estimating function (cf. [3], [1])

$$H(A) = H(a_0,a_1)) = \frac{a_1}{a_0}, \;\; a_0 > 0, \tag{4}$$

where $A = (a_0,a_1)$, $a_g = a_g(u) = \int q^g f(q,u)dq$, $g = 0,1$, $f(q,u) = f(z)$, $z \in R^{m+s+1}$, is p.d.f. of the random vector $(Y_t, U_t) = (Y_t, Y_{t,m}, X_{t,s}) = Z_t$ in stationary conditions.

In this paper, we study the mean square convergence of estimator (3) to the function Ψ determining ARX-process (1).

1 The MSE for Plug-in Estimator of Ψ

We introduce the following notation: $f_{1(i+1)(i+j+1)(i+j+k+1)}(z,\upsilon,u,w)$ is $4(m+s+1)$-dimensional p.d.f. of sample vectors $Z_1, Z_{(i+1)}, Z_{(i+j+1)}, Z_{(i+j+k+1)}$,
$a^+_{1(i+1)(i+j+1)(i+j+k+1),p}(z,y,z',y') =$
$\int_{R^4} |\nu\upsilon\nu'\upsilon'|^p f_{1(i+1)(i+j+1)(i+j+k+1)}(\nu,z,\upsilon,y,\nu',z',\upsilon',y')d\nu d\upsilon d\nu' d\upsilon', \; Q{+}1 \le i,j,k < n,$

$i+j+k \le n-1$; $a_{1(1+j)(i+j+k),p}(z,y,z') = \int_{R^3} |\nu\upsilon\nu'|^p f_{1(1+j)(i+j+k)}(\nu,z,\upsilon,y,\nu',z')d\nu d\upsilon d\nu'$, $a^+_{1(i+1),p}(z,z') = \int_{R^2} |\nu\upsilon|^p f_{1(1+j)}(\nu,z,\upsilon,z')d\nu d\upsilon$, $a^+_p(z) = \int |\nu|^p f(\nu,z)d\nu$, $g = 0,1$, $L = m+s$.

Definition 1. A function $K(u)$ belongs to the class of one-dimensional kernels $K(\cdot) \in \mathcal{A}_\nu$ if $\int |K(u)|du < \infty$, $\int K(u)du = 1$, $\int |u^\nu K(u)|du < \infty$, $T_j = \int u^j K(u)du = 0$, $j = 1,\dots,\nu-1$, $T_\nu \ne 0$, and $K(u) = K(-u)$.

Below, to study convergence of estimators, we use the same bandwidth h_n for each variable from $m+s$ variables in Theorem and the product of one-dimensional kernels as multidimensional kernels of proper dimensions.

Denote $\omega_{g\nu}(u) = \dfrac{T_\nu}{\nu!}\sum_{j=1}^{L} \dfrac{\partial^{\,\nu} a_g(u)}{\partial u_j^\nu}$, $H_g = \dfrac{\partial H(A)}{\partial(a_g)}$, $\sup\limits_{u\in R^L} = \sup\limits_u$.

Definition 2. A function $H(\cdot) : R^L \to R^1$ belongs to the class $\mathcal{N}_\nu(z)$ ($H(\cdot) \in \mathcal{N}_\nu(z)$) if it is continuously differentiable up to the order ν at the point $z \in R^L$. A function $H(\cdot) \in \mathcal{N}_\nu(R)$ if it is continuously differentiable up to the order ν for any $z \in R^L$.

Theorem. Assume that for the function $H(A)$ in (4) and integers $g,p = 0,1$, $\beta = 0,4$, $l,\dots,q = \overline{1,L}$, the following conditions hold:

1) (Z_i) satisfies the s.m. condition, $\int_0^\infty \tau^2[\alpha(\tau)]^{\frac{\delta}{2+\delta}}d\tau < \infty$, $0 < \delta < \infty$;

2) $a_{g+p}(\cdot) \in \mathcal{N}_0(R)$, $a^+_{g(2+\delta)}(\cdot) \in \mathcal{N}_0(z)$; $\sup\limits_u a^+_{g+p}(u) < \infty$, $\sup\limits_u a^+_{g\beta}(u) < \infty$;

3) $K(\cdot) \in A_\nu$, $\sup\limits_{u\in R^1} |K(u)| < \infty$;

4) $a_0(u) > 0, a_g(\cdot) \in \mathcal{N}_\nu(R), \sup\limits_u |a_g(u)| < \infty, \sup\limits_u \left|\dfrac{\partial^\nu a_g(u)}{\partial u_l \dots \partial u_q}\right| < \infty$;

5) a non-increasing sequence (h_n) is such that $(d_n) = \left(h_n + \dfrac{1}{nh_n^L}\right) \downarrow 0$;

6) $\sup\limits_u a^+_{1(i+1)(i+j+1)(i+j+k+1),g}(u,u,u,u) < \infty$,
$\sup\limits_u a^+_{1(i+1)(i+j+1),g(2+\delta)}(u,u,u) < \infty$, $\sup\limits_u a^+_{1(i+1),g(2+\delta)}(u,u) < \infty$,
$\sup\limits_{u,u'} a^+_{1(i+1),g+p}(u,u') < \infty$ for any $i,j,k \ge 1$;

7) for all possible values of $Y_1,\dots,Y_n$, $X^1_1,\dots,X^1_n,\dots$, $X^p_1,\dots,X^p_n$ the sequence $\{|H(A_n)|\}$ is dominated by the sequence of numbers $(C_0 d_n^{-\gamma})$, where C_0 is a constant, $0 \le \gamma \le 1/4$.

Then

$$E[H(A_n) - H(A)]^2 = \sum_{g,p=0}^{1} H_g H_p \left[\frac{a_{t+p}(x)}{nh_n^L}\left(\int K^2(u)du\right)^L + \omega_{g\nu}(z)\omega_{p\nu}(z)h_n^{2\nu}\right]$$

$$+ O\left(\left[h_n^{2\nu} + \frac{1}{nh_n^L}\right]^{\frac{3}{2}}\right).$$

Note that in this formula according to (4) $H_0 = -\frac{a_1}{a_0^2}$, $H_1 = \frac{1}{a_0}$. The proofs of Theorem is based on the results, presented in [10], [1], [4]–[6].

2 Comparison of Parametric and Nonparametric Algorithms

Computer modeling is started by generation sequences of dependent observations, using the following processes:
$M(1) : Y_n = 0.2Y_{n-1} + 0.11X_n^1 + 0.15X_{n-1}^1 + 0.3X_n^2 + 0.2X_{n-2}^3 + \xi_n$,
$M(2) : Y_n = e^{0.1Y_{n-1}+0.2X_n^1+0.1X_{n-1}^1+0.01X_{n-1}^2+0.03X_n^3} + \xi_n$.

Here, variables X^1, X^2, and X^3 take values from uniform distributions on the corresponding intervals [2, 2.5], [5, 6], [8, 10], and random variables ξ_n are distributed according to normal distributions with zero mean and variances, calculated for models M(1), M(2) by the formula

$$\sigma^2 = \frac{Y_{max} - Y_{min}}{6} \cdot a, \tag{5}$$

where the multiplier a is a level of noise ξ_n, which takes the values 0.01, 0.05, 0.1, 0.15, 0.2, 0.5. In simulation we use the following sizes of observations: 50, 100, 200, and 500.

Note that condition (2) holds for model $M(1)$ (see subsection 3.5.9 in [1]).

Identification algorithms for functions Ψ in (1) were obtained by the least squares method (LSM), by iterative weighted least squares method (WLSM), and by nonparametric approach. The LSM and WLSM estimators are computed by making use of MATLAB built-in functions.

Simulation of nonparametric algorithms is also based on MATLAB. As a kernel $K(u)$, we use the standard Gaussian density. The bandwidths are defined in two ways. In accordance with [13], [2], [9], [8], the bandwidths are calculated by the cross-validation (CV) method. The second method of finding the bandwidths use the estimate $\overline{Y}_n$, based on the following empirical criteria:

$$h_{j,Empiric} = C_0\sigma_j n^{-\frac{1}{4+L}}, \quad C_0 = \operatorname{argmin}_{0<C<\infty}\left|Y_{n-1} - \frac{\sum\limits_{i=Q}^{n-1} Y_i K_L\left(\frac{U_n - U_i}{h}\right)}{\sum\limits_{i=Q}^{n-1} K_L\left(\frac{U_n - U_i}{h}\right)}\right|,$$

$j = \overline{1, L}$, where L is the dimension of function Ψ, σ_j^2 is the sample variance of observations for the j-th variable, $h = (h^y, h^x)$.

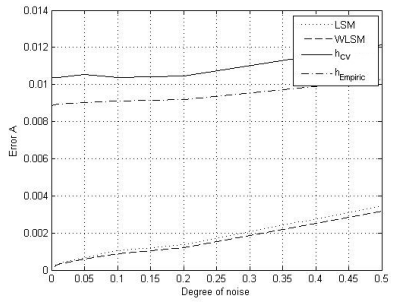

Figure 1: Averaged identification errors for M(1), $n = 50$

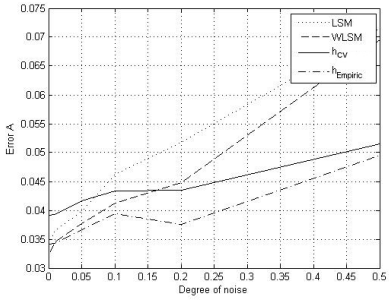

Figure 2: Averaged identification errors for M(2), $n = 50$

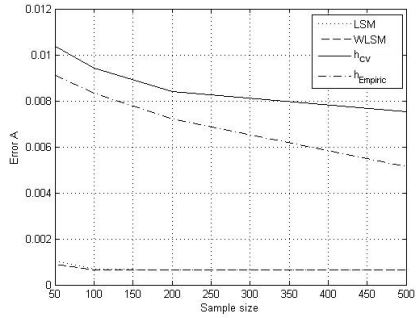

Figure 3: Dependence of the identification quality of M(1) on the size of observations for the noise level 0.15

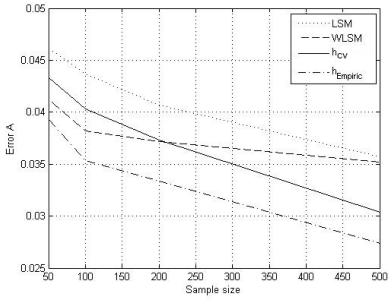

Figure 4: Dependence of the identification quality of M(2) on the size of observations for the noise level 0.15

For models M(1) and M(2) the values of the relative identification errors

$$A_{n,j} = \frac{1}{n-Q} \sum_{i=Q+1}^{n} \left| \frac{Y_i - \overline{Y}_i}{Y_i} \right|, \quad j = 1, 2,$$

$n = 50$, are presented in Figures 1 and 2. The results for other sizes of observations are shown in Figures 3 and 4. All the simulation results are averaged over 20 samples of the same size.

According to Figures the identification quality for the two models and all the methods decrease with increasing the level of noise. Further, for non-linear model M(2) nonparametric algorithms have advantages over parametric algorithms because of their adaptability. There is a tendency of reduction of identification errors for all models by increasing sizes of observations.

3 Real data processing

We examine the dependence of Russian Federation's Industrial Production Index (IPI) Y (see Figure 5) on the dollar exchange rate X^1, import X^2, and direct in-

vestments X^3 from September 1994 to January 2013. The data are available from http://www.gks.ru and http://sophist.hse.ru/. Apply (3) under $U_i = (Y_{i,1}, X_{i,4}) =$

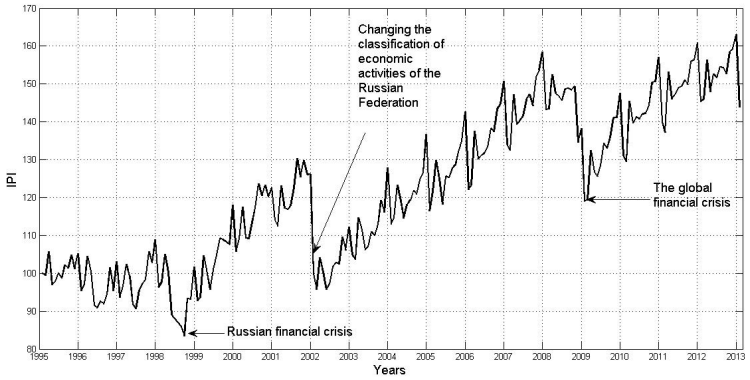

Figure 5: IPI for the period from January 1995 to January 2013

$(Y_{i-1}, X_i^1, X_i^2, X_i^3, X_{i-1}^3)$. Due to the fact that the classification principles of economic activities were changed in 2002, we consider two series of the data from September 1994 to December 2002 and from January 2003 to January 2013. We take the Gaussian density as the kernel $K(u)$, the bandwidths $h_j = 1.1\sigma_j n^{-1/9}$ for the data from September 1994 to December 2002, and $h_j = 1.23\sigma_j n^{-1/9}$ for the data from January 2003 to January 2013, $j = 1, 2, 3, 4, 5$, where the constants 1.1 and 1.23 are obtained by the above empirical criteria.

To compare the nonparametric algorithms (3) with the LSM-estimators, we have calculated the relative errors A_n and the relative average annual errors $A(t)$, $t = 1994, \ldots, 2013$, for both the approaches: $A_n = \dfrac{1}{n}\sum\limits_{i=1}^{n}\left|\dfrac{Y_i - \hat{Y}_i}{Y_i}\right|$, $A(t) = \dfrac{1}{12}\sum\limits_{i=1}^{12}\left|\dfrac{Y_i(t) - \hat{Y}_i(t)}{Y_i(t)}\right|$, where Y_i is the true value of the IPI and $\hat{Y}_i$ is its estimate. The results of such a comparison are given in Figure 6 and Figure 7.

The results of 1998 and 2009 can be explained by Russian financial crisis ("Ruble crisis") in August 1998 and Global financial crisis in 2009.

To predict the IPI Y for the data from 2002 to 2013 (cf. [12]), we apply (3) under

$$U_i = (Y_{i,1}, X_{i,4}) = (Y_{i-1}, X_{i-1}^1, X_{i-1}^2, X_{i-1}^3, X_{i-2}^3).$$

Here the bandwidths are equal to $h_{jt} = 0.94\hat{\sigma}_j t^{-1/9}$, where $\hat{\sigma}_j$, $j = 1, 2, 3, 4, 5$.

The similarity of identification algorithms and forecasting algorithms leads one to expect the both should behave similarly. For the relative average annual errors $A(t)$, seen in Figure 7 and Figure 8, one indeed observed that.

4 Control

Let the outputs Y of models M(1) and M(2) should be levelled to the given values $Y_*(k)$, $k = 1, \ldots, 5$ (see Figure 9 and 10) by making use of the controlled inputs X^2 and X^3 for the corresponding models.

For instance, outputs can be some qualities of produced goods and inputs – certain parameters of technology of production. Note, we can take direct investments as the controlled input in Section 3.

There is a problem of finding X^2_*, X^3_* for M(1), M(2), and X^3_* for the case of real data in Section 3. The variables X^1, X^2, and X^3 take values from uniform distributions on the corresponding intervals [2,2.1], [5,6], and [0,0.2] for M(1), and [2,2.1], [0,0.2], and [0,2] for M(2). Random variables ξ_n were generated from normal distributions with zero mean and variances, calculated by the formula (5).

The wanted outputs $Y_*(k)$ were taken as $Y_*(1) = \min\{Y_3, \ldots, Y_{50}\}$, $Y_*(2) = \dfrac{1}{48}\sum_{i=3}^{50} Y_i$,
$Y_*(3) = \max\{Y_3, \ldots, Y_{50}\}$, $Y_*(4) = (Y_*(1) + Y_*(2))/2$, $Y_*(5) = (Y_*(2) + Y_*(3))/2$.

In the case of model M(1), for example, analogously to (3) in accordance with [7], from the sequence of $(Y_i, Y_{i-1}, X^1_i, X^1_{i-1}, X^2_i, X^3_{i-2})$, $i = \overline{3, n}$, using the given value $Y_*(1n) = Y_*(1)$, $n = 51, \ldots, 70$, one can construct the corresponding estimates of $X^2_*(1n)$:

$$X^2_*(1n) = \frac{\sum\limits_{i=3}^{n} X^2_i K_2\left(\dfrac{Y_{*,2}(1n) - Y_{i,2}}{h_y}\right) K_2\left(\dfrac{X^1_{n,2} - X^1_{i,2}}{h_{1x}}\right) K\left(\dfrac{X^3_{n-2} - X^3_{i-2}}{h_{33}}\right)}{\sum\limits_{i=3}^{n} K_2\left(\dfrac{Y_{*,2}(1n) - Y_{i,2}}{h_y}\right) K_2\left(\dfrac{X^1_{n,2} - X^1_{i,2}}{h_{1x}}\right) K\left(\dfrac{X^3_{n-2} - X^3_{i-2}}{h_{33}}\right)}, \quad (6)$$

where $Y_{*,2}(1n) = (Y_*(1n), Y_{n-1})$, $Y_{i,2} = (Y_i, Y_{i-1})$, $X^1_{i,2} = (X^1_i, X^1_{i-1})$, $h_y = (h_1, h_2)$, $h_{1x} = (h_{11}, h_{12})$.

Here, we use the Gaussian kernel and bandwidths founded on the base of the empirical criteria as in Section 2. The relative control errors are defined by the formula

$$A(k) = \frac{1}{20}\sum_{i=51+(k-1)20}^{70+(k-1)20} \left|\frac{\hat{Y}_i - Y_*(k)}{Y_*(k)}\right|, \quad k = 1, \ldots, 5,$$

where, for example, $\hat{Y}_i = 0.2Y_{i-1} + 0.11X^1_i + 0.15X^1_{i-1} + 0.3X^2_*(1i) + 0.2X^3_{i-2}$, $i = 51, \ldots, 150$. The results are also obtained for model M(2) (see Table 1).

Acknowledgements

This investigation was supported by the Russian Foundation for Basic Research (project no. 13-08-00744).

Table 1: Errors of Control

Models	$M(1)$	$M(2)$
A_1	0.024	0.021
A_2	0.014	0.012
A_3	0.024	0.026
A_4	0.011	0.009
A_5	0.020	0.015

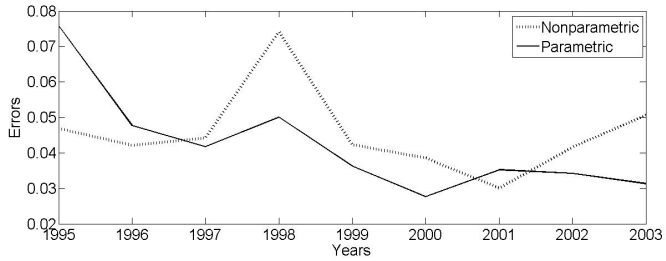

Figure 6: Identification relative errors $A(t)$

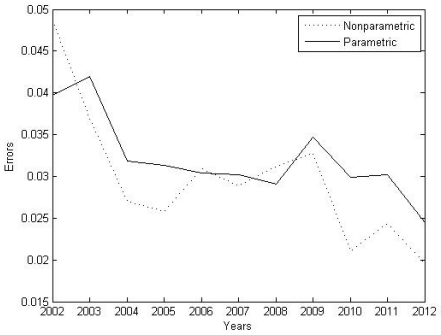

Figure 7: Identification relative errors $A(t)$

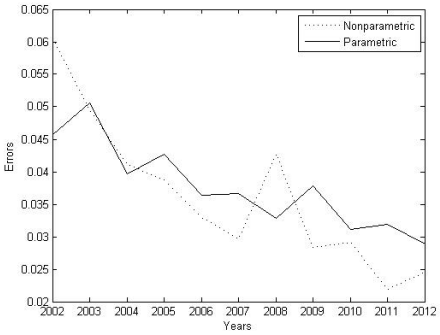

Figure 8: Forecasting relative errors $A(t)$

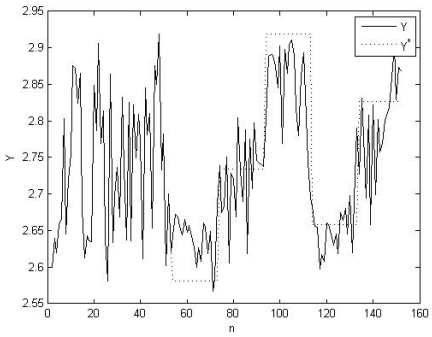

Figure 9: Control model $M(1)$

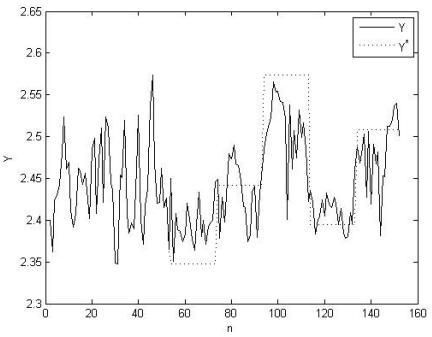

Figure 10: Control model $M(2)$

References

[1] A.V. Dobrovidov, G.M. Koshkin and V.A. Vasiliev. *Non-Parametric State Space Models.* Kendrick Press, Inc., Heber, UT 84032, USA, 2012.

[2] P. Hall. Asymptotic properties of integrated square error and cross-validation for kernel estimation of a regression function. *Z. Wahrscheinlichkeitstheorie Verw*, Geb. 67, 175–196, 1984.

[3] A.V. Kitaeva and G.M. Koshkin. Semi-recursive nonparametric identification in the general sense of a nonlinear heteroscedastic autoregression. *Automation and Remote Control*, 2, 92–111, 2010.

[4] A.V. Kitaeva and G.M. Koshkin. Nonparametric semirecursive identification in a wide sense of strong mixing processes. *Problems of Information Transmission*, 46, 1, 22–37, 2010.

[5] G.M. Koshkin. Deviation moments estimates of substitution and its piecewise-smooth approximations. *Siberian Mathematical Journal*, 40, 3, 605–618, 1999.

[6] G.M. Koshkin and I.Yu. Glukhova. Nonparametric identification of nonlinear ARX-processes. *Journal of Control and Computer Science. Tomsk State University*, 3, 20, 55–61, 2012 (in Russian).

[7] G.M. Koshkin and F.P. Tarasenko. Nonparametric algorithms for identifying and control of continuous-discrete stochastic objects. *8-th IFAC-IFORS Symposium on Identification and System Parameter Estimation.* Beijing: Pergamon Press, 2, 882–887, 1988.

[8] D. Leung. Cross-validation in nonparametric regression with outliers. *Annals of Statistics*, 33, 2291–2310, 2005.

[9] Q. Li and J.S. Racine. Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14, 485–512, 2004.

[10] E. Masry. Probability density estimation from sampled data. *IEEE Trans. Inf. Theory*, IT–29, 5, 696–709, 1983.

[11] E. Masry and D. Tjostheim. Nonparametric estimation and identification of nonlinear ARCH time series *Econometric Theory*, 11, 2, 258–289, 1995.

[12] V.A. Simakhin. Adaptive robust nonparametric prediction algorithms. *Journal of Control and Computer Science. Tomsk State University*, 1, 14, 45–54, 2011 (in Russian).

[13] C.J. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of Royal Statistical Society*, 36, 111–147, 1974.

Stochastic Models of Periodically Correlated Non-Gaussian Processes

Nina A. Kargapolova[1] and Vasily A. Ogorodnikov[1]
[1] *Institute of Computational Mathematics and Mathematical Geophysics, Siberian Brunch of Russian Academy of Sciences, Novosibirsk State University, Novosibirsk, Russia*
e-mail: nkargapolova@gmail.com, ova@osmf.sscc.ru

Abstract

Two algorithms for numerical modeling of piecewise constant periodically correlated non-Gaussian processes are proposed in this paper. Using first algorithm it is possible to obtain a process with time-independent distribution function. Second algorithm lets us to model a process with density of distribution that is a mixture of given densities and parameters of mixture are periodical functions of time.

Keywords: Inhomogeneous Markov chains, periodically correlated processes, piecewise constant random processes.

Introduction

In solution of various applied problems related to the study of actual time series (for example, meteorological or oceanologic) arises necessity of simulation of random processes with different periodical properties. Usually correlation function is periodic. There are two main approaches to modeling of periodically correlated processes. Vector autoregression processes are most often used for this purpose [1]. Second approach is based on some kind of point processes [3]. A new method for modeling of binary periodically correlated sequences is suggested in [2]. Such time series are simulated as inhomogeneous Markov chains with special type of inhomogeneity.

In this paper we suggest two algorithms for modeling of piecewise constant periodically correlated time series with given distribution densities. Both algorithms are based on inhomogeneous Markov chains studies in [2].

1 Modeling of processes with time-independent distribution function

Let $\xi_t, \quad t = 0, 1, 2, \ldots$be a binary inhomogeneous Markov chain with range space $M = \{0, \ 1\}$, initial distribution vector

$$A = \{a_1, \ a_0\} = \{a, \ 1 - a\}$$

and transition probability matrix $\Pi(t)$, that possesses 2 properties: $\Pi(t)$ is periodic function of discrete argument t, takes on a value

$$P_k = \begin{pmatrix} p_{11}[k] & p_{10}[k] \\ p_{01}[k] & p_{00}[k] \end{pmatrix} = \begin{pmatrix} p_k & 1-p_k \\ 1-q_k & q_k \end{pmatrix}$$

when $t = un+k > 0$, where $n \geq 2$ is a length of period. Here $p_{fg}[k]$ is a probability of transition from state f into state $g \quad (f, g \in \{0,\ 1\})$ when $t = un + k, \quad a_f$ is initial probability of state f.

Let's consider a process η_t, constructed via

Algorithm 1.

1. We simulate defined above Markov sequence $\xi_t,\ \ t = 0, 1, 2, \ldots$
2. Random variable ω with distribution function $F(x)$ is independently simulated, $\eta_0 = \omega$.
3. For $t \geq 1$:

if $\xi_t = \xi_{t-1}$: $\eta_t = \eta_{t-1}$;
if $\xi_t \neq \xi_{t-1}$: we independently simulate ω with distribution $F(x)$, $\eta_t = \omega$.

This process η_t is piecewise constant. It is obvious that distribution function of η_t is $F(x)$.

Example 1. Figure 1 shows a realization of Markov chain ξ_t and corresponding realization of η_t, in case if

$$n = 2,\ \ a = 0.5,\ \ P_0 = \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix},\ \ P_1 = \begin{pmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{pmatrix},\ \ F(x) = N_{0,1}.$$

Equality

$$corr(\eta_t, \eta_{t+h}) = P(\xi_t = \xi_{t+1} = \ldots = \xi_{t+h}).$$

holds for correlation function $corr(\eta_t, \eta_{t+h})$ of the process η_t. It means that correlation function $corr(\eta_t, \eta_{t+h})$ is completely specified by parameters of process ξ_t and is independent of $F(x)$. Due to Markov property of ξ_t value of $P(\xi_t = \xi_{t+1} = \ldots = \xi_{t+h})$ can be easy found, for example,

$$corr(\eta_{mn}, \eta_{mn+(un+v)}) = (b_n + (a-b_n)\, d_n^m) \left(\prod_{j=0}^{n-1} p_j^u \right) \left(\prod_{f=0}^{v-1} p_f \right) +$$
$$+ (1 - b_n - (a-b_n)\, d_n^m) \left(\prod_{j=0}^{n-1} q_j^u \right) \left(\prod_{f=0}^{v-1} q_f \right),$$

where

$$0 \leq v < n,\ \ b_n = \tfrac{c_k - q_0 t_n}{1-d_n},$$
$$t_n = \det \left(\prod_{i=1}^{n-1} P_i \right),\ \ d_n = \det \left(\prod_{i=0}^{n-1} P_i \right),\ c_n = \left\{ \prod_{i=1}^{n-1} P_i \right\}_{11}.$$

Function $corr(\eta_t, \eta_{t+h})$, as function of t, is oscillating. Since $|d_n| < 1$, then $corr(\eta_t, \eta_{t+h})$ converges to its limiting value, and this value is a periodic function of argument t.

Example 2. Correlation coefficients, as functions of time and shift, of the process η_t, defined in Example 1, are shown on Fig. 2.

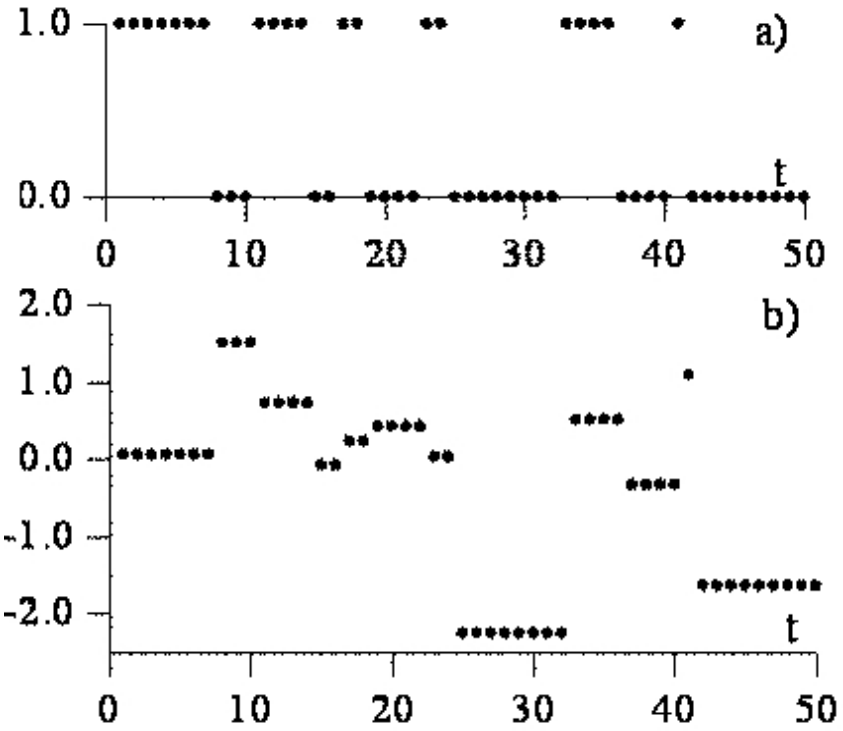

Figure 1: a) – realization of process ξ_t; b) – corresponding realization of η_t.

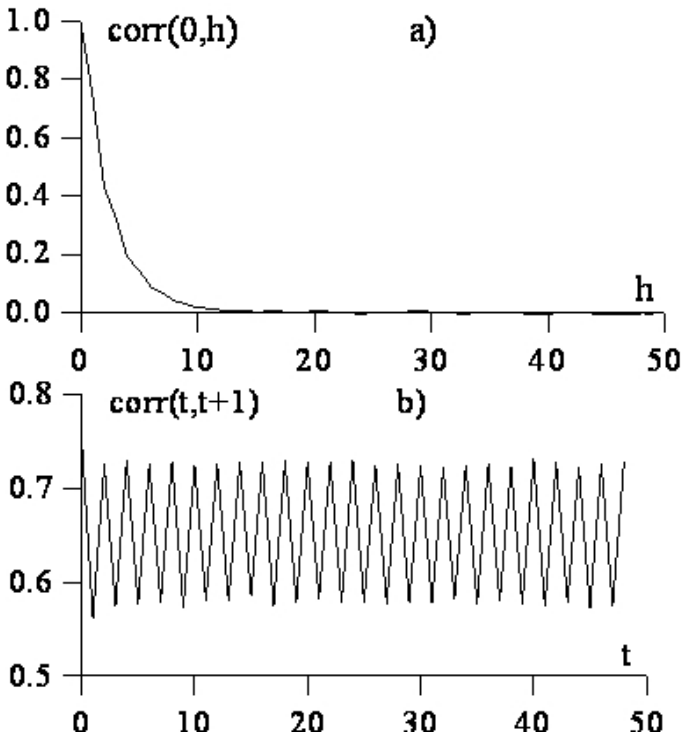

Figure 2: Correlation coefficients of process η_t as function of a) – shift; b) – time.

2 Modeling of processes with periodic distribution density

Now we are going to consider one more type of piecewise constant processes, based on described above Markov chains. Let $f_0(x), f_1(x), \ldots, f_{n-1}(x)$ be one-dimensional densities of distribution.

Algorithm 2.

1. We simulate defined above Markov sequence ξ_t, $t = 0, 1, 2, \ldots$
2. Random variable ω with distribution density$f_{n-1}(x)$ is independently simulated, $\upsilon_0 = \omega$.
3. For $t \geq 1$:

if $\xi_t = \xi_{t-1}$, then $\upsilon_t = \upsilon_{t-1}$;
if $\xi_t \neq \xi_{t-1}$, then we independently simulate ω with density $f_i(x)$, where $i \equiv t \pmod{n}$, $\upsilon_t = \omega$.

For process υ_t next statement can be proved.

Statement. Process υ_t has one-dimensional distribution density that can be described as mixture of $f_0(x), f_1(x), \ldots, f_{n-1}(x)$. Weights in his mixture are oscillating functions of time. When $t \to \infty$ weights converge to periodic functions and the period is equal to n. Correlation function, as function of time, oscillates and converges to periodic function with period n.

Proof of this statement is based on theory of covering runs in inhomogeneous Markov chains.

Example 3. It is possible to obtain theoretical formulas that describe distribution density of process υ_t for every t. Figure 3 shows these densities for 4 different moments$t$, when $n = 2$, $f_0 = \varphi_{0,1}$, $f_1 = \varphi_{5,1}$ and

$$a = 0.5, \; P_0 = \begin{pmatrix} 0.9 & 0.1 \\ 0.8 & 0.2 \end{pmatrix}, \; P_1 = \begin{pmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{pmatrix}.$$

Example 4. Figure 4 shows correlation coefficients $corr(\eta_t, \eta_{t+1})$ of process υ_t when $f_0 = \varphi_{0,1}$, $f_1 = \varphi_{5,1}$ (normal distribution) and parameters of ξ_t are showed below:

$$n = 2, \; a = 0.5, \; P_0 = \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix}, \; P_1 = \begin{pmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{pmatrix}.$$

Conclusion

Discussed algorithms let us to model piecewise constant processes with given properties of correlation function and one-dimensional density of distribution. Such processes can be used, for example, during investigation of meteorological time-series.

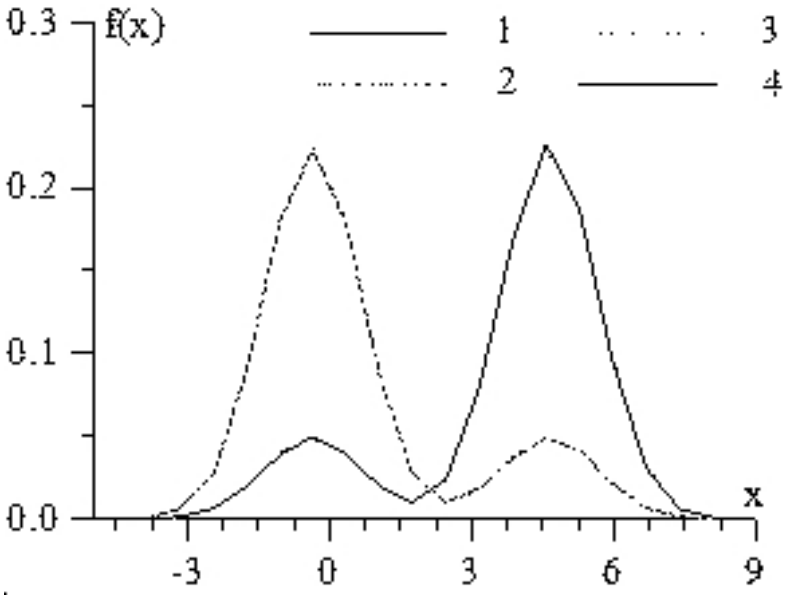

Figure 3: Probability density $f(x)$ of process υ_t when: 1 – $t = 20$, 2 – $t = 21$, 3 – $t = 24$, 4 – $t = 25$.

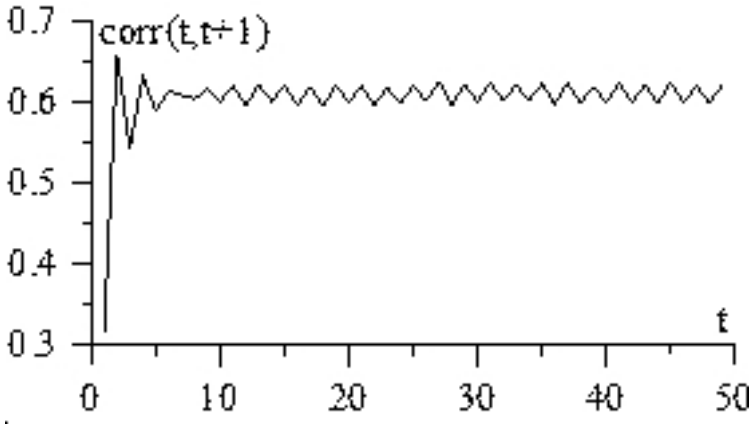

Figure 4: Correlation coefficients $corr(\upsilon_t, \upsilon_{t+1})$ of process υ_t.

Acknowledgements

This work is supported by Russian Foundation for Basis Research (grant 11-01-00641).

References

[1] Derenok K.V., Ogorodnikov V.A. (2008.)Numerical simulation of significant long-term decreases in air temperature. *Russian Journal of Numerical Analysis and Mathematical Modelling. Vol. 23, Iss. 3*, pp. 223-277.

[2] Kargapolova N.A., Ogorodnikov V.A. (2012).Inhomogeneous Markov chains with periodic matrices of transition probabilities and their application to simulation of meteorological processes. *Russian Journal of Numerical Analysis and Mathematical Modelling. Vol. 27 , Iss 3*, pp. 213-228.

[3] Ogorodnikov V.A., Prigarin S.M. (1996). On stochastic interpolation of discrete random processes and fields. *Russian Journal of Numerical Analysis and Mathematical Modelling. Vol. 11 , Iss. 1*, pp. 49-69.

Nonparametric Algorithm of Nonlinear Dynamical Systems Identification

N.V. Koplyarova[1] and N.A. Sergeeva[1]
[1] *Siberian State Aerospace University*
Krasnoyarsk, Russian Federation
e-mail: koplyarovanv@mail.ru, sergena@list.ru

Abstract

The problem of nonlinear dynamical systems of Wiener and Hammerstein type identification is considered. The linear dynamical part of the system is in nonparametric uncertainty conditions. The common type of nonlinearity is assumed to be known with the set of parameters. Presented algorithm allows to create the adequate in the sense of mean-square criterion models. The proposed method of dynamic objects modeling is based on the nonparametric estimation of linear and nonlinear parts of the system.

Keywords: nonlinear system, nonparametric, Wiener and Hammerstein models.

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. 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 or Hammerstein

type [1]. 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.

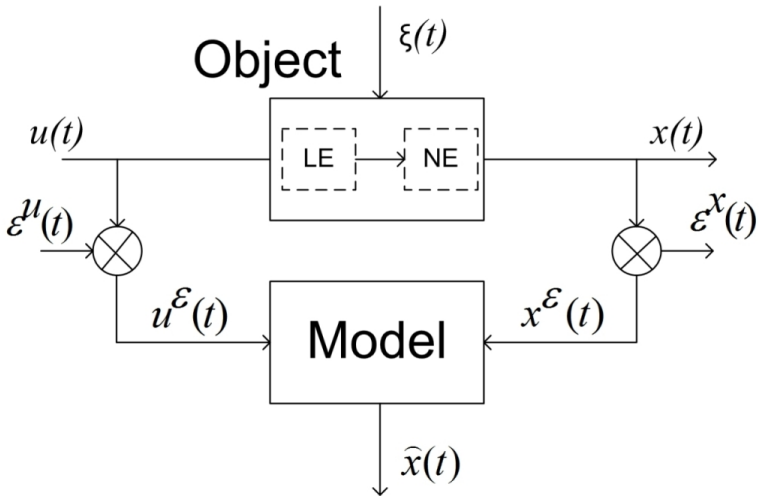

Figure 1: The general scheme of the identification problem

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 objects measures of s size.

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. The following nonlinear elements are considered in the paper:

- quad

$$x(t) = aw(t)^2. \tag{1}$$

- link saturation (with saturation parameter b1):

$$x(t) = \begin{cases} w(t), & \text{if } w(t) < b_1; \\ b, & \text{if } w(t) > b_1; \\ -b, & \text{if } w(t) < -b_1. \end{cases} \tag{2}$$

Problem of nonlinear system identification can be divided into two tasks. At first part we consider the identification of a linear element.

2 Nonparametric model of linear dynamical system

The reaction of linear dynamical system $x(t)$ to the input signal $u(t)$ is described with the Duhamel integral[2]:

$$x(t) = k(0)u(0) + \int_0^t h(t-\tau)u(\tau)d\tau = k(0)u(0) + \int_0^t k\prime(t-\tau)u(\tau)d\tau \qquad (3)$$

where $h(t)$-impulse response (weight function) of the system, and $k(t)$-step response of this system (transient function).

In this case we can calculate the output value of object $x(t)$ only if its weight function $h(t)$ is known. But in practice, as a rule, it is impossible to get the weight function of the object. Therefore we can write the estimation of step response of the system as a stochastic approximation of regression as follows:

$$k_s(t) = \frac{1}{sC_s}\sum_{i=1}^{s} k_i H\left(\frac{t-t_i}{C_s}\right) \qquad (4)$$

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].

$$c_s > 0, s = 1, 2..., \lim_{s\to\infty} c_s = 0, \lim_{s\to\infty} cc_s = \infty$$

$$\int_{\Omega(u)} H\prime(u)du = 0, c_s \int_{\Omega(u)} H\prime(u)udu = -1, u = \frac{\tau - t}{c_s} \qquad (5)$$

$$\lim_{s\to\infty} c_s^{-1} H\left(\frac{\tau-t}{c_s}\right) = \delta(\tau - t)$$

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:

$$h_s(t) = k\prime(t) = \frac{1}{sC_s}\sum_{i=1}^{s} k_i H\prime\left(\frac{t-t_i}{C_s}\right) \qquad (6)$$

Linear dynamic system can be described with the following mathematical formula:

$$x_s(t) = \frac{1}{sC_s}\sum_{i=1}^{s}\sum_{j=1}^{t/\Delta\tau} k_i H\prime\left(\frac{t-t_i-\tau_j}{C_s}\right) u(\tau_j)\Delta\tau \qquad (7)$$

Then consider the algorithm of nonlinear system identification.

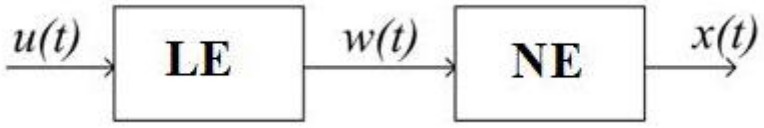

Figure 2: Wiener model, LE linear dynamical and NE - nonlinear parts of the system, $u(t)$, $x(t)$- input and output action, $w(t)$ - intermediate part output(is not measured)

3 Nonparametric model of Wiener type

Lets consider the system that can be represented as a model of Wiener type (Fig. 2.)[5]

If the structure of the nonlinear element in the model is known with a set of parameters, then the object output is calculated as a function of the Duhamel integral:

$$x(t) = f(w(t), \alpha) \tag{8}$$

where $w(t)$ - the linear part output (is not measured); $f()$ - nonlinear function.

The mathematical model of the nonlinear object can be represented as a set of equations (7), (8), in that, instead of the weight function $h(t)$ and the parameters α are used their statistical estimations. To obtain this estimation it is necessary to generate the sample $\{u_i, w_i, i = 1, s\}$. In the case when for some classes of nonlinear elements, the equation (8) can be solved for $w(t)$,we have [4]:

$$w(t) = f^{-1}(x(t), \alpha) \tag{9}$$

In this case, the nonparametric model of nonlinear object is the following:

$$\hat{x}(t) = f(\hat{w}(t), \alpha) \tag{10}$$

$$\hat{w}(t) = \int_0^t \hat{k\prime}(t-\tau)u(\tau)d\tau \tag{11}$$

Thus, we get the algorithm for modeling of Wiener type nonlinear dynamical systems.

a Identification of nonlinear systems 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 $x(t) = ak(t)^2$.

That is, the step response of a linear element can be represented by the output of the process as follows:

$$\hat{k}(t) = \sqrt{x(t)/a}) \tag{12}$$

For an arbitrary input action the output of linear part of the system is described by the equation (7). 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 $x1_i$ - 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):

$$2x\prime\prime(t) + 0.3x\prime(t) + 1.5x(t) = u(t)$$

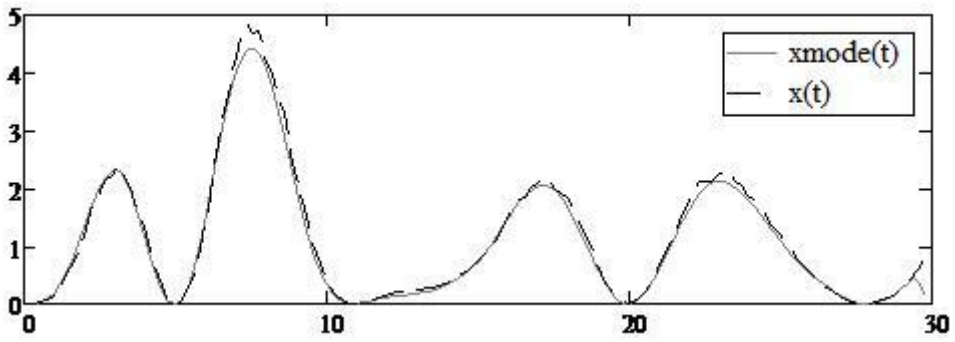

Figure 3: The model of system: $xmodel(t)$ - a model of a nonlinear system, $x(t)$ - the system output, sample size s = 250, sampling interval $h = 0.2$, noise 5%, input action $u(t) = 2\cos(0.4t)$, error 1.7%

b 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 (2). 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 experimentally by several static experiments [4]. We get the following algorithm to construct the model:

1. to carry out some experiments (m) under the following conditions: input actions $u_{ij} = c_j$, $c_j = const$. The result is a sample $\{u_{ij}, x_{ij}\}$, $i = 1, s$, $j = 1, m$.

2. to find the distance between two consecutive measurements: $h1_j = |x_{ij} - x_{i-1j}|/h$, where h - sampling interval.

3. $\hat{b} = x_{ij}$, if $h1_k < \varepsilon, \varepsilon > 0$.

4. if $x_{ij} = \hat{b}$, then $x_j = y_{i-1j}$, $\hat{a} = M\{a\}$

- 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.

- 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, b1=1.5) and the differential equation (simulating object):

$$7.4x''(t) + 2.5x'(t) + 2.43x(t) = u(t)$$

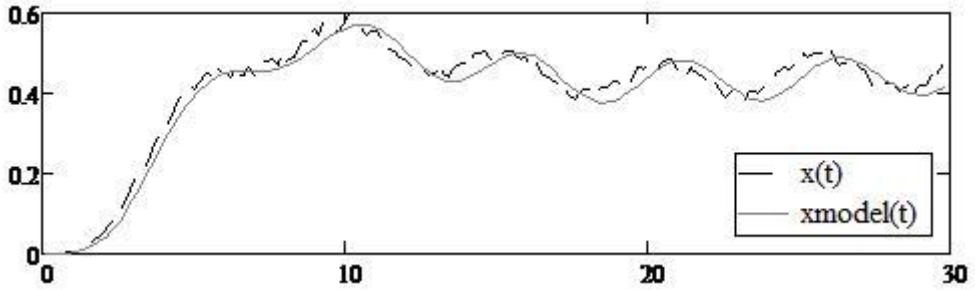

Figure 4: The model of system: $xmodel(t)$ - a model of a nonlinear system, $x(t)$ - the system output, sample size s = 250, sampling interval $h = 0.15$, noise 5%, input action $u(t) = 0.9 \sin 2(0.6t)$, the relative average error of simulation 5.2%

4 Nonparametric model of Hammerstein type

Lets consider the system that can be represented as a model of Hammerstein type (Fig. 5.)

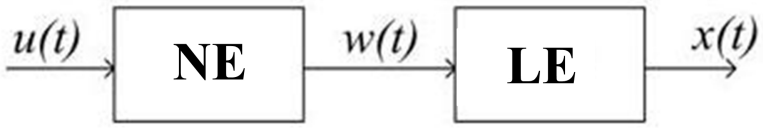

Figure 5: Hammerstein model, LE linear dynamical and NE - nonlinear parts of the system, $u(t)$, $x(t)$- input and output action, $w(t)$ - intermediate part output(is not measured)

In this case the results of the nonlinear element output measurement are not available. It is assumed that the parameterized structure of linear element (LE)

is unknown, and the nonlinear characteristic is known with a set of parameters. According to Figure 5 the relationship between the input $u(t)$ and output $x(t)$ of the object with zero initial conditions can be described by equations of the following form [1,4]

$$w(t) = f(u(t)) \tag{15}$$

$$\hat{x}(t) = \int_0^t \hat{h}(t-\tau)w(\tau)d\tau \tag{16}$$

Or, excluding variable $w(t)$,

$$\hat{x}(t) = \int_0^t \hat{h}(t-\tau)f(\alpha, u(\tau))d\tau \tag{17}$$

where $h(t)$-impulse response of the dynamic element, $f(\alpha, u)$ - nonlinear function with unknown parameters.

Let $x1(t)$- response of nonlinear object to the input signal in the form of the Heaviside function $u(t) = 1(t)$, $x(t)$ - the reaction of the object on some test signal. The measurements of signals $x1(t)$ and $x(t)$ in discrete time are formed in samples of observations $\{1, x1_i\}$ and $\{u_i, x_i\}$, $i = 1, s$ respectively. The reaction of nonlinear element to the step input $u(t) = 1$ is also a step function, but the amplitude of signal is changed. In this case $w1(t) = f(\alpha, 1(t)) = c1$, where $c1$-constant. The nonlinear system output $x1(t) = \int_0^t h1(t-\tau)1(\tau)d\tau$ can be regarded as a transfer function of a linear system with the impulse function $h1(t)$. The estimation of the impulse function can be obtained from the sample $\{x1_i, t_i\}$, $i = 1, s$.

The nonlinear system model (17) according to the estimation $\hat{h}1(t)$ can be written as:

$$\hat{x}(t) = \int_0^t \hat{h}(t-\tau)f(\alpha, u(\tau))d\tau \tag{18}$$

The parameters α are estimated on the base of the sample $\{u_i, x_i\}$, $i = 1, s$ of the input and output objects measurements as a solution of the extremal problem. Then the nonparametric model of Hammerstein system can be described with the following formula:

$$\hat{x}(t) = \frac{1}{sC_s}\sum_{i=1}^{s}\sum_{j=1}^{t/\Delta\tau} k1_i H\prime\left(\frac{t - t_i - \tau_j}{C_s}\right) f(u(\tau_j))\Delta\tau \tag{19}$$

where $k1$ - step response of the linear element, $\hat{f}(t)$ - estimation of the nonlinear element.

Example. Consider a nonlinear dynamical system of Hammerstein type that consisting of a link saturation (with parameters $a = 1, b = 2$) and the differential equation (simulating object):

$2x\prime\prime(t) + 0.3x\prime(t) + x(t) = u(t)$

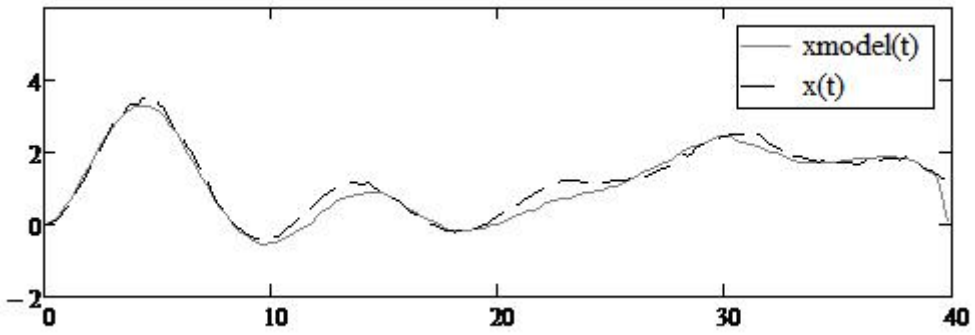

Figure 6: The model of system: $xmodel(t)$ - a model of a nonlinear system, $x(t)$ - the system output, sample size s = 150, $h = 0.2$, noise 5%, $u(t) = e^{\cos(0.2t)}$, error of simulation 4.1%

Having made the analysis of the nonparametric model of the nonlinear dynamic object with the quad and link saturation nonlinearity, we can say that the model adequately describes systems in terms of noise communication channels at different sample sizes and different input actions.

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 and Hammerstein 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 and Hammerstein type in the cases of quad and link saturation nonlinearity.

Acknowledgements

Authors would like to express the sincere gratitude to their scientific advisor Professor Medvedev A.V. for the continuous support of their research, for his patience, motivation, enthusiasm, and immense knowledge.

References

[1] Chaika S.N. (1989). *Identification of dynamic systems with partially parameterized model structure.*Izd Gorky State. University Press, Gorky.

[2] Eykoff P. (1975). *The basis of control systems identification.* Springer-Verlag.

[3] Medvedev A.V., (1983). - *Nonparametric system adaptation.* Nauka, Novosibirsk.

[4] Medvedev A.V., (1979). - *Nonparametric algorithms of nonlinear dynamical systems identification. / Sat. Stochastic control system..* Nauka, Novosibirsk.

[5] Popkov U.S. (1976). *Identification and optimization of nonlinear stochastic systems,* -. Pub. M: Energiya, Moscow.

About Data Analysis In Non-Parametric Identification Problem

A.A.Korneeva[1], N.A.Sergeeva[1]and E.A.Chzhan[2]
[1] *Siberian State Aerospace University, Krasnoyarsk, Russia Federation*
[2] *Siberian Federal University, Krasnoyarsk, Russia Federation*
e-mail: anna.korneeva.90@mail.ru, ekach@list.ru

Abstract

In this paper a new class of stochastic processes with components dependence of the input (H-processes) is discussed. Standard modeling approach may not be effective in conformity to the H-process. Therefore, a new class of models is introduced. They differ from the well-known by indicator functions. Numerous computational experiments show the high efficiency of simulation of discrete-continuous processes "tubular" type.

Keywords: nonparametric and parametric identification, "tubular" processes, H-model, an indicator function, a priori information, noninertial processes.

Introduction

Discrete and continuous processes prevail in industry and manufactures with continuous technological types of the process. This type of processes is typical for objects of ferrous and non-ferrous metallurgy, construction, power engineering, refining and etc.

Certainly the problem of modeling and identification remains one of the central problems of cybernetics for a long time. While formulating the identification and control problems the level of a priori information is of the special emphasis. It depends on both a priori knowledge of the process, available control tools and the technology of variable measurement.

Another peculiarity of building adaptive and learning models while developing systems of automation and control by technological processes is stochastic dependence of the vector component of input variables. It leads to the fact that in the space of "input-output" variables determined by the technological regulations a real process has a "tube" structure. It is necessary to enter an indicator of a special form to build adaptive models of discrete and continuous processes. The investigations showed that even in the case of a great number of variables describing a manufacturing process it is possible to build a learning model with samples existence of a small volume.

1 Problem Formulation

The general scheme of the researched process is shown in Fig.1.

On Fig.1 the table of symbols are accepted: A is an unknown object operator, $x(t) \in \Omega(x) \subset R^1$ is an output variable of the process, $u(t) = (u_1(t), u_2(t), ..., u_m(t)) \in$

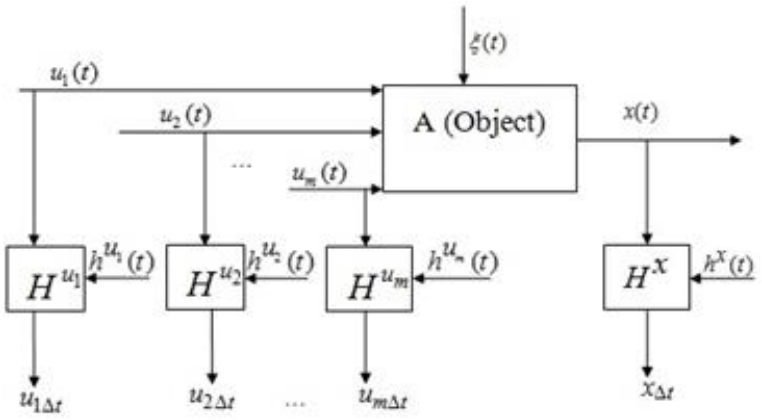

Figure 1: The General Scheme of the Researched Process and Tools of Control

$\Omega(u) \subset R^m$ is a control action, $\xi(t)$ is a vector random action, (t) is continuous time, H^u, H^x are channels of connection corresponding to different variables and including control tools, $h^u(t), h^x(t)$ are random noises of measurements corresponding to variables of the process with zero means and limited variance.

2 Identification in the "Narrow" and "Wide" Senses

When modeling of various discrete-continuous processes the identification theory in the "narrow" sense [1] dominates now. Its essence is that at the first stage the parametric class of operators A^α is defined on the basis of available a priori information, for example:

$$\tilde{x}_\alpha(t) = A^\alpha(u(t), \alpha). \tag{1}$$

where A^α is a parametric structure of the model, and α is a vector of parameters. At the second stage the estimation of parameters is realized on the basis of an available sample $\{x_i, u_i, i = \overline{1,s}\}$, s is a sample volume. The estimation of parameters can be executed with numerous recurrent procedures, in particular with a method of stochastic approximations or ordinary least squares [2]. In this case the success of the identification problem solution essentially depends on that how "successfully" the operator A^α is defined. Now the theory of parametric identification is the most developed [2].

The identification in the "wide" sense assumes the absence of a choice of a parametric class of the operator. Often it is much more simply to define a class of operators on the basis of qualitative character data, for example, linearity of the process or type of nonlinearity, etc. In this case the identification problem consists of the estimation of this operator on the basis of a sample $\{x_i, u_i, i = \overline{1,s}\}$ in the shape:

$$\tilde{x}_s(t) = A_s(u(t), \overrightarrow{x_s}, \overrightarrow{u_s}). \tag{2}$$

where $\overrightarrow{x_s} = (x_1, x_2, ..., x_s)$, $\overrightarrow{u_s} = (u_1, u_2, ..., u_s)$ are temporary vectors. The estimation of the operator A_s can be realized with means of nonparametric statistics [3]. It is remarkable that there is no stage of a choice of parametric. Thereby it is possible to affirm that in this case the identification (and it is the identification in the "wide" sense) is more adequate for real problems.

3 Nonparametric Estimates of the Regression Function on Observations

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 [4] are used for the backing up $\tilde{x} = M\{x|u\}$

$$x_s(u) = \sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi(c_s^{-1}(u^j - u_i^j)) / \sum_{i=1}^{s} \prod_{j=1}^{m} \Phi(c_s^{-1}(u^j - u_i^j)), \tag{3}$$

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 [4]. In this case the triangular kernel was used as the bell-shaped function $\Phi(c_s^{-1}(u^j - u_i^j))$, $i = \overline{1,s}$, $j = \overline{1,m}$:

$$\Phi(c_s^{-1}(u^j - u_i^j)) = \begin{cases} 1 - |c_s^{-1}(u^j - u_i^j)|, if c_s^{-1}(u^j - u_i^j) \le 1, \\ 0, if c_s^{-1}(u^j - u_i^j) > 1. \end{cases} \tag{4}$$

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 isnt considered in the model:

$$R(c_s) = \sum_{k=1}^{s} (x_k - x_s(u_k, c_s))^2 = \min_{c_s}, k \neq j. \tag{5}$$

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.

4 H-models

In practice we often deal with the processes having the stochastic dependence among components of input variables. Let's say that objects with this feature have "tubular" structure. Let's consider a process with "tubular" structure presented on Fig. 2. As we can see on the figure, the range of the process progress $\Omega(u,x)$ corresponds to the unit hyper cube where $u \in R^2$, $x \in R^1$.

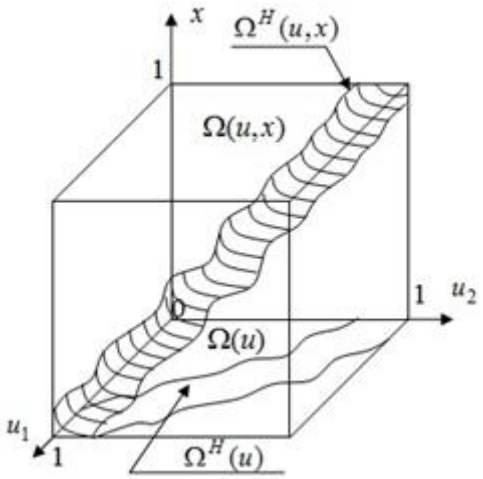

Figure 2: The Object with the "Tubular" Structure

However if the researched process has a "tubular" structure, the range of the process progress is limited not only in the hyper cube space $\Omega(u, x)$ but also its subrange $\Omega^H(u, x) \in \Omega(u, x)$ which is never known to us. As $\Omega^H(u, x)$ is never known, the existence of the "tubular" structure isn't known to us also. And it is necessary to notice that the hyper cube volume, as we can see on the figure above, exceeds the "tube" volume appreciably. Let's consider the modeling of processes having this structure. Usually in the problem of the inertia less object identification some parametric model, representing a surface in the space of "input-output" variables, is supposed to be available:

$$\hat{x}(u) = \hat{f}(u, \alpha_s). \tag{6}$$

where α_s is a vector of parameters. In that case when components of the input vector are statistically dependent, i.e. we deal with "tubular" structure object, it is necessary to set the indicator $I_s(u)$. The model (1) should be corrected as follows:

$$\hat{x}(u) = \hat{f}(u, \alpha_s) I_s(u), \tag{7}$$

where the indicator is identified according to the rule:

$$I_s(u) = sgn(sc_s)^{-1} \sum_{i=1}^{s} \Phi\left(c_s^{-1}(x_s(u) - x_i)\right) \prod_{j=1}^{k} \Phi\left(c_s^{-1}(u^j - u_i^j)\right), \tag{8}$$

where

$$x_s(u) = \sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi\left(c_s^{-1}(u^j - u_i^j)\right) \Bigg/ \sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(c_s^{-1}(u^j - u_i^j)\right), \tag{9}$$

where the kernel function $\Phi\left(c_s^{-1}(u^j - u_i^j)\right)$ and the smoothing parameter have some convergence properties [2].

The logic of this indicator is that at any chosen value of the current variable $u = u'$ indicator $I_s(u)$ takes the value of one if the value u' belongs to the "tubular" structure defined by the available sample $\left\{x_i, u_i, i = \overline{1,s}\right\}$ and if u' taken to a value outside the "tube", the indicator is zero. Note that if a process is described by the surface in space, the model (1) and (2) are the same. If the process has a tubular structure in this space, you need to use the model (2).

5 The Computing Experiment of H-Model

Let's consider results of the numerical experiments. Let the researched object be described by the system of the equations:

$$\begin{cases} x = 0.7u_1 + 0.3u_2 + \xi, \\ u_2 = u_1 + \psi, \end{cases} \tag{10}$$

where ξ and ψ are random numbers uniformly distributed over on a symmetric interval $[-0.05, 0.05]$, $u_1, u_2 \in [0; 3]$. In this case the equation of the object is specified to identify initial samples of "input-output" variables. When building the model on the basis of the samples, the structure of the dependence of the output variable x from the input variables u is accepted with parameters. In the parameters estimating the method of least squares is used. So, the sample of statistically independent observations $\left\{x_i, u_i, i = \overline{1,s}\right\}$ is discovered, where x is the measured output variable, $u = (u_1, u_2)$ is the vector control action, s is the sample size. We build the parametric model (1) of the object. The modelling results are shown on the Fig.3 ($s = 100$).

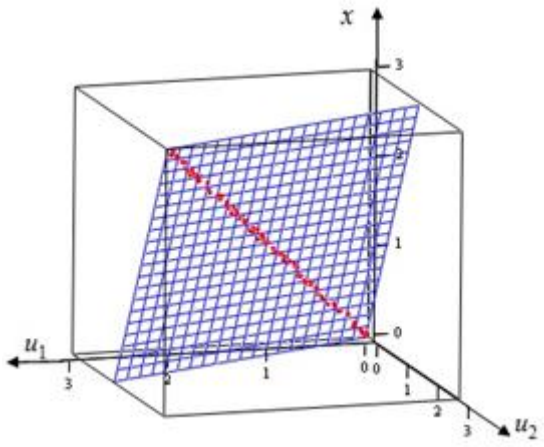

Figure 3: The Object with a "Tubular" Structure and its Parametric Model

On the Fig. 3 the red dots denote the researched object, blue - the resulting parametric model and the black cube variables values determined by the technological regulations. As we can see on the figure, the researched object is a "tube." The model (1) is the plane lying on the object. As we know, a straight, in this case, "tube" can

be approximated by an infinite number of planes. Suppose there are 5 samples of statistically independent measurements $\{x_i, u_{1i}, u_{2i}, i = \overline{1,s}\}$ of size $s = 100$. For each case model (7) was used. So we have 5 models. Modelling results are shown on Fig. 4.

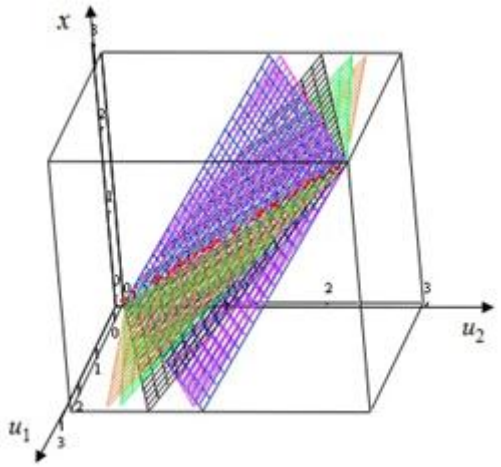

Figure 4: The Set of Parametric Models of the Object With a "Tubular" Structure

Now, for the object (10) will be used the model (8) containing the indicator. Type of indicator is described by (9). The results of numerical simulation of the object (8) for the sample size $s = 1000$ are shown in Fig. 5.

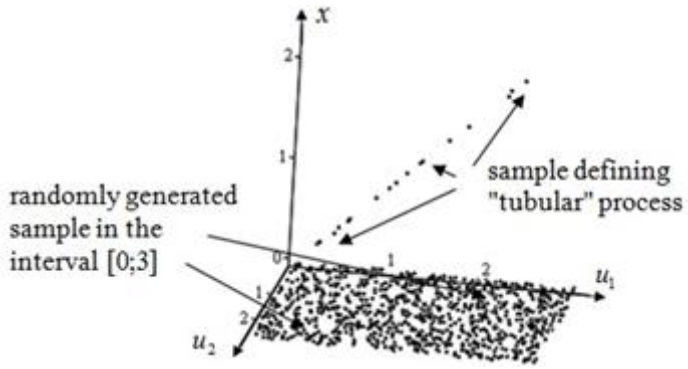

Figure 5: Sample Measurements of the "Input-Output" Variables

As we see, when constructing the model, the indicator function took into account only those points which belong to the space of the "tubular" process progress, i.e. $I_s(u) = 1$. In other points of the sample the output value wasn't recovered, i.e. $I_s(u) = 0$. In particular, only 20 points from the total sample size $s = 1000$ belong to the tube space.

In some cases, a point $x_s = 0$ can belong to "tubular" object, so it is advisable to assume that the points at which the value of the indicator is zero, the value of the model output is not reconstructed at such points.

Let's change the system of equations, which describes the researched object, on the following one:

$$\begin{cases} x = 2\sin u_1 + 0.5u_2 + \xi, \\ u_2 = u_1 + \psi, \end{cases} \tag{11}$$

The results of similar experiments are shown on the following figures:

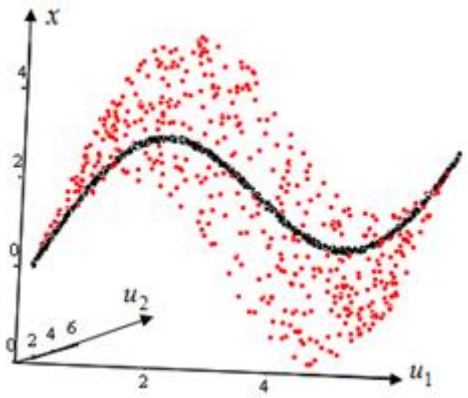

Figure 6: Parametric Model of the "Tubular" Process

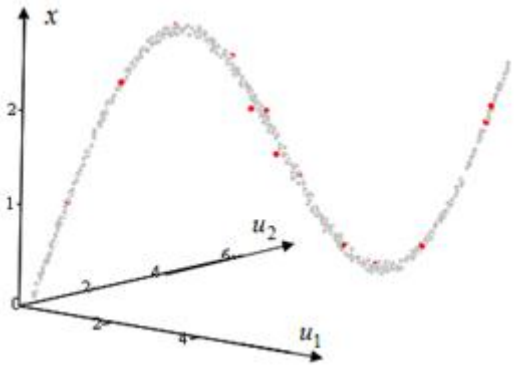

Figure 7: Parametric Model with Indicator of the "Tubular" Process

In fig. 6 the object is described by the system (11). It has the "tubular" structure (black dots on the color graph). The plane (Fig. 6) shows the parametric model recovered using OLS.

Also the parametric model with the indicator function (8) is shown (figure 7). As in the previous experiment it shows that the indicator considers only those points that belong to the "tube" space ($I_s(u) = 1$), in other points the indicator equals to zero and in future, these points are not considered in the model construction.

Conclusions

We offer H-models with dependent input variables ("tubular"). The domain of these processes is always unknown but it must be defined while modelling. H-models differ from the generally excepted models of noninertial systems by the indicator function existence which defines the domain of the "tubular" process progress. The offered method of data analysis and adaptive H-models can be applied while designing and developing different computer systems of automation processes in manufacturing branches..

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] Tsipkin Y.Z. (1984). *Fundamentals of information theory of identification* . Nauka, Moskow.

[2] Eykhoff P. (1975). *Identification fundamentals control systems*. Mir, Moskow.

[3] Medvedev A.V. (1983). *Nonparametric systems of adaptation* . Nauka, Novosibirsk.

[4] Nadaraya E.A (1983) *Nonparametric estimation of the probability density and the regression curve*.Izdatelstvo Tbiliskogo yniversiteta, Tbilisi.

A Regression Model for Prices of Second-Hand Cars

Artyom Kovalevskii[1]
[1] *Novosibirsk State Technical University,*
Novosibirsk State University,
Novosibirsk, Russia
e-mail: kovalevskiii@gmail.com

Abstract

We construct and test a regression model for prices of second-hand cars *Toyota Camry* in Novosibirsk, Russia. Data are from a local internet site ngs.ru. The single significant factor for cars with right wheel is a year of production. Our statistical criterium is based on an empirical bridge of regression residuals. It proves different model parameters for 1991-99 and 2000-08 years of production. This approach gives an algorithm of estimation of car price.

Keywords: regression model, significant factor, empirical bridge.

Introduction

We suggest a linear model for logarithms of car price against regressors. Kuiper [5] used a similar model for data from www.kbb.com. Unlike of [5], our approach contains correspondence analysis on a base of an empirical bridge approach [3], [4]. Statistical test is based on weak convergence of the empirical bridge to a gaussian process in $C(0,1)$ (see [1]).

Section 1 contains empirical analysis. We introduce and use the statistical test in Section 2.

1 Empirical Data Analysis

We analyse ads about sales of Toyota Corolla cars at www.ngs.ru on 02.06.2012. There are 525 ads. We explore a regression of logarithm of a sale price against a date of the ad, a steering wheel position (left or right), a year of production, an engine volume, gearboxes type, milage. Standard regression analysis gives p-values lesser than 0.01 only for a steering wheel position and a year of production. The number of cars with a left wheel is relatively small, so we choose right-wheeled cars (382 ads). We investigate dependence of production year Y_i against production year X_i. Ads are ordered by the year. The order is random for cars of a same year.

The model is (see Pic. 1)

$$\ln Y_i = aX_i + b + \varepsilon_i,\ i = 1, \dots, n. \tag{1}$$

Here ε_i are independent and identically distributed, have zero mean and non-zero finite variance. Estimations of a and b are approximately (all values are up to 4

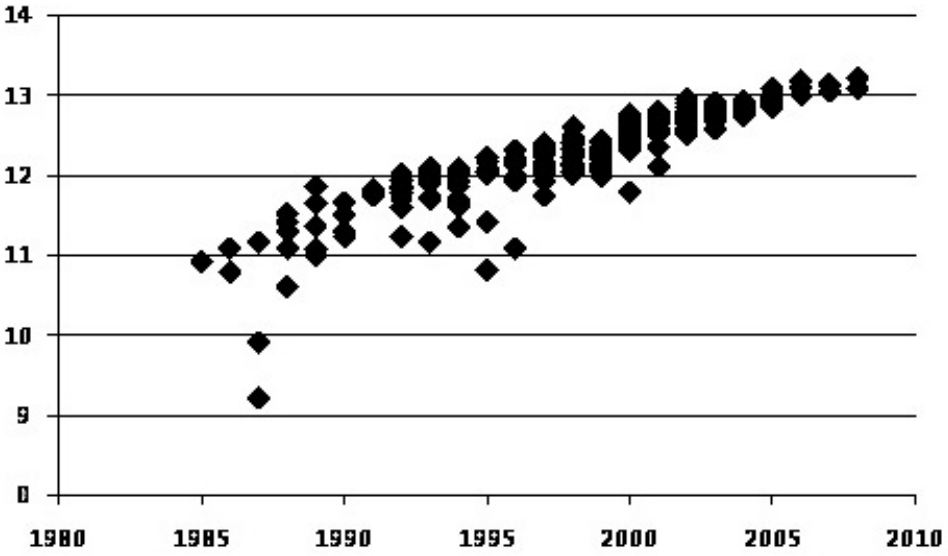

Figure 1: Dependence of $\ln Y_i$ against X_i for right wheels, 382 ads

numbers)

$$\widehat{a} = 0.1089,\ \widehat{b} = -205.3.$$

We estimate Y_i and calculate regression residuals. The sample standard deviation of regression residuals is $S = 0.2469$.

Then we delete consequently ads with regression residuals that are absolutely greater then 3-multiplied sample standard deviation (which is recalculated after each ad deletion). 364 ads remains after deletion, parameters estimations for it

$$\widehat{a} = 0.09558,\ \widehat{b} = -178.7,\ S = 0.1291.$$

We have decreased S almost twofold.

We calculate an empirical bridge of regression residuals as in [3], [4]. It is in Pic. 2. There are years at the absciss axis for clarity.

The basic hypothesis is model (1). As distribution of an empirical bridge is close to distribution of a standard brownian bridge under this hypothesis, we neglect the basic hypothesis on level 0.01 on a base on maximal deviation of the empirical bridge on Pic.2 from an absciss axe. Therefore we propose a new hypothesis: model (1) for each of intervals between sharp peaks of the empirical bridge, that is, for intervals from 1 to 16, from 17 to 154, from 155 to 364. We don't analyse points from 1 to 16 due to a small number of points. We slightly correct intervals to correspond production years. Interval from 17 to 148 corresponds to years from 1991 to 1999, and interval from 149 to 364 corresponds to years from 2000 to 2008.

We analyse each of these two intervals. For the first one

$$\widehat{a} = 0.06484,\ \widehat{b} = -117.3,\ S = 0.1356.$$

For the second one

$$\widehat{a} = 0.07144,\ \widehat{b} = -130.3,\ S = 0.08699.$$

We analyse the suggested models in the next section.

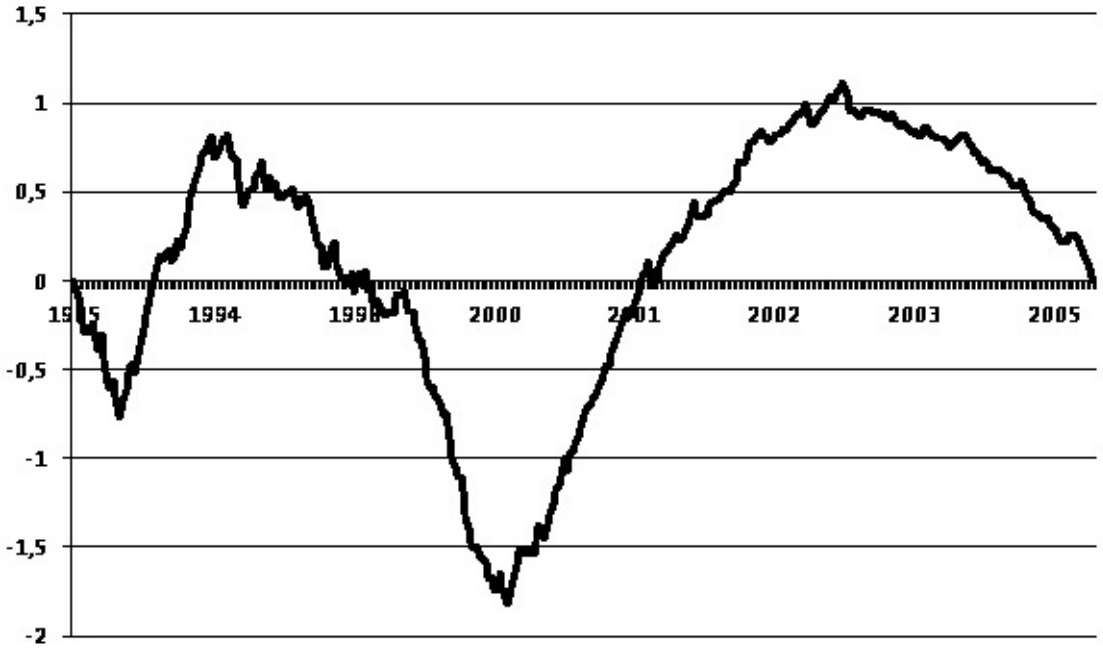

Figure 2: The empirical bridge for regression residuals (364 ads)

2 Statistical Testing of Models

Let $\widehat{Y}_i = \widehat{a} + \widehat{b}X_i$, $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$, $\widehat{\Delta}_i^0 = \widehat{\varepsilon}_1 + \ \ldots + \ \widehat{\varepsilon}_i$.

The empirical bridge is a random polygon $\widehat{Z_n}$ with nodes

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta}_k^0 - \frac{k}{n}\widehat{\Delta}_n^0}{\sqrt{\widehat{\sigma^2}n}}\right) = \left(\frac{k}{n},\ \frac{\widehat{\Delta}_k^0}{\sqrt{\widehat{\sigma^2}n}}\right)$$

where $\widehat{\sigma^2} = \overline{\widehat{\varepsilon}^2} - (\overline{\widehat{\varepsilon}})^2 = \overline{\widehat{\varepsilon}^2}$

Denote by $GL_F(t) = \int\limits_0^t F^{-1}(s)\,ds$ *a theoretical general Lorenz curve* (see [2]) where $F^{-1}(s) = \sup\{x : F(x) < s\}$ be a quantile function (a generalized inverse function) of a distribution function $F(x)$. Denote by $GL_n(t) = \frac{1}{n}\sum_{i=1}^{[nt]} \xi_{i:n}$ *an empirical Lorenz curve*. Goldie [2] proved a fundamental fact: an empirical Lorenz curve converges to a theoretical one in a uniform metric almost surely. Let $GL_F^0(t) = GL_F(t) - tGL_F(1)$ be a centered theoretical general Lorenz curve.

By $\Longrightarrow$ we denote a weak convergence (a convergence in distribution) in an appropriate space. So, we write $\Longrightarrow$ in theorems below to indicate a weak convergence in the space $C(0,1)$ with the uniform metric ([1], p. 82).

We use the next theorem from [3].

Theorem 1 *Let $X_i = \xi_{i:n}$ be order statistics generated by sample $(\xi_1, \ldots, \xi_n)$ with distribution function F, sequences $\{\varepsilon_i\}$ and $\{\xi_i\}$ are independent. If $0 < \mathbf{Var}\xi_1 < \infty$ then*

$\widehat{Z_n} \Longrightarrow Z_F^0$ where Z_F^0 is a centered Gaussian process with a covariance kernel $K_F^0(t,u)$, given by

$$K_F^0(t,u) = \min\{t,u\} - tu - \frac{GL_F^0(t)GL_F^0(u)}{\mathbf{Var}\xi_1},\ t,u \in [0,1].$$

We change $GL_F^0(t)$ by its estimation $GL_n^0(t) = GL_n(t) - tGL_n(1)$. We substitute sample variance S^2 for variance $\mathbf{Var}\xi_1$. Let

$$K_n^0(t,u) = \min\{t,u\} - tu - \frac{GL_n^0(t)GL_n^0(u)}{S^2},\ t,u \in [0,1].$$

Then $K_n^0(t,u) \to K_F^0(t,u)$ uniformly on $t,u \in [0,1]$ as $n \to \infty$.

Our statistical test use values of the empirical bridge in d points: let

$$\mathbf{a} = (a_1, \ldots, a_d) = \left(\frac{1}{d+1}, \ldots, \frac{d}{d+1}\right),$$

$$G = \left(K_F^0(a_i,a_j)\right)_{i,j=1}^d,\quad G_n = \left(K_n^0(a_i,a_j)\right)_{i,j=1}^d,$$

$$q = (\widehat{Z_n}(a_1), \ldots, \widehat{Z_n}(a_d))^T.$$

If G^{-1} exists then $q^T G^{-1} q$ is a quadratic form, that is, a continuous functional of $\widehat{Z_n}$. Therefore we have

Corollary 1 *Let conditions of Theorem 1 be satisfied. If G^{-1} exists then $q^T G_n^{-1} q$ converge weakly to χ^2-distribution with parameter d.*

P-value for the test is $\alpha^* = 1 - F_{\chi_d^2}(q^T G_n^{-1} q)$.

We choose $d = [n^{1/3}] + 1$. We have $n = 364$, $d = 8$ for all the sample, $n_1 = 132$, $d_1 = 6$, $n_2 = 216$, $d_2 = 7$ for its 1st and 2nd parts (corresponding years 1991–1999 and 2000–2008).

Caculations give $\alpha^* << 10^{-4}$ for all the sample, $\alpha_1^* = 0.1677$ for its 1st part and $\alpha_2^* = 0.07505$ for its 2nd part. Therefore the test rejects the basic hypothesis in all the time interval at the 10^{-4} level and accepts it in intervals 1991–1999 and 2000–2008 at the 0.07 level.

So one can calculate estimated prices of cars in these intervals using models with corresponding coefficients (see Table 1). There is a gap between 1999 and 2000.

Table 1: Estimated prices in thousands of roubles

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
Price	130	139	148	158	169	180	192	205	219
Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Price	290	312	335	360	386	415	446	479	514

Conclusions

We suggest the statistical test for analysis of correspondence of the simple regression model to data that is ordered by a regressor. We have constructed and tested the linear regression model for logarithms of car prices. The model works in intervals 1991–1999 and 2000–2008. The price difference is about 6% per year in 1991–1999 and about 7% per year in 2000–2008. Cars of 1999 cheaper than cars of 2000 approximately 1.32 times.

Acknowledgements

The work is partially supported by RFBR (grant 13-01-00661). Author thanks Nikolay Arkashov for helpful discussions and Alexandra Kovalenko for some calculations.

References

[1] Billingsley P. (1968). *Convergence of Probability Measures*. New York: John Wiley & Sons.

[2] Goldie C. M. (1977) Convergence theorems for empirical Lorenz curves and their inverses *Advances in Applied Probability* Vol. **9**, pp. 765–791.

[3] Kovalevskii A., Shatalin E. (2013) A limit process for a sequence of partial sums of residuals of a simple linear regression against order statistics *(in preparation)*.

[4] Kovalevskii A., Shatalin E. (2011) Asymptotic distribution of empirical bridge for regression on order statistics *V International Conference "Limit Theorems in Probability Theory and Their Applications"*, p. 26.

[5] Kuiper S. (2008) Introduction to Multiple Regression: How Much Is Your Car Worth? *Journal of Statistics Education* Vol. **16**, No.3.

Quantifying the Price of Uncertainty in Bayesian Models

Milovan Krnjajić[1], David Draper[2]
[1] *National University of Ireland, Galway*
[2] *University of California at Santa Cruz*
e-mail: milovan.krnjajic@nuigalway.ie

Abstract

During the exploratory phase of a typical statistical analysis it is natural to look at the data in order to narrow down the scope of the subsequent steps, mainly by selecting a set of families of candidate models (parametric, for example). One needs to exercise caution when using the same data to assess the parameters of a specific model and deciding how to search the model space, in order not to underestimate the overall uncertainty, which usually occurs by failing to account for the second order randomness involved in exploring the modelling space. In order to rank the models based on their fit or predictive performance we use practical tools such as Bayes factors, log-scores and deviance information criterion. Price for model uncertainty can be paid automatically when using Bayesian nonparametric (BNP) specification, by adopting weak priors on the (functional) space of possible models, or in a version of cross validation, where only a part of the observed sample is used to fit and validate the model, whereas the assessment of the calibration of the overall modelling process is based on the as-yet unused part of the data set. It is interesting to see if we can determine how much data needs to be set aside for calibration in order to obtain an assessment of uncertainty approximately equivalent to that of the BNP approach.

Keywords: model uncertainty, Bayesian non-parametric specification, cross validation, model choice

Introduction

When faced with a task of analyzing a data set, statisticians usually take a standard, data-analytic (DA), approach to model specification. In DA approach we explore the space of models in search for the 'right' model, using all of the available data and then using the same data to draw inferential or predictive conclusions conditional on the results of the search. This amounts to using the data twice and often yields poorly calibrated (too narrow) predictive intervals.

There seem to be only two principled solutions to this problem: (1) Bayesian nonparametric (BNP) modelling (with enough data) in which prior distributions are specified on the entire model space, therefore avoiding some of the search and the use of data to specify error distributions, response surfaces, etc., and (2) A version of Bayesian cross-validation (we call it 3-way out-of-sample predictive cross-validation, or 3CV, in a manner somewhat related to a method used in machine learning; 3CV is a modification of DA search in which the data are partitioned into 3 subsets $(S_1; S_2; S_3)$,

rather than the usual 2, and where a DA search is undertaken iteratively, modeling with S_1 and predictively validating with S_2; S_3 is not used in quoting final uncertainty assessments, but is instead used to evaluate predictive calibration of the entire modeling process. It looks as if the approach (2) resolves the problem by paying the "right" price for shopping around in the modelling space in terms of setting aside a part of the data.

BNP modeling is often characterized as providing "insurance" against mis-specified parametric models for the following reason: (a) You can generate data from a known ("true") parametric model M_1 and fit M_1 and a BNP model to the simulated data sets; both will be valid (both will reconstruct the right answer averaging across simulation replications) but the BNP uncertainty bands will typically be wider. (b) You can also generate data from a different model M_2 and fit M_1 and BNP to the simulated data sets; often now only BNP will be valid. People refer to the wider uncertainty bands for BNP in (a) as the "insurance premium" you have to pay with BNP to get the extra validity of BNP in (b). But this is not a fair comparison: the simulation results in (a) and (b) were all conditional on a known "true" model, and don't immediately apply to a real-world setting in which you don't know what the "true" model is. However, when you pay an appropriate price for shopping around for the "right" parametric model (as in 3CV), the discrepancy between the parametric and BNP uncertainty bands vanishes.

The approach described above begs the question – can we quantify in a general way (and how exactly) the price of model uncertainty? One idea involves a comparison of *how much data* Bayesian parametric and nonparametric models need *to achieve the same inferential accuracy* about the main quantity of interest. In order to quantify the price of model uncertainty we may proceed as follows: specify a BNP model centered at an a priori plausible parametric model using all n data values and perform the inference; then find out how many data points $n_{DA} < n$ are needed by the best parametric model, discovered with a DA search, to achieve the same inferential accuracy as the BNP model; the difference $(n - n_{DA})$ is how much data should be reserved in 3CV subset S_3.

The plan of the paper is as follows: in the first section we describe the simulation setup with a parametric (Poisson based) model and its BNP counterpart and explain differences in estimated inferential and predictive uncertainty. Section 2 describes an attempt to gauge out what fraction of the data set should be used in the calibration stage of a DA model that results in an assessment of uncertainty approximately equivalent to that of the BNP approach.

1 Bayesian parametric Poisson based model vs. a BNP model

Assume that we have a data set (of size n) consisting of counts coming from an unknown data generating mechanism. The first thing to try parametrically with

count data is usually a fixed-effects Poisson (FEP) model (for $i = 1, \ldots, n$):

$$\begin{array}{rcl} (y_i|\theta) & \overset{\text{ind}}{\sim} & \text{Poisson}[\exp(\theta)] \\ (\theta|\mu,\sigma^2) & \overset{\text{iid}}{\sim} & N(\mu,\sigma^2) \\ (\mu,\sigma^2) & \sim & p(\mu,\sigma^2). \end{array} \tag{1}$$

This specification uses a Lognormal prior for $\lambda = e^\theta$ rather than conjugate Gamma choice; the two families are similar, and the Lognormal generalizes more readily. In practice data often exhibit heterogeneity resulting in (extra-Poisson variability), manifesting as variance-to-mean ratio, $VTMR > 1$. A natural parametric extension to FEP would be to try a random effects Poisson model (REP):

$$\begin{array}{rcl} (y_i|\theta_i) & \overset{\text{ind}}{\sim} & \text{Poisson}[\exp(\theta_i)] \\ (\theta_i|G) & \overset{\text{iid}}{\sim} & G \\ G & \equiv & \text{N}(\mu,\sigma^2) \\ (\mu,\sigma^2) & \sim & p(\mu,\sigma^2), \end{array} \tag{2}$$

Here, $i = 1, ..., n$, and we assume a cumulative distribution function (CDF) of latent variables (random effects), θ_i, to be parametric (Gaussian).

The problem is that the mixing distribution, G, in the population to which it is appropriate to generalize may be multimodal or skewed, which a single Gaussian can't capture. If so, this REP model can fail to be valid. Moreover, this would usually be diagnosed with something like a density trace of posterior means of θ_i, looking for need to use mixture of Gaussians instead of single one, but choosing G to be Gaussian will tend to make diagnostics support Gaussian model even when it's not right.

Therefore, it would be good to remove the assumption of a specific parametric family (Gaussian) for the mixing distribution G of the random effects, by allowing G to be random and specifying a prior model on the space of $\{G\}$. This BNP model may be centered on a Gaussian model, $N(\mu,\sigma^2)$, but would permit adaptation/learning. Specifying a prior for an unknown distribution requires a stochastic process with realizations (sample paths) that are CDFs. We use the Dirichlet process (DP): $G \sim DP(\alpha\, G_0)$, where G_0 is the center or base distribution of the process and α is a precision parameter, see [3]. DP mixture Poison model (DPMP: this paper's BNP model):

$$y_i \mid G \overset{ind}{\sim} \int \text{Poisson}(y_i; e^\theta) \text{d}G(\theta), \tag{3}$$

where G is a random mixing distribution. For a data set $y = (y_1, ..., y_n)$ the BNP model is:

$$\begin{array}{rcl} y_i \mid \theta_i & \overset{ind}{\sim} & \text{Poisson}(e^{\theta_i}) \\ \theta_i \mid G & \overset{iid}{\sim} & G \\ G & \sim & \text{DP}(\alpha, G_0(\psi)), \end{array} \tag{4}$$

where $\psi = (\mu, \sigma^2), G_0 \equiv N(\cdot; \mu, \sigma^2)$ and $i = 1, ..., n$. Additional model stages are introduced by placing priors on α and ψ. MCMC implemented for a marginalized version of DP mixture. Key idea: G is integrated out over its DP prior, resulting in a marginalized version of (4) that follows Pólya urn structure, as shown in [2].

Further references and details of DP mixture modelling along with the description of the simulations with a number of data sets can be found in [4]. Here, it suffices to say that the sample sizes were $n = 300$ and that the data sets were generated based on a variety of unimodal (symmetric and skewed) and bimodal distributions of latent variables (random effects) resulting in data samples with increased variability, nontrivial tails, and densities which were unimodal or with a slight to a noticeable bimodality.

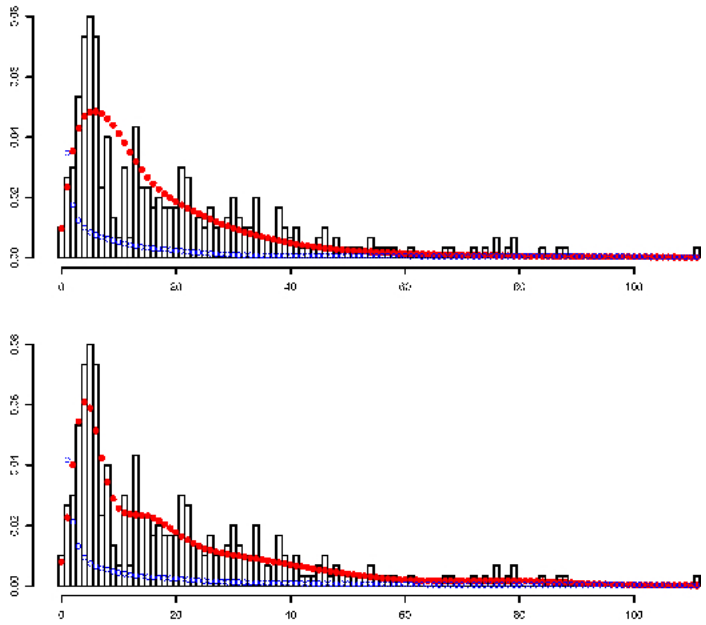

Figure 1: *Prior (blue) and posterior (red) predictive distribution from REP model (top) and BNP model (bottom).*

Figure 1 shows the posterior predictive distributions obtained from the parametric REP model and a BNP model with a DP prior, where the data set of counts was generated by a model with a bimodal distribution of latent variables (random effects). (The posterior predictive distribution is always obtained as $p(y^*|y) = \int_\Theta p(y^*|\theta)p(\theta|y)d\theta$) It is obvious from the graphs that the REP model can't adapt to bimodality or skewness without remodelling (say) as a mixture of Gaussians on the latent scale, whereas the BNP modelling smoothly adapts to the data-generating mechanism. A formal comparison of the parametric and the BNP model (using log-scores and deviance information criterion, DIC) showed clear preference for the BNP model when the data were generated with non-Gaussian distribution of random effects.

It is interesting to analyze what is happening on the scale of latent variables which come from random mixing distribution G. We can do this since the BNP model permits obtaining posterior draws of G, $P(G \mid \text{ data })$. Based on these draws we can compute estimates such as the mean functional, $E[y \mid G)]$, and in fact, obtain the entire distribution of $E[y \mid G)]$. [1] derived an important result for the posterior distribution of the random mixing distribution G. It turns out that for $G \sim DP(\alpha, G_0(\psi))$, the posterior of G is as follows:

$$(G|\text{data}) \sim \int P(G|\theta, \alpha, \psi)\mathrm{d}Q(\theta, \alpha, \psi \mid \text{ data }), \tag{5}$$

where $P(G \mid \theta, \alpha, \psi)$ is also a DP with parameters $\alpha^{'} = \alpha + n$ and

$$G_0^{'}(\cdot|\psi) = \frac{\alpha}{\alpha + n} G_0(\cdot|\psi) + \frac{1}{\alpha + n} \sum\nolimits_{i=1}^{n} 1_{(-\infty, \theta_i]}(\cdot), \tag{6}$$

and $Q(\theta, \alpha, \psi \mid \text{ data })$ is the joint posterior distribution. Using (5), (6) along with the definition of DP we obtain posterior sample paths from $P(G \mid \text{data})$ in a computationally efficient way.

2 Parametric vs BNP models: the price of model uncertainty

Posterior estimates of the means of random effects distribution G along with the 90% point-wise uncertainty bands are shown in Figure 2. It is obvious that the REP model can't capture the skewness and bimodality of the CDF (of the distribution of random effects, G), what is not surprising since REP assumes a Gaussian here. Yet, what is somewhat remarkable in a negative way is the very narrow uncertainty bands. On the other hand, the BNP model captures well both non-standard shapes of the CDF-s as expected, albeit with wider uncertainty bands around the mean estimate.

We have seen that when REP is incorrect model, it continues to yield narrower uncertainty bands that fail to include the truth, whereas BNP model adapts successfully to the data-generating mechanism, as is illustrated in Figure 2. However, the Gaussian assumption on the latent variables scale in the REP model, although wrong, can make the model look plausible when it's not: Diagnostic checking of REP model would make it look appropriate when it's not; by contrast BNP correctly captures the bimodality or skewness of the random effects distribution.

One way to pay the right price for conducting a data-analytic search to arrive at a final parametric model is three-way cross-validation (3CV) which proceeds along the following lines: (1) Partition data at random into *three* subsets S_i, of size n_i (respectively). (2) Fit tentative {likelihood + prior} to S_1. Expand initial model in feasible ways suggested by the data exploration using S_1. Iterate until fit is good (for example). (3) Use final model (fit to S_1) from (2) to create predictive distributions for all data points in S_2. Compare actual outcomes with these distributions, checking the predictive performance. Go back to (2), change likelihood or re-tune the prior as

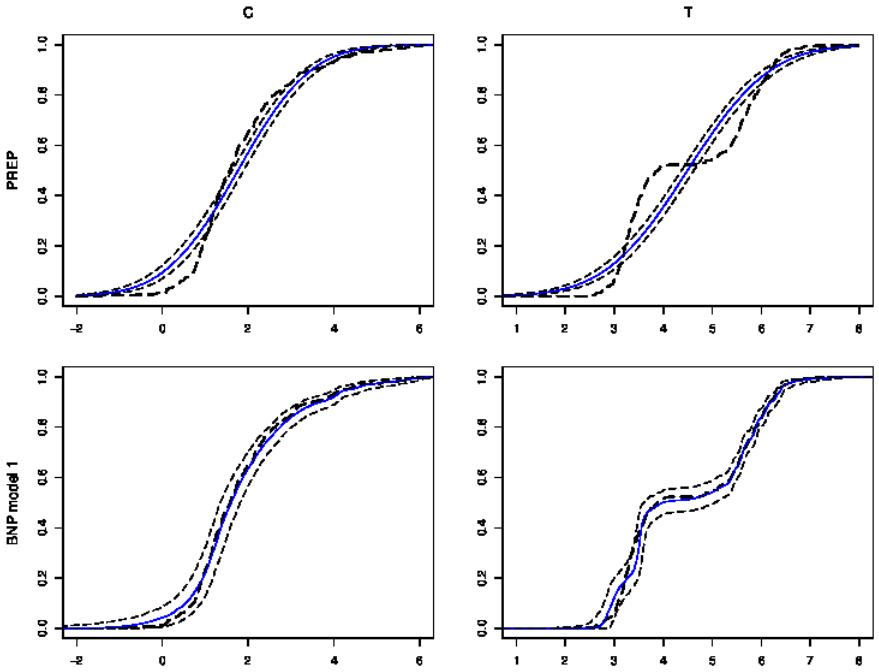

Figure 2: *Posterior MCMC estimates of the means of random effects distributions G, with* 90% *uncertainty bands. REP model, first row; BNP model, second row. Data sets generated using a model with skewed (left panels) and bimodal (right panels) distributions of random effects. The CDF-s of these (true) distributions are represented with thick dashed lines.*

Table 1

	REP				DPMP			
n	Area	r_A	MaxDiff	r_D	Area	r_A	MaxDiff	r_D
200	0.2256	1.403	0.11510	1.440	0.5556	1.160	0.3910	1.415
400	0.1608	1.433	0.07992	1.372	0.4788	1.141	0.2763	1.007
800	0.1122	1.427	0.05827	1.456	0.4195	1.090	0.2745	1.123
1600	0.0786		0.04002		0.3849		0.2445	

necessary, to get good calibration. Iterate until the predictive performance is OK (for example). (4) Announce final model (fit to $S_1 \cup S_2$) from (3), and report predictive calibration of this model on data points in S_3 as an indication of how well it would perform with new data.

In practice, with large n we probably only need to do this once, whereas with small and moderate n it may be necessary to repeat (1–4) several times and (perhaps) combine results in some way (for example, through model averaging). Note again that n_3 observations in S_3 are *not* to be used in summarizing inferential uncertainty about the quantities of interest but are instead used to estimate calibration of the data-analytic modeling process.

We need to find a way to determine or estimate sizes n_i of three data subsets (S_1, S_2, S_3). In order to approach this task of quantifying the price of model uncertainty it is useful (a) to regard Bayesian parametric models as just BNP models with a stronger prior. For example: REP model takes $G \equiv N(\mu, \sigma^2)$ while DP mixture model takes $G \sim DP(\alpha G_0), G_0 \equiv N(\mu, \sigma^2)$. Notice that larger sample sizes and stronger prior information often lead to narrower uncertainty bands.

Therefore, it is natural that a BNP model, on account of its vague prior on a large space of distribution functions, would require more data (sample size n_{BNP}) to achieve (about) the same inferential or predictive accuracy as the best-fitting (best-predictive) *parametric* model (in terms of sample sizes, $n_{BNP} >$ sample size $= n = n_{Param}$. It is then reasonable to recommend $n_3 = n(1 - n/n_{BNP})$ as the size of the calibration subset S_3. Combining this with the typical cross-validation practice that you should put about twice as much data in the modeling subset as in the validation subset yields

$$(n_1, n_2, n_3) = \left[\frac{2n^2}{3n_{BNP}}, \frac{n^2}{3n_{BNP}}, n\left(1 - \frac{n}{n_{BNP}}\right)\right]. \tag{7}$$

Therefore, for a data set with $n = 1000$ observations, if it takes about $n_{BNP} = 1200$ observations to achieve BNP accuracy equivalent to that of the best parametric model on the main quantities of interest, the subsets S_i should have about $(550, 275, 175)$ observations in them.

Implementing this idea (obviously) requires estimating n_{BNP}. As a data-generating mechanism we use a REP model (with Gaussian G); and generate four samples of sizes
$n = (200, 400, 800, 1600)$. To quantify the effect of (doubling) sample size, we compute (1) the areas between the 0.05 and 0.95 point-wise quantiles of the posterior

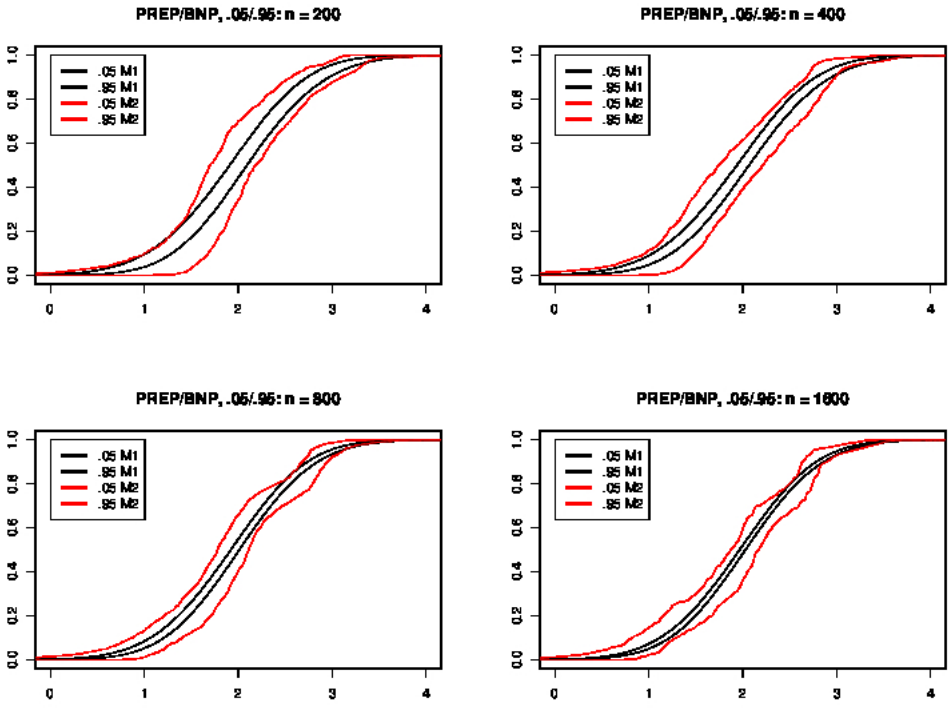

Figure 3: *95% point-wise uncertainty bands of posteriors of G, produced by the REP model (smooth & black) REP, and DPMP model (jagged & red).*

realization of the CDF-s of G, and (2) the maximum differences between two quantiles. The results are summarized numerically in Tables 1 and 2. Figure 3 shows estimates of the 90% uncertainty bands of the posterior distribution of the CDF of G for parametric and BNP model and different sample sizes.

We see that the REP model learns about G at a substantially faster rate than the DPMP model. Noting the values of r_A and r_D, the ratios of the consecutive values of"Area" and "MaxDiff" it appears that the REP learning rate follows a square root law, but the DPMP rate does not. However, if the data-generating mechanism was non-REP the REP model would continue to "learn" the wrong CDF at a same $\sqrt{n}$ rate, whereas the DPMP model would (somewhat slower) learn the right G.

Besides looking at the scale of latent variables, a similar comparison can be made on the data scale and for that purpose we use the mean functional:

$$E(y \mid G) = \sum_{y=0}^{\infty} yF(y;G) = \sum_{y=0}^{\infty} y \int \text{Poisson}(y;\theta)G(\mathrm{d}\theta) = \int e^{\theta}G(\mathrm{d}\theta).$$

The mean functional has a closed form in case of REP model, whereas for the DPMP model we use MCMC draws from the joint posterior distribution of all parameters to compute it. The results for $E(y \mid G)$ are summarized in Table 2. It was unexpected to see that DPMP appears to learn about the posterior mean on the data scale at a faster rate than REP, although the difference between the two decreases for larger

Table 2

	90% Interval Width For $E(y|\text{data})$			
n	REP		DPMP	
200	1.793	1.433	1.679	1.424
400	1.251	1.564	1.179	1.483
800	0.800	1.396	0.795	1.438
1600	0.573		0.553	

sample sizes. The result is counterintuitive, but an explanation may be given based on how the standard MCMC estimate of the posterior mean on the data scale is computed:

$$u_j = \sum_{k=1}^{K} \exp(t_k) \left[G_j(t_k) - G_j(t_k^-) \right] \tag{8}$$

for each MCMC iteration j, where $\{t_1, \ldots, t_K\}$ is a grid of points at which $G_j(\cdot)$, the current MCMC iteration estimate of G, is evaluated; the many flat segments in G_j when the sample size is small can result in the uncertainty assessment on the low side.

Conclusion

To summarize the results, we can say that BNP models adapt well at latent and data levels and have superior predictive performance. It is interesting to see that a weaker prior information (provided by specifying priors on space of distributions in BNP) does not necessarily lead to weaker inferential statements on the data scale. Stronger prior information (when wrong, but difficult to diagnose) can lead to wrong inference in a somewhat striking manner.

Inferential uncertainty measured on the latent scale was smaller for parametric models and decreased with sample size, however, the uncertainty on the data scale was smaller for the BNP model. It means that this search for data equivalence between parametric and BNP models leads eventually in oposite directions and cannot be used to estimate the desired amount of data to use for calibration in DA approach.

The concept of data equivalence, if it worked, could lead to a fairly general way of quantifying the price of uncertainty for a data-driven search of the model space. However, the structure of the space of latent variables in BNP models changes non-trivially with the sample size and also reflects the features of the data set (such as skewness and multimodality), making the comparison with parametric problems a challenge. In general, $p(y \mid x) = \int p(y \mid x, M) p(M \mid x) dM$, where M is a space of models, $p_1(M)$ may be a weaker prior than $p_2(M)$, and yet $p_1(M)$ may concentrate on models with better predictive accuracy, $p(y \mid x, M)$ than $p_2(M \mid x)$ does, leading

to stronger inference from $p_1(y \mid x)$ than from $p_2(y \mid x)$.

References

[1] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to non-parametric problems, *Annals of Statistics*, 2: 1152-1174.

[2] Blackwell and MacQueen (1973). Ferguson distributions via Plya urn schemes. *Annals of Statistics*, 1: 353-355.

[3] Ferguson, T. (1974). Prior distributions on spaces of probability measures, *Annals of Statistics*, 2: 615-629.

[4] Krnjajić, M., Kottas A. and Draper, D. (2008). Parametric and nonparametric Bayesian model specification: a case study involving models for count data. *Computational Statistics and Data Analysis*, 52: 2110-2128.

Robust Estimation of Mixed Response Regression Models

Daniil V. Lisitsin
Novosibirsk State Technical University
Novosibirsk, Russian Federation
e-mail: dalis2@yandex.ru

Abstract

In paper earlier developed theory of an optimum estimation of unknown parameters of statistical model in the presence of multivariate nonhomogeneous data is applied to a case of regression models with responses of the mixed type (polytomous quantal and continuous). The theory is based on use of the weighted L_2-norm of Hampel's influence function. Estimators provide a robustness to a deviation of observations distribution from postulated distribution.

Keywords: parameter estimation, regression, mixed outcomes, robustness, influence function.

Introduction

The classical statistic procedures are based on a number of assumptions which can't be fulfilled in practice. Under such conditions a lot of widespread statistic procedures lose their positive qualities. For instance, the procedures, which rest on the maximum likelihood method. But this problem can be solved by using robust estimators [2, 7]. For example, Shurygin's approach [7] based on use of the weighted L_2-norm of Hampel's influence function (see [5]) allows to get the estimators possessing a high robustness and efficiency.

Generally robustness theory has been developed for the quantitative random variables modeling. Qualitative and mixed variables modeling are paid much less attention. Originally the estimators within Shurygin's approach were formed only for continuous random variables models.

Shurigin's approach has been extended in [5] to a case of multivariate nonhomogeneous outcomes represented in numerical form, but not necessarily are numerical, for example, they can be qualitative or mixed; the separate consideration is given to a case of the incomplete data; conditions that make the missing-data mechanism ignorable are obtained. Cases of the scalar count and qualitative outcomes are investigated earlier [1, 3].

The purpose of our study is application of the theory from [5] to regression models with polytomous quantal and continuous responses. Further the case of scalar polytomous variable is considered. Case of a vectorial polytomous variable can be always reduced to the previous case.

Our model is the general location model introduced in [6] and extended in [4] to a case of regression model with polytomous quantal and continuous responses.

1 Model specification and method of estimation

Assume that polytomous quantal response variable ζ has a fixed number of acceptable values $\{1, 2, ..., J\}$. Distribution of ζ is set of probabilities

$$\Pr\{\zeta = j | x, \alpha\} = \pi_j(x, \alpha),$$

where x is a vector of input variables, α is a vector of parameters.

Given that $\zeta = j$, the vector of n continuous response variables has a conditional distribution with density $g_j(y, x, \phi)$, where ϕ is a vector of parameters, $g_j(y, x, \phi) > 0$ for $y \in R^n$.

Define the sample as $(y_t, z_t, x_t), t = 1, \ldots, N$, where y_t is the tth observation of continuous response variables, z_t is the tth observation of quantal response variable, and x_t is the tth observation of input variables.

M-estimate $\hat{\Phi}$ of parameter vector $\Phi = (\alpha^T, \phi^T)^T$ is obtained by solving system of equations

$$\sum_{t=1}^{N} \psi(y_t, z_t, x_t, \hat{\Phi}) = 0,$$

where $\psi(y_t, z_t, x_t, \hat{\Phi})$ is a vectorial score function satisfying further condition for all t:

$$\mathbf{E}\, \psi(y_t, z_t, x_t, \Phi) = 0, \tag{1}$$

$\mathbf{E}$ is an expectation operator.

2 Optimal estimation

One of the major indicators of estimator's robustness is Hampel's influence function [2] which in the case under some regularity conditions takes the form

$$\mathrm{IF}(y, z, x_t, \Phi) = M^{-1} \psi(y, z, x_t, \Phi),$$

where

$$M = -\sum_{u=1}^{N} \frac{\partial}{\partial \tilde{\Phi}^T} \mathbf{E}\, \psi(y_u, z_u, x_u, \tilde{\Phi}) \Big|_{\tilde{\Phi}=\Phi} =$$

$$= \sum_{u=1}^{N} \sum_{j=1}^{J} \int_{R^n} \psi(y, j, x_u, \Phi) \frac{\partial}{\partial \Phi^T} [\pi_j(x_u, \alpha)\, g_j(y, x_u, \phi)]\, dy.$$

Indicator of estimation badness can be written as square of the weighted L_2-norm of an influence function, namely,

$$\Psi_s(\psi) = \sum_{t=1}^{N} \sum_{j=1}^{J} \int_{R^n} \mathrm{IF}^T(y, j, x_t, \Phi)\, W\, \mathrm{IF}(y, j, x_t, \Phi) s(y, j, x_t, \Phi)\, dy,$$

where $W = W(\Phi)$ is some symmetric positive definite weight matrix, $s(y, j, x, \Phi)$ is a weight function, $s(y, j, x, \Phi) > 0$ for $y \in R^n$.

Note that function s can be interpreted on the basis of Bayesian point-mass contamination model [7]. In this case two first argument of influence function are the random variables (contamination variables). Assume that their distribution is formed on the basis of the general location model in sample series when probabilities $\sigma_j(x_t, \alpha), j = 1, ..., J$, for quantal contamination variable and conditional density $s_j(y_t, x_t, \phi)$ for vector of continuous contamination variables are defined. Then function s and functional Ψ_s are of the form represented

$$s(y, j, x, \Phi) = \sigma_j(x, \alpha)\, s_j(y, x, \phi), \tag{2}$$

$$\Psi_s(\psi) = \sum_{t=1}^{N} \mathbf{E}_{s_t} \mathrm{IF}^T(y_t, z_t, x_t, \Phi)\, W\, \mathrm{IF}(y_t, z_t, x_t, \Phi),$$

where $\mathbf{E}_{s_t}$ is an expectation with use of contamination variable distribution under observation t.

Optimal score function is a solution of minimization problem:

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

Function ψ_s^* is represented as follows:

$$\psi_s^*(y, z, x, \Phi) = C \left\{ \frac{\partial}{\partial \Phi} \ln \left[\pi_z(x, \alpha)\, g_z(y, x, \phi) \right] + \beta \right\} \frac{\pi_z(x, \alpha)\, g_z(y, x, \phi)}{s(y, z, x, \Phi)}, \tag{3}$$

where $C = C(\Phi, W)$ is some nonsingular matrix, vector $\beta = \beta(x, \Phi)$ provides fulfillment of the condition (1).

Generalized radical estimator is an example corresponding to the case

$$s(y, z, x, \Phi) = \left[\pi_z(x, \alpha)\, g_z(y, x, \phi) \right]^{1-\lambda} / \Delta,$$

where λ is estimator parameter ($\lambda \geq 0$), Δ is additional weight function with suitable arguments, it can be used for a normalisation of densities s, σ_j, and g_j. Note that the case of $\lambda = 0$ matches maximum likelihood estimation.

The case of $\lambda = 1$, $\Delta = 1$ with weight function $s(y, z, x, \Phi) = 1$ strictly corresponds to concepts of the robustness theory. It is use of unweighted L_2-norm of an influence function, but function s is not interpreted as a probability density.

Generalized radical score function is represented as follows:

$$\psi_\lambda^*(y, z, x, \Phi) = C \left\{ \frac{\partial}{\partial \Phi} \ln \left[\pi_z(x, \alpha)\, g_z(y, x, \phi) \right] + \beta \right\} \left[\pi_z(x, \alpha)\, g_z(y, x, \phi) \right]^{\lambda} \Delta.$$

Conditional optimal score function is a solution of minimization problem $\min\limits_{\psi} \Psi_{s_1}(\psi)$ under condition

$$\Psi_{s_2}(\psi) \leq D(\Phi) \tag{4}$$

or

$$\Psi_{s_2}(\psi) = D(\Phi) \tag{5}$$

($s_1 \neq s_2$).

Corresponding optimum score functions under some regularity conditions are given by (3) with

$$s(y, z, x, \Phi) = s_1(y, z, x, \Phi) + \gamma(\Phi) s_2(y, z, x, \Phi) > 0,$$

where $\gamma(\Phi)$ provides fulfillment of the condition (4) (together with conditions $\gamma(\Phi) \geq 0$ and $\gamma(\Phi)\left[\Psi_{s_2}(\psi) - D(\Phi)\right] = 0$) or (5) respectively.

3 Example

Assume that under the condition $z_t = j$ regression model for continuous modelling variables and continuous contamination variables is of the form represented

$$y_t = \eta_j(x_t, \theta) + e_t,$$

where η_j is a response function, θ is a vector of parameters, e_t is the tth error.

Assume that function $s(y, z, x, \Phi)$ can be written in the form (2) and errors have a distribution with null location and a density $g(e, \tau)$ ($s(e, \tau)$), where τ is a vector of parameters, for all values j and continuous modelling variables (continuous contamination variables). In addition following condition is fulfilled:

$$\int\limits_{R^n} \frac{\partial g(e, \tau)}{\partial e} \frac{g(e, \tau)}{s(e, \tau)}\, de = 0.$$

It is easily shown that optimal score functions for α and ϕ are of the form represented

$$\psi^*_{s,\alpha}(y, z, x, \Phi) = \psi^*_{s,\alpha}(z, x, \alpha) \frac{g(y - \eta_z(x, \theta), \tau)}{s(y - \eta_z(x, \theta), \tau)},$$

$$\psi^*_{s,\phi}(y, z, x, \Phi) = \psi^*_{s,\phi,z}(y, x, \phi) \frac{\pi_z(x, \alpha)}{\sigma_z(x, \alpha)},$$

where $\psi^*_{s,\alpha}(z, x, \phi)$ is optimal score function for estimation of parameter α with use of observed quantal responses only, $\psi^*_{s,\phi,j}(y, x, \phi)$ is optimal score function for estimation of parameter ϕ with use of observed continuous responses for subsample with $z_t = j$ only.

Such property allows to construct score functions for a case of the mixed responses, knowing score functions for a case of the quantal responses and a case of the continuous responses.

Let's continue our example. For modeling dependence of nominal response from input variables polytomous logistic regression is often used. Corresponding probabilities are of the form

$$\pi_j(x,\alpha) = \exp\left[f^T(x)\alpha_j)\right]\left\{1+\sum_{k=1}^{J-1}\exp\left[f^T(x)\alpha_k\right]\right\}^{-1},$$

where $f(x)$ is a vector of regressors, α_j is a subvector of α (subvectors α_j, $j = 1,2,...,J-1$, are not intersected), α_J is a null vector.

For model of continuous variables we consider $\eta_j(x_t,\theta) = F_j(x_t)\theta_j$ and $e_t \sim N_n(0,\Omega)$, where $F_j(x_t)$ is a matrix of regressors, θ_j is a subvector of θ (subvectors θ_j, $j = 1,2,...,J$, are not intersected), 0 is a null vector, Ω is a covariance matrix. Thus, $g(e,\tau) = g(e,\operatorname{vech}\Omega) = (2\pi)^{-n/2}|\Omega|^{-1/2}\exp(-\frac{1}{2}e^T\Omega^{-1}e)$, where vech is a half-vec operator.

Generalized radical estimates of α_j, θ_j, and Ω obtain by solving system of equations

$$\sum_{t=1}^{N}\hat{w}_t\left\{\delta_{jz_t}-[\pi_j(x_t,\hat{\alpha})]^{1+\lambda}/\textstyle\sum_{k=1}^{J}[\pi_k(x_t,\hat{\alpha})]^{1+\lambda}\right\}f(x)=0,$$

$$\hat{\theta}_j = \left[\sum_{t:z_t=j}\hat{w}_t F_j^T(x_t)\hat{\Omega}^{-1}F_j(x_t)\right]^{-1}\sum_{t:z_t=j}\hat{w}_t F_j^T(x_t)\hat{\Omega}^{-1}y_t,$$

$$\hat{\Omega} = \frac{1+\lambda}{\sum_{t=1}^{N}\hat{w}_t}\sum_{t=1}^{N}\hat{w}_t\left[y_t-F_{z_t}(x_t)\hat{\theta}_{z_t}\right]\left[y_t-F_{z_t}(x_t)\hat{\theta}_{z_t}\right]^T,$$

where $\hat{w}_t = w(y_t,z_t,x_t,\hat{\Phi}) = \left[\pi_{z_t}(x_t,\hat{\alpha})\,g(y_t-F_{z_t}(x_t)\hat{\theta}_{z_t},\operatorname{vech}\hat{\Omega})\right]^{\lambda}\Delta$, δ is Kronecker delta. Note that two last subsystems define some modification of Meshalkin's estimator [7].

References

[1] 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.

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

[3] 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.

[4] Lisitsin D.V. (2005). Parameter estimation of multifactorial model in presence of mixed responses. *Nauchniy Vestnik NGTU*. No. **1(19)**, pp. 11-20 (in Russian).

[5] Lisitsin D.V. (2013). Robust estimation of model parameters in presence of multivariate nonhomogeneous incomplete data. *Nauchniy Vestnik NGTU*. No. **1(50)**, pp. 17-30 (in Russian).

[6] Olkin I., Tate R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.* Vol. **32**, pp. 448-465.

[7] Shurygin A.M. (2009). *Mathematical methods of prediction*. Goryachaya Liniya — Telecom, Moscow (in Russian).

Observations-Based Stochastic Simulation of the Sea Surface Undulation and Extreme Ocean Waves

Kristina V. Litvenko
The Institute of Computational Mathematics and Mathematical Geophysics,
Siberian Branch of Russian Academy of Sciences,
Novosibirsk, Russia
e-mail: litchristina@gmail.com

Abstract

This paper deals with the numerical simulation of formation and development of the extreme ocean waves by using specific models of random processes and fields. The estimates of the frequency of extreme wave occurrence are studied based on the random fields outliers theory.

Keywords: numerical simulation, sea surface undulation, extreme waves, rogue waves, autoregression scheme, spectral models, random fields.

Introduction

The extreme waves, known as rogue or freak waves, represent a poorly understood natural phenomenon whose existence was distrusted because of the absence of reliable evidence. In contrast to tsunami waves, the solitary extreme waves are of 20 m, 30 m, or ever higher, thus essentially exceeding the heights of other waves, occurring suddenly, and vanishing far from the shore without visible causes; sometimes this occurs in a quiet sea with a relatively light wind. For the first time the rogue wave was instrumentally detected only in 1995, and nowadays strenuous efforts are aimed at observing the extreme waves and studing the rogue wave phenomenon both theoretically and experimentally.

In this paper conditional spectral models of random fields proposed in [3] are used for the numerical simulation of extreme waves. It is assumed that the sea surface roughness is sufficiently well described by a spatial-temporal random Gaussian field, which is stationary with respect to time and homogeneous in space. Numerical models of the sea surface based on this assumption were used, in particular, for the solution of applied ocean optics problems by Monte Carlo method [1]. Along with the spectrum of the extreme wave, the simulation of the extreme wave requires additional information concerning the wave profile, i.e., the field of the sea surface elevation should be specified at certain points at given time moments. Numerical experiments have established an extreme high sea level at the points on the wave crest. The results of these experiments are presented here. Conditional numerical models allow us to simulate numerically the set of independent spatial-temporal implementations of the sea level passing through given points and hence to study typical features of the development and propagation of extreme waves.

An autoregressive model is well known among the most common models of Gaussian stationary processes to obtain a time series. To estimate the autoregressive

parameters we made use of real data. Autoregressive model provides a wave such that it's height is 2 times greater than the height of a typical wave.

In this paper stochastical algorithms are used for the simulation of the sea roughness, the results obtained are presented, and the estimates of the frequency of extreme wave occurrence are studied based on the random fields outliers theory.

1 Calculation of the mean number of extreme waves over a fixed time period

The results of the outliers theory of random processes and fields allow us to calculate the mean number of extremely high waves above a given level in a specified region or for a known time interval according to the sea roughness model considered here. The corresponding formulas have the form (see, e.g., [6]):

$$N_{xy}(C) = \frac{C}{2\pi\sqrt{2\pi}\sigma_0^2}\sqrt{\frac{K_{22}K_{33} - K_{23}^2}{\sigma_0^2}}\exp\Big(-\frac{C^2}{2\sigma_0^2}\Big)\Delta S, \tag{1}$$

$$N_t(C) = (2\pi)^{-1}(\sigma_1/\sigma_0)\exp\big(-C^2/(2\sigma_0^2)\big)\Delta T, \quad C > 0. \tag{2}$$

Here $N_t(C)$ is the mean number of waves above the level C in the time interval ΔT, $N_{xy}(C)$ is the mean number of waves above the level C on the rough area ΔS (at a fixed time moment), σ_0^2 is the variance of the random field

$$K_{22} = \int_0^\infty\int_{-\infty}^\infty \lambda_1^2 f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2, \quad K_{33} = \int_0^\infty\int_{-\infty}^\infty \lambda_2^2 f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2,$$

$$K_{23} = \int_0^\infty\int_{-\infty}^\infty \lambda_1\lambda_2 f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2,$$

$$\sigma_1^2 = \int_0^\infty \mu^2 S(\mu)d\mu.$$

Here $f(\lambda_1,\lambda_2)$ is the spatial spectral density and S is the frequency spectrum. The level C means the height above the fixed level of the rough surface.

In this section, we consider the spectrum of the wind driven sea surface undulation, which is represented in [5]. Statistical properties of the sea roughness are determined within this model by the two parameters: the wind velocity v and the frequency of spectral maximum $\mu_{\max}$. Some values of $N_t(C)$ and $N_{xy}(C)$ for the model of the sea roughness with the spectrum and parameters

$$\mu_{\max} = 0.3\text{sec}^{-1}, \quad v = 10\text{m/sec}. \tag{3}$$

Table 1: The mean number $N_t(C)$ of waves above the level C within a year and the mean number $N_{xy}(C)$ of waves above the level C on the area of 100 square kilometers (at a fixed time moment) for parameters (3)

C (m)	$N_t(C)$	$N_{xy}(C)$
5.0	2.162×10^3	1.306×10^1
5.5	4.799×10^2	3.162×10^0
6.0	9.230×10^1	6.571×10^{-1}
6.5	1.538×10^1	1.174×10^{-1}
7.0	2.221×10^0	1.805×10^{-2}
7.5	2.778×10^{-1}	2.391×10^{-3}
8.0	3.011×10^{-2}	2.730×10^{-4}
8.5	2.828×10^{-3}	2.687×10^{-5}
9.0	2.301×10^{-4}	2.282×10^{-6}
9.5	1.622×10^{-5}	1.673×10^{-7}
10.0	9.911×10^{-7}	1.058×10^{-8}

are presented in Table 1. Table 2 contains the values for a stronger roughness with the parameters

$$\mu_{\max} = 0.4\text{sec}^{-1}, \quad v = 7\text{m/sec}. \tag{4}$$

We can give the following interpretation of the values of $N_t(C)$ for $C = 13$, 14, and 15m from Table 2. Suppose that 100 sea vessels navigate sufficiently far from each other with the sea roughness with the spectrum and parameters (4). Then the average number of meetings with the extreme wave per a year is 1 for the wave height exceeding 30 m (the wave height h corresponds to the level $C = h/2$), 11 - for the waves higher than 28 m, and 109 for waves higher than 26 m.

2 Modeling of giant waves on the basis of autoregressive schemes

In statistics and signal processing, an autoregressive model is a representation of a type of a random process; as such, it describes certain time-varying processes in nature. The autoregressive (AR) model specifies that the output variable depends linearly on its own previous values. The AR model is defined as

$$w(t_n) = \sum_{j=1}^{P} a_j w(t_{n-j}) + \epsilon_n, \tag{5}$$

where P is order of an autoregressive model, a_j, $j = 1, ..., P$ are autoregressive coefficients, ϵ_n is white noise.

Table 2: The mean number $N_t(C)$ of waves above the level C within a year and the mean number $N_{xy}(C)$ of waves above the level C on the area of 100 square kilometers (at a fixed time moment) for parameters (4)

C (m)	$N_t(C)$	$N_{xy}(C)$
5	2.484×10^5	7.439×10^2
6	9.681×10^4	3.475×10^2
7	3.179×10^4	1.330×10^2
8	8.795×10^3	4.200×10^1
9	2.050×10^3	1.099×10^1
10	4.027×10^2	2.395×10^0
11	6.663×10^1	4.351×10^{-1}
12	9.291×10^0	6.604×10^{-2}
13	1.091×10^0	8.385×10^{-3}
14	1.081×10^{-1}	8.915×10^{-4}
15	9.009×10^{-3}	7.944×10^{-5}
16	6.329×10^{-4}	5.936×10^{-6}

The Draupner wave or the New Year's wave was the first rogue wave to be detected by a measuring instrument, occurring on the Draupner platform in the North Sea off the coast of Norway on 1 January 1995. This wave is shown in Figure 1. Record of the New Year's wave is often used in the simulation of extreme waves. We used this recording of the waves to estimate autocorrelation function and the spectral density of the process (see Figure 2).

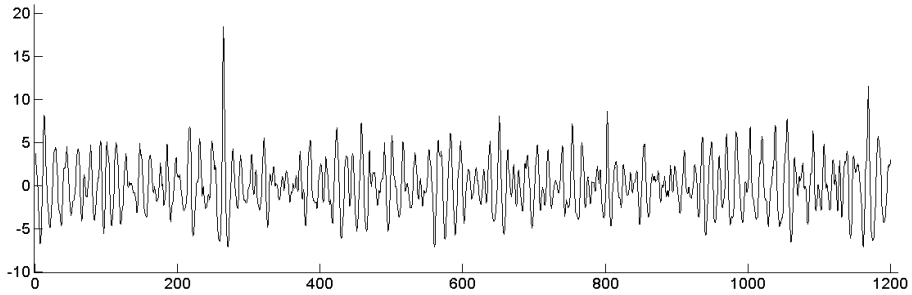

Figure 1: The Draupner wave or the New Year's occurring on the Draupner platform in the North Sea off the coast of Norway on 1 January 1995

Several methods are available to estimate an autoregressive model. In this paper, the Yule-Walker method is employed. The sector of the recording of the Draupner wave that does not contain an extreme wave was chosen to estimate autoregression coefficients. The autoregressive scheme of order 1990 was used to simulate a Gaussian random process. Figure 3 shows the implementation of the freak waves, which were

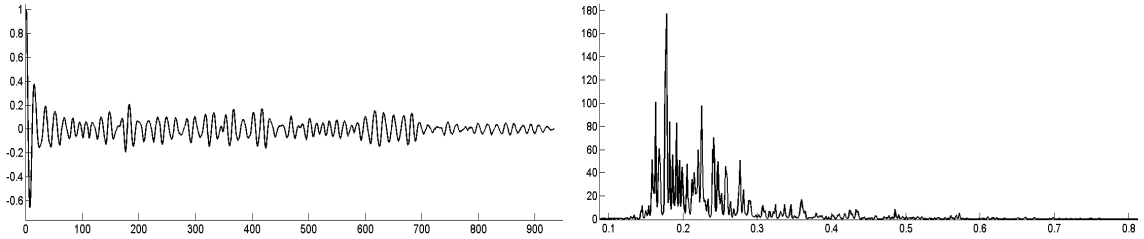

Figure 2: The autocorrelation function (left) and the spectral density (right) estimated from real data

obtained experimentally.

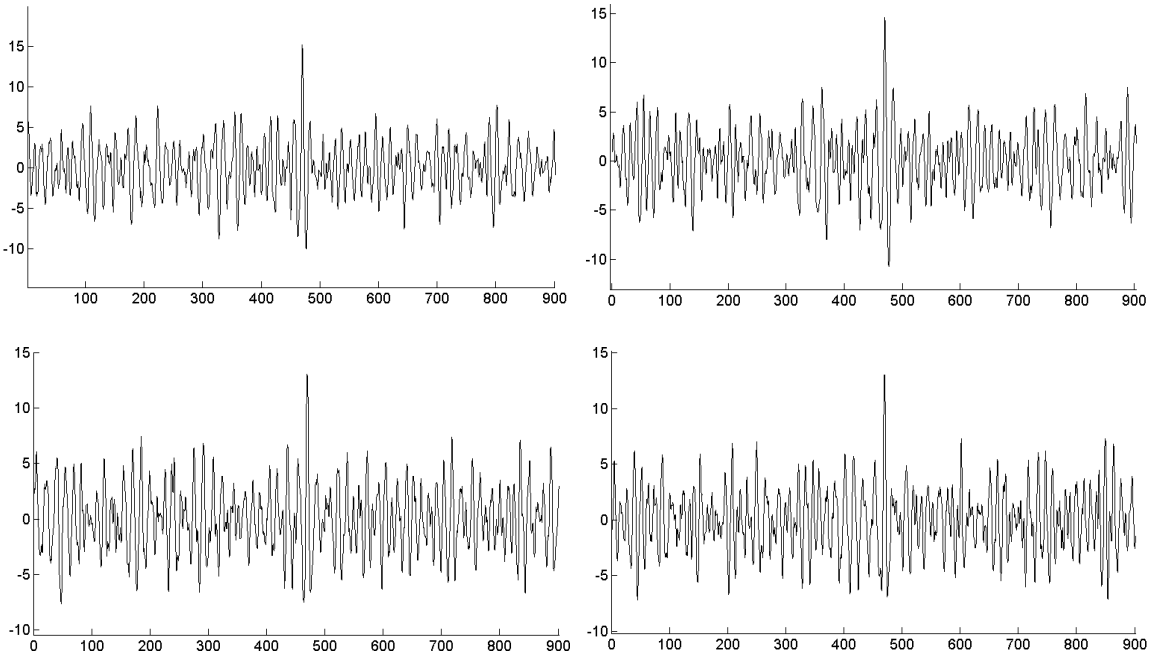

Figure 3: Elevation of the sea surface related to the mean level depending on time (at a fixed point of the plane). The maximal values correspond to the passage of extremely high waves

We calculated the mean number of extremely high waves above a given level according to the sea surface roughness autoregressive model considered here. Some values of $N_t(C)$ are given in Table 3. The result obtained can be interpreted as follows. Assume that a platform is located at a fixed point of the ocean with stationary undulation during 1 year. Then on average 92 waves exceeding 26 m, 18 waves higher then 28 m and 3 waves can reach a height of 30 m will fall on this platform.

Table 3: The mean number $N_t(C)$ of waves above the level C within a year for the sea surface model (5) that were tuned according to the parameters of a real wave

C (m)	$N_t(C)$
5	6.2479×10^5
6	3.1865×10^5
7	1.4379×10^5
8	5.7407×10^4
9	2.0278×10^4
10	6.3378×10^3
11	1.7526×10^3
12	4.2879×10^2
13	9.2810×10^1
14	1.777×10^1
15	3.0126×10^0

3 Conditional spectral models and spatial-temporal simulation of extreme waves

In this Section we present a method of constructing random spatial-temporal implementations of the sea surface with waves of abnormal height. The method proposed is based on the conditional spectral models developed in [3, 4].

Consider a real homogeneous random Gaussian field $w(x)$, $x \in R^k$, with zero mean and unit variance. The spectral representations of a random field can be written in the form:

$$w(x) = \int\limits_P \cos < x, \lambda > \xi(d\lambda) + \int\limits_P \sin < x, \lambda > \eta(d\lambda), \tag{6}$$

where $\xi(d\lambda)$, $\eta(d\lambda)$ are real orthogonal stochastic Gaussian measures of the spectral space P (i.e., P is a measurable set in R^k such that $P \cap P = 0$), $\eta(d\lambda)$ is the spectral measure of the random field $w(x)$, and $< . , . >$, denotes the scalar product in R^k.

The main idea underlying the spectral models is to use an approximation of stochastic integral (6) for the simulation of a random field $w(x)$. In particular, a spectral model can be constructed in the following way. Let us fix some splitting of the spectral space:

$$P = \sum_{j=1}^{n} Q_j, \quad Q_j \cap Q_i = \emptyset \quad \text{for } i \neq j.$$

For approximating (6) consider

$$w_n(x) = \sum_{j=1}^{n} a_j \left[\xi_j \cos < \lambda(j), x > + \eta_j \sin < \lambda(j), x >\right], \quad a_j = \nu^{1/2}(Q_j) \tag{7}$$

Here n is the number of harmonics of the spectral model, ξ_j, η_j are independent standard normal variables, the vectors $\lambda(j) \in P$ belonging to the corresponding sets of the partitioning Q_j. A particular numerical algorithm is determined by the partitioning of the spectral space P and by the choice of $\lambda(j) \in Q_j$. In [2], randomized spectral models were developed so that the vectors $\lambda(j)$ are chosen randomly according to the distribution induced by the spectrum. The randomization allows one to exactly reproduce the spectrum of a random field, however, in this case the numerical model is Gaussian only asymptotically. Non-randomized spectral models (such that the vectors $\lambda(j)$ are not random) have Gaussian finite-dimensional distributions and a discrete spectrum. The information concerning different modifications of spectral models, their properties and convergence can be found, e.g., in [4].

Assume we have to construct a numerical model of a Gaussian random field $w(x)$ with zero mean and the additional condition

$$w(x_m) = b_m, m \in \{1, 2, ..., M\}. \tag{8}$$

A conditional spectral model is a numerical approximation of (7) constructed in the following way. At the first stage we determine the values a_j and simulate the vectors $\lambda(j)$ according to the chosen spectral model. At the second stage we simulate the Gaussian random vector with the components $\xi(j)$,$\eta(j)$, $j = 1, ..., n$, and the conditional distribution generated by equalities (8).

The spectral model of freak waves was constructed using the temporal spectrum and the value of the highest elevation of the New Year's wave with an additional assumption about the angular spectrum. A realization of a spectral model of the freak wave is presented in Figure 4.

Conclusion

It is worth noting that autoregressive and spectral models can be constructed for different types of the sea wind roughness and swell. It is important that the information concerning the roughness spectrum of the sea surface should be as complete as possible. Moreover, conditional spectral models can take into account additional data on the form of an extreme wave if this information is available. In this paper we have described the numerical experiments where the extreme height of the water surface was specified at a single point. In the general case, the conditional spectral model allows us to fix a set of points on the sea surface.

The adequacy of the estimates of the mean number and the probability of extremely high waves obtained on the basis of the outliers theory of random fields is of particular interest. However, one needs extensive statistical data concerning the occurrence of extreme waves in the ocean and the sea roughness spectrumbthat resulted in these waves have been observed in this case.

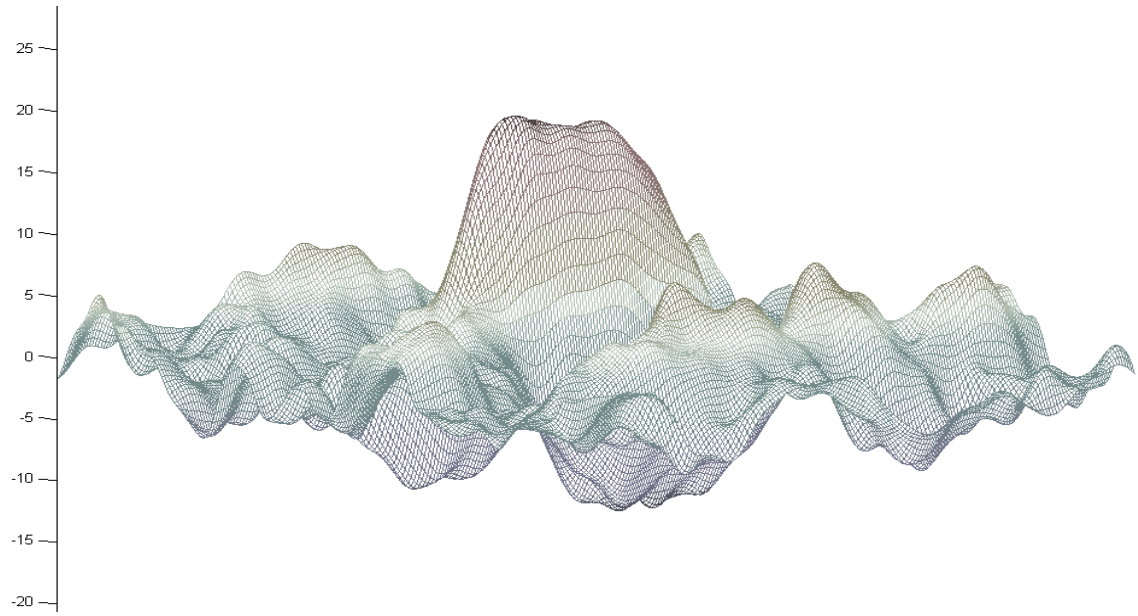

Figure 4: An example of the simulated topography of the freak wave (a spectral model). The wind direction is parallel to the horizontal axis

Acknowledgements

The authors are grateful to A.V. Voitishek, G. A. Mikhailov V. A. Ogorodnikov, S.M. Prigarin for useful remarks and E.N. Pelinovskii, A.V. Slyunyaev the provision of the New Year's wave recording. We take this opportunity to thank the Russian Foundation for Basic Research for the financial support (grants 11-01-00641, 12-05-00169).

References

[1] Anvarov S.R., Prigarin S.M. (1994). Numerical simulation of the spatio-temporal structure of the sea swell surface in optical problems. *Atmospheric and Oceanic Optics*. Vol. 7, No. 5, pp. 361-364.

[2] Mikhailov G.A. (1978). Numerical construction of a random field with given spectral density, *Dokl. USSR Ac. Sci.*, 238, No 4, pp.793–795.

[3] Prigarin S.M. (1998). Conditional spectral models of Gaussian homogeneous fields. *Russian Journal of Numerical Analysis and Math. Modelling*. Vol. 13, No. 1, pp. 57–68.

[4] Prigarin S.M. (2001), *Spectral Models of Random Fields in Monte Carlo Methods*. VSP, Utrecht.

[5] Prigarin S.M., Litvenko K.V. (2011). Numerical simulation of the sea surface and extreme ocean waves with stochastic spectral models. *Proc. of the AMSA-2011 International Workshop Applied Methods of Statistical Analysis Simulations and Statistical Inference, Novosibirsk: Publishing house of NSTU.* pp. 394-402.

[6] Sveshnikov A.A. (1968). Applied Methods of the Theory of Random Functions. *Nauka, Moscow.*

Universal Coding and Decision Trees for Nonparametric Prediction of Time Series with Large Alphabets

Lysyak A.S., Ryabko B.Ya.
Institute of computational technologies SB RAS, Novosibirsk, Russia
Novosibirsk state university, Novosibirsk, Russia
e-mail: accemt@gmail.com, boris@ryabko.net

Abstract

We suggest and experimentally investigate a method to construct forecasting algorithms based on universal measure and decision trees. By the example of predicting currency exchange rates, consumer price index and producer price index we show that the precision of thus obtained predictions is relatively high.

Keywords: universal measure, decision trees, forecasting, prediction, R-method, time series.

Introduction

Tasks of prediction during the modern time had important application-oriented character. Prediction of various processes allowed to solve a wide spectrum of tasks in the modern science and technique. To them the number can carry the analysis of social, economic, geophysical events, a prediction of the natural phenomena and economic events.

Methods of prediction served for research of system communications and regularities of operation and development of objects and processes with usage of the modern methods of information processing and was important means in the analysis of difficult application-oriented systems, operation with the information, purposeful influence of the person on objects of research, for the purpose of increase of effectiveness operation are more their. Most widespread setting of the task of prediction are the task of prediction of time series, . . functions, particular time for axes. The g type of prediction are very significant owing to that the class of tasks correspond to it are widely connect with many problems of economy, geophysics and other areas. These methods played a key role in increases of effectiveness, reliability and quality of technical, economic, biological, medical and social systems.

In the last 2nd decades there were a set of methods of the prediction which had show the extra enough efficiency. In particular, models of machine tutoring [6] who beg to represent a serious competition to classical statistical models in community of experts in prediction [1,2,3].

Despite presence of considerable enough quantity of effective and various methods the predictions connect with power mathematical apparatus (the spectral analysis, the auto regression analysis, Monte Carlo's methods and many other things) and the algorithms connect with creation of expert estimations (the so-called recursive

strategy, which description can be found in [4, 5]), many problems were still far from the resolution.

In this article the new approach to prediction of the one-dimensional and multidimensional time series, bas on particular models of the information theory and methods of data analysis are described. Also, new approaches to a combination of various mathematical methods of prediction, and also methods of acceleration of any mathematical methods of prediction without loss of exactitude of the forecast was offer. Results of experimental researches were described.

1 Setting of the task of prediction

In a general view, the task of prediction of time series could be formulate as follows. Let there are some source generates sequence x_1x_2 of elements from some set A, called alphabet. The alphabet could be the discrete or represent some finite sequence of the continuous interval. Capacity of the alphabet we will designate, as N. It are supposed that allocation of probabilities of characters of this source did not change in due course and did not depend on specific implementation of process (the source are stationary and ergodic). Let the source generated the message of a type $x_1x_(t-1)x_t, x_i \in A \forall i, i = 1, 2, , t$, also are required n following elements. Posteriori deflection amount of the forecast from the real state of object are called as an error of the forecast. Hereinafter under an error of prediction n elements we will understand an average error of the forecast of each of n elements separately. It are clear that the error of the forecast characterized quality of prediction.

It are obvious that if allocation of probabilities of outcomes of process are known in advance the task of prediction of the following values dared simply enough. However in the majority of practical tasks similar prior convergence missed, and not always the given allocation explicitly existed. In the g operation we will consider such case. In the present state of affairs for the decision of the task of prediction it are possible to use the precise estimations of the specified values received by means of statistical techniques, the correlations of serial outcomes of process construct on the basis of the analysis and detection of regularities.

In more common setting of the task of prediction elements x_i could be not only specific numbers (whole or real), and vectors of dimensionality k, where the first element of a vector - value of the predict characteristic of a row, and remain $(k-1)$ attributes - any characteristics of process known for all elements of a row. Give an example. Let there are a row of values of gross national product of the country with an interval in a month. Gross national product was influenc by such parameters, as a rate of inflation, a consumer price index, objyoma of industrial production and many other things. Values of all such characteristics also, as well as value of gross national product, will be known for every month of a row.

Thus, the task of prediction could be, as one-dimensional, and it are a lot of attributive.

2 Method of prediction which is based on the universal measure

In operations [4,7] as the approach for the decision of the task of prediction of time series it are offered to use the methods bas on the general-purpose measure.

Let's result determination of the general-purpose measure, and also we will explain communication between g and described in the previous point approaches. On determination a measure μ are called general-purpose, if for any stationary and ergodic source P the following equalities was true:

$$\lim_{t\to\infty} \frac{1}{t} \left(-\log_2 P\left(x_1...x_t\right) - \log_2 \mu\left(x_1...x_t\right)\right) = 0$$

with probability 1, and

$$\lim_{t\to\infty} \frac{1}{t} \sum_{u\in A^t} P(u) \log_2 \left(P(u)/\mu(u)\right) = 0$$

The g equalities shown that, somewhat, a measure μ are a nonparametric estimation for (unknown) measure P. For this reason the general-purpose measures could be us for an estimation of statistical characteristics of process and prediction.

Now we will describe the universal measure R, which were used as the basis for a method of prediction in our work. The choice of this measure are connect with that it are based on asymptotically optimum universal code of R that are proofed in [8].

Generally, as the general-purpose measure the measure of Krichevsky were t $K_m \geq 0$, which are general-purpose for set of Markov sources with storage, or compendency, m, $m \geq 0$; if $m = 0$, that are a source independent and equally the distributed characters. Somewhat this measure are optimal for this set (see [53]). On determination,

$$K_m\left(x_1...x_t\right) = \begin{cases} \frac{1}{|A|^t}, t \leq m, \\ \frac{1}{|A|^m} \prod\limits_{\vartheta\in A^m} \frac{\prod\limits_{a\in A} \left(\Gamma(\nu_x(\vartheta a)+1/2)/\Gamma(1/2)\right)}{\left(\Gamma(\overline{\nu}_x(\vartheta)+|A|/2)/\Gamma(|A|/2)\right)}, t > m; \end{cases} \tag{1}$$

where $\nu_x(\vartheta)$ - number of sequences ϑ, me in x, $\overline{\nu}_x\left(\vartheta\right) = \sum\limits_{a\in A} \nu_x\left(\vartheta a\right)$, $x = x_1...x_t$, and $\Gamma\left(\right)$ - gamma-function.

Let's define also allocation of probabilities $\{\omega = \omega_1, \omega_2, ...\}$ for whole $1, 2, ...$ as

$$\omega_i = 1/\log(i+1) - 1/\log(i+2) \tag{2}$$

Let's use further it are allocation.

Measure R it are define as

$$R(x_1...x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1...x_t) \tag{3}$$

To count the infinite total in actual algorithms it will not be possible, therefore all calculations we will do according to that each following item in the total (3)

imported all the smaller contribution to total value R. It are possible to prove the g fact without effort how theoretically, and to check up empirical by. Owing to the g property we will use the first m the compos totals, and number m called a depth of the analysis.

So, value of a measure R calculate on the basis of the formula (3), could serve as an estimation of probability of an outcome of process and be us for the decision of the task of prediction.

Let's consider now the circuit of prediction on the basis of the general-purpose measure both for sources on the discrete, and on the continuous alphabet the continuous.

In the beginning we will consider a source generates values from the finite alphabet. In this case the circuit of actions are sufficient are simple. Let x_1x_t - available sequence. For everyone $a \in A$ let's construct sequence x_1x_t a also we will calculate the conditional probability on the basis of a measure of R:

$$R\left(a|x_1...x_t\right) = R\left(x_1...x_ta\right)/R\left(x_1...x_t\right)$$

It are possible to use received by thus way elements $a \in A$ values as estimations of appropriate unknown probabilities $P(x_1x_ta)$. Value a, has the maximum estimation also will be look-ahead value.

Let's consider now the circuit of prediction for a source from the continuous interval. Let there are the continuous interval $[A, B]$. Then let $\{\Pi_n\}, n \geq 1$ - increas sequence of finite partitions of an interval $[A, B]$ on n parts. In our case the partition of intervals were produced uniformly on equal subintervals, i.e. the size of each subinterval are defined, how $h = (B - A)/n$. The substantiation of a choice of such method will be g further.

Let's define now an estimation of density of probabilities r as:

$$r(x_1...x_t) = \sum_{s=1}^{\infty} \omega_s R\left(x_1^{[s]}...x_t^{[s]}\right) \tag{4}$$

As shown in [4,8], density $r(x_1x_t)$ are an estimation of unknown density $p(x_1x_t)$, and appropriate conditional density

$$r\left(a|x_1...x_t\right) = r\left(x_1...x_ta\right)/r\left(x_1...x_t\right) \tag{5}$$

are a suitable estimation $p(a|x_1x_t)$. The amount of items in the total (4) at implementation of the algorithm described further, as well as in the discrete case, are equal to depth of the analysis m.

3 A method of prediction on the basis of decision trees

In a general view setting of the task for decision trees looked as follows. Let there's a set of objects A (set A consists of N the objects ma so-called learning sequence), which

possess particular independent characteristics (attributes with finite set of values; all are available (M+1) attributes). Set of the first M attributes we will designate, as Q. For the given set A all (M+1) attributes was known. For other (new) elements on known to the first M attributes need to find target (M+1)-th attribute. Thus on an input number N (elements in learning sequence), number M, parameter $m \in M$.

As a rule, the described method are applicable to tasks of classification and clustering. In the g operation the approach who showed a method of data application of trees to prediction of time series are offer. The tree of decision-making are under construction on the algorithm described more low.

Firstly, let's define some important definitions.

Definition 1. Entropy: S – target attribute. $H(A,S) = -\sum\limits_{i=1}^{Sn} \frac{|A_i|}{|A|} \log_2 \frac{|A_i|}{|A|}$ where A_i – elements from A, at which the attribute of S are equal to i (and $|A| = N$).

Definition 2. The increase of the information – are define for each attribute from Q in relation to target attribute of S and showed, what of attributes of Q gives the maximum increase of the information concerning value of attribute of S (i.e. concerning a class of an element). The increase of the information are defined by the following formula:

$$Gain(A,Q) = H(A,S) - \sum_{i=1}^{Qn} \frac{|A_i|}{N} H(A_i,S).$$

Further, we will describe immediately one of most effective algorithms of creation of the decision tree, called ID3, dependent on set A, target attribute S and sets of attributes Q:

1. To create a root of a tree.
2. If S are equal to any q on all elements from A, to mark in a root q and to quit.
3. If $Q = \{\emptyset\}$, than select such q from the set S, which is the number, which is equal to the greatest quantity of elements from A, and to deliver q in a root and than quit.
4. To select $q \in Q$, for which Gain(A, q) it are maximum.
5. To mark in a root of a tree of q.
6. For each value q_i attribute of q:
   (a) To add the new descendant and to mark a proceed edge with a label q_i.
   (b) If in A there was no elements for whom value of q are equal to q_i, that to arrive according to item 3.
   (c) Else launch ID3 (Aq_i, S, Q\q) and to add it result, as the subtree with a root in this offspring.

The tree are under construction before exhaustion of learning set or to emptiness of set of Q. Also, in offer implementation of the g algorithm it are possible to restrict depth of a tree artificially – separate parameter. After the maximal depth of a tree are achieved, we go to the step 3 of the ID3 algorithm.

4 The results

Both described a method implement on a supercomputer and test on forecasts of the actual data. Results of the g forecasts was result in following tables. It are necessary to note that at implementation of algorithm ID3 there restrictions on the maximum depth of a tree about whom it were t in chapter 3. It are restriction it are showed in the table by the first in a third column. After a sign "/" there are a parameter m methods R and decision tree, which shows the depth of a sequence analysis (see chapter 2 and 3). Thus all researches it were le in two modes: on-line, which meant prediction for 1 step forward on 10 different sequences with the subsequent averaging of an error and prediction for 10 steps forward on 10 different sequences with the subsequent averaging of an error. Thus prediction for 10 steps forward were considered as follows: the next value and sequence were predict were replenish by look-ahead value and so we continued to a 10th look-ahead element then we considered an error. It are important to mark that prediction were fulfill not absolute values of sequence, and a difference between adjacent elements: such approach allowed to lower essentially the necessary size of the continuous interval in whom look-ahead values lain; and also allowed to reveal the linear trends (and the periods on them). It were impossible at prediction of absolute values of a temporary row.

In each of the following below tables parameters was result: size of sequence, amount of parts of a partition of the continuous interval and size of an error of a method.

In the Table 1 the data of the forecast of exchange currency rates of euro/dollar with a time interval (timeframe) one day (D1) and the period are cit 6/20/2012 - 7/11/2012. The graph of this time series are contained in a Figure 1.

In the following two tables the data of the forecast of the common overall consumer price index on territory of the USA during the period are reflect 01.1990-02.2013 with a time interval 1 month. Values predict with 02.2012 to 02.2013. Thus in Table 2 results of the forecast of the last part of a row in which the dispersion of sequence are explicit more than in the middle (it are visible from a figure 2) was reflect. In Table 3 results of the forecast of the central part of the schedule which looked more smooth was show. Apparently from tables, results of the forecast of a row with more low dispersion (table 3) are notable more better (despite the smaller size of sequence) that are quite natural.

.

In Table 4 the data of the forecast of the common index of the industrial prices for territories of the USA during 03.2002 - 02.2013 are cit.

Table 1

Size of sequence L	Partitioning n	Max. depth of tree/ m	Decision tree On-line	R-measure On-line	Decision tree 10 steps	R-measure 10 steps
500	10	2/2	0,0079	0,0084	0,0103	0,0299
		5/5	0,0095	0,0084	0,0151	0,0299
	20	2/2	0,0088	0,0083	0,0105	0,0159
		5/5	0,0084	0,0083	0,0105	0,0159
	50	2/2	0,0089	0,0083	0,0119	0,0187

Table 2

Size of sequence L	Partitioning n	Max. depth of tree/ m	Decision tree On-line	R-measure On-line	Decision tree 10 steps	R-measure 10 steps
277	5	2/2	0,628	0,733	1,670	1,671
	10	2/2	0,602	0,602	0,572	0,573
		2/5	0,825	0,602	1,717	0,573
	20	2/2	0,551	0,701	0,718	0,884
		2/5	0,609	0,701	1,075	0,884
	100	2/2	0,615	0,655	0,753	1,497
		2/5	0,812	0,655	1,858	1,497

Table 3

Size of sequence L	Partitioning n	Max. depth of tree/ m	Decision tree On-line	R-measure On-line	Decision tree 10 steps	R-measure 10 steps
157	5	2/2	0,495	0,364	1,684	1,684
	10	2/2	0,394	0,328	1,255	1,255
		2/5	0,394	0,328	1,255	0,255
	20	2/2	0,408	0,310	0,424	0,424
		2/5	0,403	0,310	0,522	0,424
	100	2/2	0,408	0,310	0,715	0,715
		2/5	0,408	0,310	0,715	0,715
	240	2/5	0,337	0,297	0,397	0,853

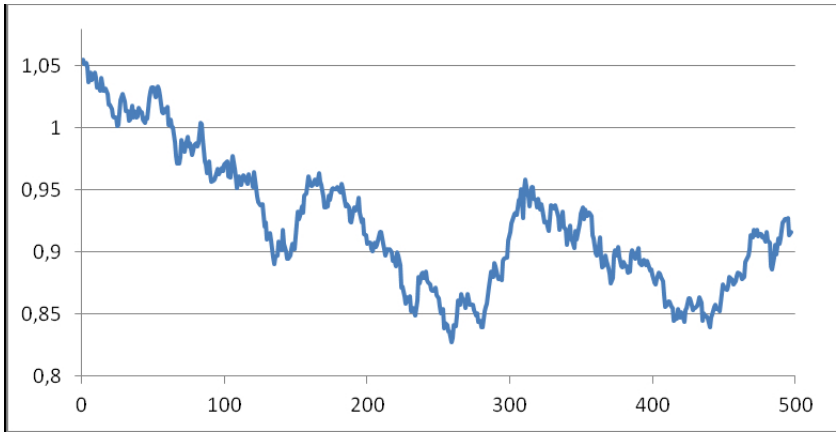

Figure 1: Exchange currency rates EUR/USD (timeframe: D1)

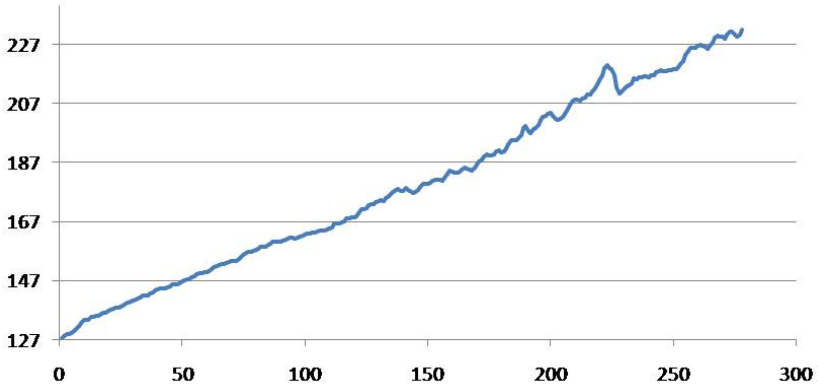

Figure 2: Producer price index

Conclusions

Apparently from the results described above, the both method demonstrate rather high degree of accuracy on the actual data. It is also true speaking about comparison to other methods of prediction that explicitly are show in [4]. Besides, more effective implementation of a method were invented for R, which allows to reduce complexity determine of various variants of look-ahead values from the alphabet. Complexity decreasing to constant concerning length of the alphabet (early, it were linear) that allowing reach the same work speed of the algorithm that earlier were reach only using a supercomputer.

Besides, the method based on the universal measure gives about the same accuracy, as decision trees, but thus it allows to generalize easily the forecast to multidimensional time series, where each element had not only value, but still other known attributes.

Table 4

Size of sequence L	Partitioning n	Max. depth of tree/ m	Decision tree On-line	R-measure On-line	Decision tree 10 steps	R-measure 10 steps
277	5	2/2	0,954	0,978	1,87	3,276
	10	2/2	0,996	0,978	1,363	3,276
		2/5	1,322	0,978	1,793	3,276
	20	2/2	1,07	0,910	0,935	1,363
		2/5	1,125	0,910	1,076	1,363
	100	2/2	1,013	0,910	1,204	1,122
		2/5	1,296	0,910	1,488	1,122

References

1. Ahmed N. et al. An empirical comparison of machine learning models for time series forecasting//Econometric Reviews. 2010. V. 29, Issue 5-6. P. 594-621.

2. Palit A. K., Popovic D. Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications (Advances in Industrial Control). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

3. Zhang G., Patuwo B. E., Michael Y. H. Forecasting with articial neural networks:: The state of the art//International Journal of Forecasting. 1998. V. 14, Issue 1. P. 35-62.

4. Prefix of the Item. And. An experimental research of a method of the prediction bas on the universal codes//Vestnik SIBGUTI, 2010, 4, s.26-35.

5. Cheng H. et al. Multistep-ahead time series prediction//Lecture Notes in Computer Science. 2006. V. 3918. P. 765-774.

6. Bontempi G. Local Learning Techniques for Modeling, Prediction and Control. Ph.d., IRIDIA-Universit de Libre de Bruxelles, BELGIUM, 1999.

7. B. Ryabko. Applications of Universal Source Coding to Statistical Analysis of Time Series. \\Isaac Woungang, Sudip Misra, Subhas Chandra Misra (Eds.), "Selected Topics in Information and Coding Theory", Edited Book, Series on Coding and Cryptology, World Scientific Publishing, 2010.

8. B.Ya.Dvazhdy's Ryabko the general-purpose coding//Problems of information transfer. 1984. . 20, 3. With. 24-28.

Simulation of Random Variates by Approximation

Oleg A. Makhotkin
The Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia,
e-mail: oam@osmf.sscc

Abstract

In this paper, the approximation of a "slow-simulated" probability density by a special "fast-simulated" density is considered. It is possible when the bias caused by the approximation is sufficiently smaller than a statistical error. For a one-dimensional case, the usefulness of finite elements is demonstrated. As an example, the simulation algorithms, using different approximation methods for the Makeham distribution density, are tested.
Keywords: random variates, simulation, approximation, B-splines, Makeham.

Abbreviations and Notations

BF - Basic Function,	DF - Distribution Function,
BRV - Basic RV: independent RV $U, U_1, U_2, \ldots$	with PDD $p_u(x) = 1,\ 0 < x < 1$,
DRV - Discrete RV,	PD - Probability Distribution,
PC - Piecewise Constant,	PL - Piecewise Linear,
PDD - Probability Distribution Density,	RV - Random Variate,
$\xi \sim p(x)$ - RV ξ **has** PDD $p(x)$,	or ξ is **simulated** according to $p(x)$,
$< f(x)\|X >= \int_X dx f(x)$,	$\chi(x\|a, b) = 1$ for $a < x < b$, else $= 0$.

1 Introduction

It is well known that efficient random number generation algorithms are the key technical elements in the large-scale statistical simulation. The tutorials describe five methods for simulation of a RVs (see, for example, [2], [3]):

1. **The inversion method:** if for one-dimensional PDD $p(x)$ with DF $F(x)$ there exists a computer algorithm for the inversion $F^{-1}(y),\ y \in [0, 1]$, then RV $\eta = F^{-1}(U)\ \sim p(x)$.

2. **The rejection method:** assume that for a multi-dimensional PDD $p(\mathbf{x})$ there exists a majorant $g_+(\mathbf{x})$ such that 1) $\bar{n} =< g_+|X > < \infty$, 2) there exists the efficient algorithm for the simulation of a RV $\xi \sim p_+(\mathbf{x}) = g_+(\mathbf{x})/\bar{n}$. Then the Markov process **Repeat** $\xi \sim p_+(\mathbf{x})$ **Until** $U \cdot g_+(\xi) \le p(\xi);\ \eta := \xi$ generates the RV $\eta \sim p(\mathbf{x})$.

3. **The composition method:** assume that a PDD can be represented in the form $p(\mathbf{x}) = \sum_{i=1}^{M} \Pi_i p_i(\mathbf{x})$, where $\Pi_i \ge 0,\ \sum_{i=1}^{M} \Pi_i = 1$, and $\{p_i(\mathbf{x})\}_{i=1}^{M}$ is a set of conditional densities. Then the algorithm $DRV\ j \sim \{\Pi_i\}_{i=1}^{M};\ \eta \sim p_j(\mathbf{x})$; generates the RV $\eta \sim p(\mathbf{x})$.

4. **The approximation method:** the main idea of this method is the approximation of a given "hard" PDD $p(x)$ by PDD $\hat{p}(x)$ with a more efficient simulation algorithm. The only one example, represented in the published works (see [3]), is the approximation of a standard normal RV by the sum of BRVs

$$\hat{\eta}_m = \sqrt{12/m}\Big[\sum_{i=1}^{m}(U_i - 1/2)\Big].$$

The second example of the approximation method is the Metropolis–Hastings algorithm [4]. In this algorithm, for the simulation of RV $\eta \sim p(\mathbf{x})$ one uses the Markov process: $\xi_0 \sim p_0(\mathbf{x})$, for $k = 1, 2, \ldots$ $\xi_k \sim q(\mathbf{x}; \xi_{k-1})$. A special selection of p_0 and q ensures the convergence: for $k \to \infty$ $\xi_k \to \eta \sim p(\mathbf{x})$. In practice, one stops the Markov process at a sufficiently large step K, using a certain empirical stopping rule. Thus, RV $\xi_K \sim p_K$ is the approximation of RV $\eta \sim p$.

5. **The method of special transformations:** in this method, one obtains a realization of m-dimensional RV η by $m \times n$ by transformation of BRVs: $\eta_i = g_i(U_1, \ldots, U_n)$, $i = 1, \ldots, m$. The most known example is 2×2 transformation (the Box–Muller algorithm)

$$\eta_1 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2), \;\; \eta_2 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2),$$

which gives two independent RVs with a standard normal distribution.

2 Statement of the problem

It is well known that every non-negative function $p(\mathbf{y}) \geq 0$, $\mathbf{y} \in Y \subset \mathbb{R}^n$ with $\int_Y p(\mathbf{y})d\mathbf{y} = 1$ can be considered as PDD. The distance between two PDDes $p(\mathbf{y}), g(\mathbf{y})$ can be defined as

$$L = ||p - g|| = \int_Y |p(\mathbf{y}) - g(\mathbf{y})|d\mathbf{y}.$$

Only this distance is invariant under the reversible transformations of the independent variables $\mathbf{y} = \mathbf{f}(\mathbf{x})$:

$$||p(\mathbf{y}) - g(\mathbf{y})|| = \int_Y |p(\mathbf{y}) - g(\mathbf{y})|d\mathbf{y} = \int_{X(Y)} |J(\mathbf{x})| \cdot |p(\mathbf{f}(\mathbf{x})) - g(\mathbf{f}(\mathbf{x}))|d\mathbf{x} =$$
$$= \int_{X(Y)} |p(\mathbf{x}) - g(\mathbf{x})|d\mathbf{x} = ||p(\mathbf{x}) - g(\mathbf{x})||,$$

where $X(Y)$ is the image of the set Y, and $J(\mathbf{x})$ is the Jacobian of transformation.

Is it possible, in principle, in some statistical problems to use a sample of approximate RV $\hat{\xi} \sim \hat{p}$ instead of a sample of exact RV $\xi \sim p$? Two examples below show that it is possible.

1. Let us consider the estimation of the integral $I = \int_X f(x)p(x)$ using Monte Carlo method. If we use the sample $\{\hat{\xi}_i \sim \hat{p}(x)\}_{i=1}^N$, then the mean-value estimator

$$\hat{\eta} = \sum_{i=1}^N f(\hat{\xi}_i)/N$$

will be biased:

$$\hat{I} = \mathbf{M}\hat{\eta} = \int_X f(x)\hat{p}(x) = I + \int_X f(x)(\hat{p}(x) - p(x)) = I + b.$$

If $c \leq f(x) \leq d$, then $|b| \leq (d-c)L/2$. For $L = \delta \ll 1$, a relative bias will also be small: $|\hat{I} - I|/|I| < \delta(d-c)/(2|I|) \ll 1$. The mean-square error of $\hat{\eta}$ is equal to

$$\mathbf{M}(\hat{\eta} - I)^2 = \mathbf{M}(f(\hat{\xi}) - \hat{I})^2/N + b^2 = \hat{\sigma}^2/N + b^2.$$

If $|b| \leq \delta_* \hat{\sigma}/\sqrt{N}$, where δ_* is an admissible bias level, then using approximate RVs do not considerably increase a common error.

2. The Scheffe theorem [6] states that

$$||\hat{p}(x) - p(x)|| = 2\sup_B |\int_B \hat{p}(x)dx - \int_B p(x)dx| = 2\sup_B |\hat{P}(B), -P(B)|,$$

where supremum is taken over all the Borel sets $B \subset X$. We cannot use an approximate RV for the estimation of $P(B)$ if $L/2 \geq |\hat{P}(B) - P(B)| \geq \delta_P P(B)$. So, only the probability of an event which satisfies the inequality $P(B) \geq L/(2\delta_P)$ will be estimated with a relative error δ_P.

State of the problem: for every PDD $p(x)$, $x \in X$ construct an approximate PDD $\hat{p}_m(x)$ with the properties :

1. **For every error level δ it is possible to find m such that $||\hat{p}_m(x) - p(x)|| \leq \delta$.**

2. **There exists an efficient algorithm for the simulation of RV $\hat{\eta} \sim \hat{p}_m(x)$ with run-time complexity, practically, independently of m.**

3 Solution of the problem

It is well known that the one-dimensional RVs have the most diverse set of PDDes. Thus let us consider in detail the approximation of one-dimensional PDD $p(x)$, defined on a standard interval $X = [0, 1]$. Other intervals on the real line can be transformed to a standard one. The following transformations can be used:

1. $Y = (a, b)$: $\quad y = a + (b - a)x$.

2. $Y = (0, +\infty)$: $y = x/(1-x)$; $y = \mathrm{tg}(\pi \cdot x/2)$. These transformations can be used for the gamma distribution.

3. $Y = (-\infty, +\infty)$: $y = z/(1-|z|)$, $z = 2x-1$; $y = \mathrm{tg}(\pi \cdot (2x-1)/2))$. These transformations can be used for the normal distribution.

4. The best transformation is $y = F^{-1}(x)$, where $F(y)$ is the DF for $p(y)$. It transforms $p(y) \to p_u(x) = 1$ for $0 < x < 1$.

For creating an efficient simulation algorithm, we should find the approximation of $p(x)$ by the sum of the BFs $\{\varphi_i(x),\ x \in X\}$:

$$\hat{p}(x) = \sum_{i=1}^{M} c_i \varphi_i(x). \tag{1}$$

The following conditions must be implemented:

1. $c_i \ \geq \ 0, \quad i = 1, \ldots, M.$

2. For $i = 1, \ldots, M \quad \varphi_i(x) \ \geq \ 0, \quad < \varphi_i(x) | X >= W_i \ < \ \infty.$

3. For the every M, the conditional PDDes $p_i(x) = \varphi_i(x)/W_i$ must have a simulation algorithm with the run-time complexity $O(1)$ with respect to M.

If these conditions are fulfilled, then sum (1) can be transformed to a weighted sum of a partial PDDes

$$\hat{p}(x) = \sum_{i=1}^{M} c_i \varphi_i(x) = \sum_{i=1}^{M} [c_i W_i][\varphi_i(x)/W_i] = \sum_{i=1}^{M} \Pi_i p_i(x). \tag{2}$$

The superposition algorithm for the simulation of RV with PDD (2) has the form:

$$\eta \sim \hat{p}(x) = \sum_{i=1}^{M} \Pi_i p_i(x) : \ DRV\ j \sim \{\Pi_i\}_{i=1}^{M}; \quad \eta := (\eta_j \sim p_j(x)). \tag{3}$$

a Properties of B-splines

The investigations carried out, have been shown that one-dimensional finite elements are the best choice for the approximation of PDDes. On the uniform grid $X_m = \{x_i = h \cdot i,\ h = 1/m\}_{i=0}^{m}$ they can be represented as uniform B–splines [1] $B_{j,k}(x),\ j = 1, \ldots, M = m + k - 1,\ k \geq 1$. These splines are defined on the extended grid $T_n = \{t_j = h(j-k),\ j = 1, \ldots n = m + 2k - 1\}$ so $B_{j,k}(x) > 0$ for $t_j < x < t_{j+k}$.
The B–splines have the useful properties:

1) $B_{j,k}(x) = B_{l,k}(x - t_j + t_l)$, $W_{j,k} = h,\ j = k, \ldots, m,$

2) $< B_{j,k}(x) | [0,1] >= W_{j,k}$, $W_{j,k} = W_{M-j+1,k},\ j = 1, \ldots, k-1.$

3) $B_{j,k}(x|T_n) = \mathcal{B}_k(y(x)|E_{k+1})$, $y(x) = (x - t_j)/h,\ j = 1, \ldots, M.$

Here $\mathcal{B}_k(y(x)|E_{k+1})$ means standard uniform B–splines defined on the grid $E_{k+1} = \{0, 1, \ldots, k\}$. In [5], it has be shown that $\mathcal{B}_k(y|E_{k+1})$ is the PDD of RV $\xi_k = U_1 + \ldots + U_k$. The RV $\eta_{j,k} \sim B_{j,k}(x)/h$ for $j = k, \ldots, m-1$ can be simulated as $\eta_{j,k} := t_j + h\sum_{l=1}^{k} U_l$. For the boundary conditional PDDes, $p_{j,k}(x) = B_{j,k}(x)\chi(x|0,1)/W_{j,k}$ with $j = 1, \ldots, k-1$ and $j = M-k, \ldots, M$, one needs special simulation algorithms. For example, for $k = 2$, there are two boundary PDDes: $p_{1,2}(x) = 2(h-x)/h$ and $p_{m,2}(x) = 2(x - x_{m-1})/h$. The boundary RVs can be simulated by algorithms[1] $\eta_{1,2} := h|U_1 - \tilde{U}_2|, \quad \eta_{m,2} := 1 - h|U_1 - \tilde{U}_2|$. For $M = m+1 \gg 1$, there are only 2 boundary RVs with non-standard simulation algorithms, other $m-1$ RVs can be simulated by a single standard algorithm.

b Examples of non-suitable basic functions

Two counter-examples show that the positivity of the BFs is not sufficient for the existence of efficient simulation algorithms.

Counter-example 1: the Bernstein polynomials are the polynomials

$$B_n(x) = \sum_{k=0}^{n} f(k/n) C_n^k x^k (1-x)^{n-k} = \sum_{k=0}^{n} f(k/n) b_{k,n}(x),$$

where $\{b_{k,n}(x)\}_{k=0}^{n}$ form the basis for the vector space of polynomial of degree at most n. For PDD $p(x) \in C^1[0,1]$ and $n \to \infty$ $\hat{p}_B(x) = B_n(x)/ < B_n(x) > \to p(x)$. Computer-aided experiments show that for sufficiently large n $||\hat{p}_B(x) - p(x)|| = O(1/n)$. But the conditional PDDes $p_k = (n+1)b_{k,n}(x)) = \gamma_k(x),\ k = 1, ldots, n-k+1)$ have no standard representation, so the complexity of the simulation algorithms grows with increasing n.

Counter-example 2: $\hat{p}_r(x)$, the rational approximation of PDD $p(x)$ on the intervals $[x_i, x_{i+1}]$, uses a local BFs $\psi_1(t) = (1-t)/(1+t), \quad \psi_2(t) = 2t/(1+t)$, where $t = (x - x_{i-1})/h$. These functions also approximate a continuous PDD $p(x)$ on $[0,1]$: for sufficiently large m $||\hat{p}_r(x) - p(x)|| = O(1/m)$. But RVs with PDDes $p_1(t) = \psi_1(t)/ < \psi_1 >,\ p_2(t) = \psi_2(t)/ < \psi_2 >$ can be simulated only by the rejection method. For the majorants $g_{1,+}(t) = 1 - t,\ g_{2,+}(t) = 2t$ the mean numbers of steps until success equals $\bar{n}_1 = 1/ < 2\psi_1 >= 1.29$ and $\bar{n}_2 = 1/ < \psi_2 >= 1.63$.

c The efficient algorithms for simulation of DRV

The simplest algorithm for the simulation of DRV j in (3) – the sequential search, uses the auxiliary array $F_i = \sum_{k=1}^{i} \Pi_k,\ i = 1, \ldots M,\ F_0 = 0$:

$$j := 0;\ U \sim p_u(x): \quad Repeat\ j := j+1\ Until\ (U \le F(j));$$

The mean number of comparisons $\bar{n} = \sum_i i\Pi_i$. For $\Pi_i = 1/M$ $\bar{n} = (M+1)/2$. A more realistic discrete distribution gives the PL-approximation of test PDD $p_s(x) = (1+s)x^s$, $s > 0$. The mean number of comparisons $\bar{n} = (s+1)h^{1+s}\sum_{i=1}^{m} i^{1+s} - (1+s)/2$.

[1] $\tilde{U}$ means BRV, used in the algorithms instead of RV $1 - U$.

Calculation of the sum using the Euler-Maclaurin formula gives $\bar{n} = [(s+1)/(C(s+2))]m + O(1)$, where $C = O(1)$.

A more efficient algorithm gives the index search, in which a second auxiliary integer array $MG[0..mi]$ is used: $MG[0] := 0$, values of $MG[k],\ k = 1, \ldots, mi$ are defined as solutions of the inequalities $F(MG(k-1)) < k/mi \leq F((MG(k)),\ k = 1, \ldots mi$.

Index search algorithm:

$$k := [mi * U];\ j := MG(k);\ Repeat\ j := j + 1\ Until\ (U \leq F(j));$$

Values of $\bar{n}$ were estimated for the PL-approximations of the test PDD $p_s(x) = (s+1)x^s,\ s > 0$. The samples of $N = 1000$ DRVs were used. The results are presented in Table 1.

Table 1: Mean comparison numbers for index search.

	m	10	20	40	80	160	320
	mi	4	8	16	32	64	128
$s = 0.5$	$\bar{n}$	2.03	2.25	2.17	2.22	2.29	2.26
$s = 3.5$	$\bar{n}$	2.59	2.55	2.52	2.50	2.85	2.59

For $s = 0.5$, $m = 10$, $mi = 10$, the value $\bar{n} = 1.46$ was obtained.
The Walker method gives the fastest algorithm for the simulation of DRVs. In this algorithm, two additional arrays are constructed for a given PD $\{\Pi_i\}_{i=1}^{M}$: the real array $F(1..M),\ 0 \leq F(i) \leq 1$, and the integer array $ia(1..M),\ 1 \leq ia(i) \leq M$(see [8] for details). The simulation algorithm uses these two arrays:

$$s := M * U_1;\ j := 1 + Trunc(s);\ U_2 := Frac(s);\ If\ (U_2 \geq F(j))\ Then\ j := ia(j);$$

Here the function $Trunc(s)$ gives the integer part of s, while $Frac(s)$ – the fractional part. The comparison of (CPU-time(N)/N) (sec) for 4 simulation methods is presented in Table 2: SS - Sequential Search, BS - Binary Search, IS - Index Search, WS - Walker Simulation. Computer Characteristics: Intel Pentium 4 , $\omega = 2.0\ GHz$, Borland Pascal , the sample volume N=50000. The test PD were: $a_i := U_i, i = 1, \ldots, m$, $\Pi_i := a_i / \sum_{l=1}^{m} a_l$.

Table 2: Comparison of different methods for simulation of DRV.

M	SS	BS	IS	mi	WS
40	6.70E-6	2.53E-6	1.86E-6	10	1.27E-6
80	1.28E-5	2.86E-6	1.92E-6	20	1.26E-6
160	2.45E-5	3.29E-6	1.98E-6	40	1.26E-6
320	5.07E-5	3.63E-6	1.97E-6	80	1.27E-6

d Approximation methods

There are two methods for constructing the approximation of PDD $p(x)$: optimal and non-optimal.

The optimal PDD approximation: coefficients in approximate PDD are found as solution the following extremal problem($W_i =< \varphi_i(x) >$)

$$\begin{aligned} \text{Criteria:} \quad & L(c) = \int_0^1 |\sum_{i=1}^M c_i\varphi_i(x) - p(x)|dx \ \rightarrow \ \min_c, \\ \text{Constrains:} \quad & c_i \ \geq \ 0, \ i = 1, \ldots, M; \quad \sum_{i=1}^M W_i c_i = 1. \end{aligned} \tag{4}$$

The optimization problem has always a solution(perhaps, not only one). The computer-aided solution of (4) needs special algorithms because values of $L(c)$ and its gradient can be calculated only by the numerical integration.

Non-optimal PDD approximation:

$$\begin{aligned} \tilde{p}(x) \ &= \ \sum_{i=1}^M a_i\varphi_i(x), \ x \in X. \\ \text{Approximation methods:} \quad & \text{for } j = 1, \ldots, M \\ \text{Interpolation:} \quad & \sum_{i=1}^M a_i\varphi_i(y_j) = p(y_j). \\ \text{Projection:} \quad & < \varphi_j(x) \cdot [\sum_{i=1}^M a_i\varphi_i(x) - p(x)]|X >= 0. \\ \text{Normalization:} \quad & d_i = \max[0, a_i], \ S = \sum_{i=1}^M d_i W_i, \\ & \hat{p}(x) = \sum_{i=1}^M \frac{d_i}{S}\varphi_i(x) = \sum_{i=1}^M c_i\varphi_i(x). \end{aligned}$$

4 Simulation of RV with the Makeham distribution

The Makeham probability distribution density(W.M. Makeham, 1889) has the form

$$\begin{aligned} p(y) \ &= \ \mu(y)\exp(-\psi(y)), \quad y \geq 0, \\ \mu(y) \ &= \ A + Hy + B\exp(\alpha y), \quad \mu(y) \geq 0, \\ \psi(y) \ &= \ \int_0^y \mu(z)\,dz = Ay + \frac{H}{2}y + \frac{B}{\alpha}(\exp(\alpha y) - 1). \end{aligned} \tag{5}$$

The change of the argument $x = y/Y_*$ transforms the density (5) again to the interval $(0, \infty)$: $p(x) = Y_* p(Y_* x)$. The transformed PDD $p(x)$ again has the form (5)

$$\begin{aligned} p(x) \ &= \ \bar{\mu}(x)\exp(-\bar{\psi}(x)), \\ F(x) \ &= \ \int_0^x p(r)dr = 1 - \exp(-\bar{\psi}(x)), \quad x \geq 0. \end{aligned} \tag{6}$$

where $\bar{\mu}(x) = Y_*\mu(Y_*x)), \quad \bar{\psi}(x) = \int_0^x dr\, \bar{\mu}(r))$.
The approximation $\breve{p}(x) = p(x)/F(1),\ 0 \le x \le 1, \quad \breve{F}(x) = F(x)/F(1)$ gives PDD on a standard interval $(0,1)$. The inversion method for the simulation of RV with PDD $\breve{p}$ needs the solution of the nonlinear equation $\bar{\psi}(x)) = S(U) = -\ln(1-F(1)U)$.

The following four methods for the approximation of the Makeham PDD were investigated:

1. PC approximation of $\bar{\mu}(x)$ on a uniform grid with m intervals. One needs to invert the PL function $\tilde{\psi}_{PL}(x) = \int_0^x \tilde{\mu}_{PC}(r)dr$.

2. PL approximation of $\bar{\mu}(x)$. One needs to invert the piecewise quadratic function $\tilde{\psi}_{PQ}(x)$.

3. PC approximation of PDD $\breve{p}(x)$. One needs to invert the PL DF $\hat{F}_{PC}(x) = \int_0^t \hat{p}_{PC}(r)dr$.

4. PL approximation of PDD $\breve{p}(x)$. One needs to invert the piecewise quadratic DF $\hat{F}_{PQ}(x)$.

The distance between the PDD $\breve{p}(x)$ and its approximation is calculated as $L = \int_0^1 |\breve{p}(x) - g(x)|dx$, where $g(x) = \tilde{\mu}(x)\exp(-\tilde{\psi}(x)))$ for cases 1,2 and $g(x) = \hat{p}(x)$ for cases 3,4. Table 3 presents the distance $L(m)$ as the function of m - number of grid intervals. The PDD (5) was used with parameters $A = 6.0e-2, H = -4.0e-3, B = 8.0e-3, \alpha = 0.08$ The distances were calculated by the 2-point compound adaptive Gauss quadrature with the relative error 1.0e-5. In the last row of Table 3, the least square approximations of $L(m)$ are presented.

Table 3: Errors of different methods for approximation of test PDD.

m	1	2	3	4
8	1.75e-1	4.33e-2	2.03e-1	7.14e-2
16	8.75e-2	1.10e-2	1.03e-1	1.87e-2
32	4.37e-2	2.77e-3	5.15e-2	4.84e-3
64	2.18e-2	6.92e-4	2.57e-2	1.21e-3
128	1.09e-2	1.73e-4	1.29e-2	3.03e-4
$L_{apr}(m)$	$1.41m^{-1.00}$	$2.74m^{-1.99}$	$1.61m^{-0.995}$	$4.38m^{-1.97}$

a The run–time complexity of simulation algorithms

The following 7 algorithms for the simulation of RV with the Makeham distribution were tested:

A1: inversion of DF

$$\breve{F}(x) = (1 - \exp(-\bar{\psi}(x))/F(1), \quad F(1) = 1 - \exp(-\bar{\psi}(1)).$$

The equation $\bar{\psi}(\xi) = S(U)$ was sold by the Newton method with a stabilization parameter $\delta = 1.0e{-}6$. The initial state ξ^0 was found as solution of the equation $\tilde{\psi}(t) = S(U), \quad \tilde{\psi}(t)) = B(\exp(\alpha Y_* t) - 1)/\alpha$ with $B > 0$, $\alpha > 0$.

A2: PC approximation of $\mu(x)$ on a uniform grid with m intervals. For the inversion of the PL function $\tilde{\psi}(t) = \int_0^t \tilde{\mu}(r)dr$, the equation $\psi_{i-1} + \mu_i(t - t_i) = S(U), \quad \psi_{i-1} < S(U) \le \psi_i$ was solved with respect to i, t.

A3: PL approximation of $\mu(x)$ on uniform grid. For the inversion of the piecewise quadratic function $\tilde{\psi}(t) = \int_0^t \tilde{\mu}(r)dr$ the equation

$$\psi_{i-1} + \tfrac{h}{2} w(\mu_{i-1}(2 - w) + \mu_i w), \quad \psi_{i-1} < S(U) \le \psi_i.$$

was solved with respect to i, w.

A4: PC approximation of the PDD $\breve{p}(x)$ on a uniform grid with m intervals. The DF $\hat{F}(x) = \int_0^x \hat{p}(r)dr$ is piecewise linear.

A5: PL approximation of the PDD $\breve{p}(x)$. The DF $\hat{F}(x)$ is piecewise quadratic. For the search in the tables, the index method was used with $mi = m\,div\,4$.

A6: PC approximation of the PDD $\breve{p}(x)$. The superposition method was used for the simulation of RV $\eta \sim \hat{p}(x) = \sum_{i=1}^m \Pi_i p_i(x)$. The DRV $j \sim \{\Pi_i\}$ was simulated with the Walker algorithm, the conditional RVs was simulated by the algorithm $\xi_j := h * (j - \tilde{U})$.

A7: PL approximation of the PDD $\breve{p}(x)$. The superposition method was used for he simulation of RV $\eta \sim \hat{p}(x) = \sum_{i=0}^m \Pi_i p_i(x)$. The DRV $j \sim \{\Pi_i\}$ was simulated with the Walker algorithm, the conditional RVs was simulated by the algorithm

$$\begin{aligned} \eta_1 &:= h|U_2 - \tilde{U}_3|; \quad \eta_m := 1 - h|U_2 - \tilde{U}_3|. \\ \eta_j &:= h(j + U_2 - \tilde{U}_3), \;\; j = 1, \ldots, m-1. \end{aligned}$$

The calculations were conducted on a computer with the tact frequency 1.68 Ggz. The PDD (6) was used with parameters $A = 6.0e-2$, $H = -4.0e-3$, $B = 8.0e-3$, $\alpha = 0.08$, $Y_* = 70$. Table 4 shows that the PC approximation of $\breve{p}(x)$ gives the

Table 4: Comparison of 7 simulation algorithms for the Makeham PDD.

Algorithm	1	2	3	4	5	6	7
t_{CPU}, sec	3.03e-5	2.58e-6	3.08e-6	9.80e-7	1.65e-6	6.00e-7	8.20e-7

fastest simulation algorithm, but it needs sufficiently more memory. For obtaining $L = 1.0e{-}4$, we need the PC and PL approximations with $m(\mathbf{A4}) = 1.6e4$, $m(\mathbf{A5}) = 2.1e2$ and for $L = 1.0e-6$ – with $m(\mathbf{A4}) = 1.6e6$, $m(\mathbf{A5}) = 2.1e4$.

5 Conclusion

In the paper, presented, author tried to show that the approximation of PDDes allows the researcher to create very fast algorithms for the simulation of RVs. This is due to the possibility to use long additional tables in the algorithms. In this paper, the one-dimensional case was considered in detail. However, the same approach is applicable to the two- and three-dimensional probability densities.

References

[1] de Boor C.(1977). Package for calculating with B-splines. *SIAM J. Numer. Anal.* Vol. **14**, pp. 441-472.

[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] Hastings W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Applications, *Biometrika*. Vol. **57**, pp. 97–109.

[5] Makhotkin O.A, Pirimkulov M.I. (1989), Using the splines in some statistical simulation problems, in *Theory and applications of statistical simulation*, ed. Michailov G.A., NCC Reports, Novosibirsk, pp. 43-53(in Russian).

[6] Scheffe H.(1947). A useful convergence theorem for probability distributions, *Annals of Math. Statistics*. Vol. **18**, pp. 434-458.

[7] Scollnik D.P.M. (1995). Simulation random variates from Makeham's distribution and from others with exact or nearly log–concave densities. *Transactions of Society of Actuaries*.Vol. **47**, pp. 409–454.

[8] Walker A.J.(1997). An efficient method for generating discrete random variables with general distributions, *ASME Transactions on Math. Software*. Vol. **3**, pp. 253-256.

On Survival Categorical Methods with Applications in Epidemiology and AIDS Research

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. Categorical tests for right censored survival data with applications to epidemiology and AIDS research are discussed. We introduce Wald's type homogeneity tests based on Kaplan–Meier and Nelson–Aalen estimators for categorical null hypothesis in right censored survival data case. Classical tests for contingency tables and the survival Wald's type tests are examined to assess their validity to use them for different kinds of statistical conclusions.

Keywords: *survival data, right censoring, independent censoring, hypothesis testing, categorical tests, contrasts, Wald's test, Nelson–Aalen estimator, Kaplan–Meier estimator, cohort analysis, AIDS research, genome wide association study (GWAS).*

1 Introduction

A common experimental design for gene association studies of disease progression is to screen a cohort of individuals for disease endpoints during some time interval. Study participants are disease free at baseline (time point zero). The goal is to quantify difference in rate of disease progression among a population of study participants .

Let T is a failure time random variable 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 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 $\gamma_T = \min_{i\in 1,\ldots,d} \sup\{x : F_i(x) < 1\}$. 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

$$H_0 : S_1(t_j) = \ldots = S_d(t_j) \quad \text{for all} \quad j = 1, \ldots, s.$$

It is clear that H_0 is closing to H_0^* if $p_{j|z} \to 0$ as $s \to \infty$.

Let (T_i, z_i), $i = 1, \ldots, n$, be the survival times and covariates of participants involved into research project. If all participants are followed up to t_s we may classify them by failure time into s groups, by z into d groups and construct contingency table. Classical methods of categorical data analysis like chi-square or likelihood ratio categorical tests are valid in this case. Unfortunately, failure time for missed at follow up before t_s participants is not specified. Classification observations of this kind to some group as well as just removing them from the analysis leads to some bias of the corresponding expected frequencies relatively to the true value of parameters in the contingency table.

The right censored survival data model is commonly used for such kind of experimental design. Let (T, U, z) be the failure time, censoring time and the covariate. The observed variables are (X, δ, z), where $X = T \wedge U$, $\delta = \mathbb{1}_{\{T \le U\}}$. Associate the corresponding variables (T_i, U_i, δ_i), $i = 1, \ldots, n$, to all participants. Then, the observed data are given by (X_i, δ_i, z_i), where $X_i = T_i \wedge U_i$ and $\delta_i = \mathbb{1}_{\{T_i \le U_i\}}$, $i = 1, \ldots, n$. The right censored survival data model will be used in sections 2 and 3. An extended model including unobserved random left truncation variable will be considered in section 4.

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 [5, 9, 10]. In section 2 we discuss several problems in application classical methods of categorical data analysis for right censored survival data.

Categorical tests using grouping for right censored survival data case are presented widely in literature. Likelihood ratio tests with grouped right censored survival data were investigated in [12]. A chi-square type test for survival data due to Habib & Thomas [4]. Advanced properties of chi-square type tests are obtained in [2, 3]. Hollander & Pena [7] consider chi-square test statistic for simple null hypotheses in censored data case and its limit behaviour. We consider contrasts based categorical tests on independence for survival data in section 3. The results are obtained immediately from limit theorems for Nelson–Aalen and Kaplan–Meier estimators.

A common problem in application of survival analysis in epidemiology is that the precise time of initiation for an individual into "at risk" group (time zero) is typically unknown and can't be estimated. One strategy involves using recruitment time as time zero. However, some of statistical conclusions on failure time may be incorrect or inaccurate in this case. In AIDS research the time of initiation for an individual into "at risk" group is the HIV infection time and it is observed for seroconverters (patients with an accurate date of HIV infection (the midpoint between a patients clinical visit that is HIV negative and the next visit which is HIV positive)). Seroprevalent participants (patients who enter the study already HIV positive) are also included into analysis and can be used to improve efficiency of the analysis. Problems of applications of survival methods in epidemiology and AIDS research will be discussed in section 4.

Clearly, the interval censored data models are more appropriate in the considered experimental design. Conversely, if the intervals of grouping are much larger

then times between visits, categorical type inference should not be sensitive on impreciseness of failure times, because the Kaplan–Meier estimator in a fixed point is dependent on order of failures and censorings observed before this time point, but not on the exact values.

2 Classical categorical tests and their applications to survival data analysis

The categorical experimental design based on classification of individuals to groups by one or several characteristics is very universal. Methods of the categorical data analysis are commonly using asymptotic normality of cell quantities as estimators of cell probabilities. It is important to take into account true cell probabilities for correct interpretation results of categorical analysis.

Consider cohort experimental design with failure time T as the response variable and categorical covariate z as the exposure variable. Assume for simplicity that two categories are specified for failure time: $T \le t_1$ and $T > t_1$. In AIDS research participants are classified to rapid and slow progressors by failure time is smaller of larger then the breakpoint t_1 respectively. Participants censored before the breakpoint can not be classified correctly because we can't say definitely whether individual of such kind is rapid or slow progressor. Different kinds of strategies for participants censored before the breakpoint t_1, including removing them from the analysis, lead to bias of the observed cell probabilities (expected frequences by the observed data) from target cell probabilities. Consider three possible strategies: all individuals censored before the breakpoint t_0 are removed from the analysis; all censored individuals including individuals without symptoms of disease at the endpoint are removed from the analysis and all individuals are classified by event time. Denote the expected cell probabilities $q_{i|z}^{(k)}$, $i = 1, 2$, $k = 1, 2, 3$.

The observed cell probabilities in this case are given by

$$q_{1|z}^{(1)} = P(T \le t_0 | \{T \le U\} \cup \{U > t_0\}; z) = P(T \le U \wedge t_0 | z) \,/\, \big(1 - P(U < T \wedge t_0 | z)\big);$$

$$q_{1|z}^{(2)} = P(T \le t_0 | T \le U; z) = P(T \le U \wedge t_0 | z) / P(T \le U; z); \quad q_{1|z}^{(3)} = P(T \wedge U \le t_0 | z)$$

and $q_{2|z}^{(k)} = 1 - q_{1|z}^{(k)}$. Assume that the distribution of censoring time U is not dependent of z and have a survival function $G(x) = P(U > x)$ with $G(0) = 0$. Then under independence of T and U,

$$q_{1|z}^{(1)} = \int_0^t G(x)\, dF_z(x) \,\Big/\, \Big(S_z(t_0)G(t_0) + \int_0^{t_0} G(x)\, dF_z(x)\Big);$$

$$q_{1|z}^{(2)} = \int_0^{t_0} G(x) dF_z(x) \,\Big/\, \int_0^{\infty} G(x) dF_z(x); \quad q_{1|z}^{(3)} = 1 - \overline{G}(t_0) S_z(t_0).$$

The corresponding biases are

$$q_{1|z}^{(1)} - p_{1|z} = S_z(t_0)\Big(G(t_0)(S_z(t_0) - 1) + \int_0^t G(x)\, dF_z(x)\Big) \,\Big/\, \Big(S_z(t_0)G(t_0) + \int_0^{t_0} G(x)\, dF_z(x)\Big) =$$

$$= S_z(t_0) \int_0^{t_0} (1 - S_z(x))\, d\overline{G}(x) \Big/ \Big(S_z(t_0)G(t_0) + \int_0^{t_0} G(x)\, dF_z(x) \Big) =$$

$$= S_z(t_0) \frac{P(T \le U < t_0)}{1 - P(U \le T \wedge t_0)},$$

where $p_{1|z} = F_z(t_0)$ and $\overline{G}(x) = 1 - G(x)$;

$$q^{(2)}_{1|z} - p_{1|z} = \int_0^{t_0} (G(x) - C) dF_z(x) \Big/ C, \quad \text{where} \quad C = \int_0^{\infty} G(x) dF_z(x)$$

and

$$q^{(3)}_{1|z} - p_{1|z} = S_z(t_0)G(t_0).$$

Remark that classical categorical tests are valid if to change true null hypothesis H_0 by

$$\widetilde{H}_0 : q_{1|1} = \ldots = q_{1|d}.$$

In most of cases the problem is to find difference between groups of individuals with different values of covariates. In this sense changing H_0 by $\widetilde{H}_0$ has no fundamental importance, but classical categorical approach is not applicable for interpretation results in terms of selected categories (rapid and slow progressors in AIDS research).

On the other hand, contamination of the estimated parameters by nuisance parameter caused by censoring may leads to missing efficiency of the test.

3 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 $1 \le k_2 < k_1 < \gamma_T$;

$$H_0 : S_1(t_l) = \ldots = S_d(t_l), \quad \text{for all} \quad l = 1, \ldots, s.$$

Denote, $\theta_{ij} = S_i(t_j)$. Introduce the parameters $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{is})'$. 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, H_0 can be rewritten in terms of contrasts

$$H_0 : \boldsymbol{\psi}_1 = \ldots = \boldsymbol{\psi}_{d-1} = 0.$$

Let $\hat{\theta}_{ij} = \hat{S}_i(t_j)$ be the estimator for θ_{ij}, where $\hat{S}_i$ is the Kaplan–Meier estimator for S_i, $i = 1, \ldots, d$. Using the convergence

$$\sqrt{n}(\hat{S}_i(t)/S_i(t) - 1) \Rightarrow W_{\sigma_i^2(t)},$$

where W_t is the standard Wiener process and $\sigma_i^2(t)$ is some positive nondecreasing function we obtain convergence

$$\sqrt{n_i}(\hat{\theta}_{i1} - \theta_{i1}, \ldots, \hat{\theta}_{is} - \theta_{is})' \Rightarrow N(0, \boldsymbol{\Sigma}_i),$$

and the asymptotic covariance matrix $\boldsymbol{\Sigma}_i = \|\sigma_{i:qr}\|$ is such that $\sigma_{i:qr} = \theta_{iq}\theta_{ir}\sigma_i^2(t_q \wedge t_r)$, $q, r = 1, \ldots, s$. We are using $\hat{\boldsymbol{\Sigma}}_i$ with $\hat{\sigma}_{i:qr} = \hat{\theta}_{iq}\hat{\theta}_{ir}\hat{\sigma}_i^2(t_q \wedge t_r)$, where $\hat{\sigma}_i^2(t)$ is a consistent estimator of $\sigma_i^2(t)$.

Introduce the random vector $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1s}, \ldots, \theta_{d1}, \ldots, \theta_{ds})'$ and the corresponding estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_{11}, \ldots, \hat{\theta}_{1s}, \ldots, \hat{\theta}_{d1}, \ldots, \hat{\theta}_{ds})'$. Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \Rightarrow N(0, \boldsymbol{\Sigma}), \tag{1}$$

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 $i = 1, \ldots, d$.

Associate with any a_{ij} diagonal matrix $\mathbf{A}_{ij} = a_{ij}\mathbf{I}_s$, where $\mathbf{I}_s$ is the identity matrix of size s and construct 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

$$H_0 : \mathbf{B}\boldsymbol{\theta} = 0.$$

Taking into account (1) 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 test can be based on Nelson–Aalen estimator of cumulative intensities $\Lambda_i(t)$ and $\Lambda_i(t) = -\log S_i(t)$ under continuous $S_i(t)$, $i = 1, \ldots, d$. Introduce parameters $\boldsymbol{\kappa}_i = (\kappa_{i1}, \ldots, \kappa_{is})'$, where $\kappa_{ij} = \Lambda_i(t_j)$. In terms of $\boldsymbol{\phi}_i = \mathbf{A}\boldsymbol{\kappa}_i$, $i = 1, \ldots, d$, where $\mathbf{A}$ is the contrasts matrix, the null hypothesis can be rewritten as

$$H_0 : \boldsymbol{\phi}_1 = \ldots = \boldsymbol{\phi}_{d-1} = 0.$$

Using martingale asymptotic result for the Nelson–Aalen estimators we obtain that

$$\sqrt{n_i}(\hat{\Lambda}_i(x) - \Lambda_i(x)) \Rightarrow W_{\tau_i^2(x)},$$

where $\tau_i^2(x)$ is the corresponding variance function, $i = 1, \ldots, d$. Then,

$$\sqrt{n_i}(\hat{\kappa}_{i1} - \kappa_{i1}, \ldots, \hat{\kappa}_{is} - \kappa_{is})' \Rightarrow N(0, \boldsymbol{\Upsilon}_i),$$

and the asymptotic covariance matrix $\boldsymbol{\Upsilon}_i = \|\tau_{i:qr}\|$ is such that $\tau_{i:qr} = \tau_i^2(t_q \wedge t_r)$, $q, r = 1, \ldots, s$. We are using $\widehat{\boldsymbol{\Sigma}}_i$ with $\hat{\tau}_{i:qr} = \hat{\tau}_i^2(t_q \wedge t_r)$, where $\hat{\tau}_i^2(x)$ be a consistent estimator of $\tau_i^2(x)$, $i = 1, \ldots, d$.

Introduce the random vector $\boldsymbol{\kappa} = (\kappa_{11}, \ldots, \kappa_{1s}, \ldots, \kappa_{d1}, \ldots, \kappa_{ds})'$ and the corresponding estimator $\hat{\boldsymbol{\kappa}} = (\hat{\kappa}_{11}, \ldots, \hat{\kappa}_{1s}, \ldots, \hat{\kappa}_{d1}, \ldots, \hat{\kappa}_{ds})'$. Then,

$$\sqrt{n}(\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}) \Rightarrow N(0, \boldsymbol{\Upsilon}), \tag{2}$$

and $\boldsymbol{\Upsilon} = \mathrm{diag}(l_1\boldsymbol{\Upsilon}_1, \ldots, l_n\boldsymbol{\Upsilon}_d)$ is the block-diagonal matrix, $l_i = n/n_i$ for $i = 1, \ldots, d$.

The null hypothesis can be written now in the following form:

$$H_0 : \mathbf{B}\boldsymbol{\kappa} = 0.$$

By (2), we obtain that under null hypothesis

$$n\,\hat{\boldsymbol{\kappa}}'\widehat{\mathbf{R}}^{-1}\hat{\boldsymbol{\kappa}} \Rightarrow \chi^2_{(d-1)s},$$

where $\widehat{\mathbf{R}} = \mathbf{B}'(\mathbf{B}\hat{\boldsymbol{\Upsilon}}\mathbf{B}')^{-1}\mathbf{B}$.

4 Applications in Epidemiology and AIDS research

General problem in application of survival methods in epidemiology is that the true failure time T is mostly unobserved and investigators are replace it by the recruitment time. Denote V is the time period from initiation of an individual into "at risk" group and recruitment time. It is reasonable to take into account two kinds of censoring: censoring associated with the individual and censoring associated with experimental conditions. We note them U_1 and U_2 respectively. We assume that (T, U_1, U_2) and V are independent. The response failure and censoring times (T^*, U^*) are identical in distribution to $(T - V, U_1 - V \wedge U_2)$ conditionally on $X_1 \geq V$, where $X_1 = T \wedge U_1$. Then, (X^*, δ^*, z) are observed for any participant, where $X^* = T^* \wedge U^*$, $\delta^* = \mathbb{1}_{\{T \leq U\}}$. Remark that

$$\mathbb{P}(T^* \in [t, t+\varepsilon) | X^* \geq t) = \mathbb{P}(T - V \in [x, x+dx) | T - V \geq x, U_1 - V \geq x, U_2 \geq x) =$$

$$= \int_0^\infty \mathbb{P}(T \in [x+v, x+v+dx) | T \geq x+v, U_1 \geq x+v, U_2 \geq x)\, dH(v).$$

Under

$$\mathbb{P}(T \in [y, y+\varepsilon) | T \geq y, U_1 \geq y, U_2 \geq y - v) = \mathbb{P}(T \in [y, y+\varepsilon) | T \geq y, U_1 \geq 0)$$

for $y > v > 0$ a.s. we obtain the independent censoring condition for the observed data

$$\mathbb{P}(T^* \in [t, t+\varepsilon) | X^* \geq t) = \mathbb{P}(T - V \in [t, t+dt) | T - V \geq t, X_1 \geq 0) = \mathbb{P}(T^* \in [t, t+\varepsilon) | T^* \geq t).$$

Then, methods of survival analysis are applicable to the observed data with the original theoretical distribution of failure time T changed by the conditional distribution of $T - V$ conditionally on $X_1 \geq V$. Under independence of T, U_1 and V we have the following survival function of T^*

$$S^*(x|z) = \mathbb{P}(T - V > x | T - V > 0) = \int_0^\infty S(x+v|z)\, dH(v) \Big/ \int_0^\infty S(v|z)\, dH(v).$$

The remarkable property of the exponential distribution is a constant hazard rate function. By this property left truncated exponential distribution coincides with the original one, i. e. $S^*(\cdot|z) \equiv S(\cdot|z)$. This property remains correct under independent random left truncation. Then, statistical conclusions obtained under the exponential distribution of survival time assumption can be applied correctly for T. On the other hand, for Weibull's parametric model, including exponential distributions as a particular cases, obtained statistical conclusions should be interpreted in terms of distribution of T^*. The exponential model is acceptable for HIV epidemic model, but for AIDS progression it is not acceptable at all.

Consider categorical design in AIDS progression study with binary classification of individuals by distribution of failure time to rapid and slow progressors. The contrasts framework discussed in section 3 as well as other methods of survival analysis are

valid for seroconverters. Seroprevalent participants can be used to increase power of homogeneity test of the original (non categorical) hypothesis H_0^*. Results obtained by using seroconverters only can be interpreted in terms of slow and rapid progressors, and results obtained by using seroconverters and seroprevalents can be interpreted in terms of differences between populations with different values of covariates. In spite of the distribution of V is not identifiable in general, comparative analysis of the estimated distributions of survival times in group of seroconverters and seroprevalents can be used to get partial information on the distribution of V in some special cases.

The methods and conclusions presented here have applications for performing more statistically robust analyses of genome wide association studies (GWAS) in AIDS research and in other complex diseases [1, 5, 8, 11]. They also bear on genome-wide statistical correction required for multiple SNPs and test hypotheses used in GWAS . Clearly the most optimal genetic epidemiological analytical methods are required to uncover the genetic determinants of disease heterogeneity in complex chronic and infectious diseases.

Acknowledgements

This research was supported in part by Russian Ministry of Education and Science Mega-grant no.11.G34.31.0068 (Dr. Stephen J. O'Brien, Principal Investigator).

References

[1] An, P., Winkler, C.A. (2010). Host genes associated with HIV/AIDS: advances in gene discovery. *Trends in Genetics* **26**(3), 119-131.

[2] 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.

[3] 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.

[4] Habib & Thomas (1986). Chi-Square Goodness-if-Fit Tests for Randomly Censored Data. *The Annals of Statistics* **14**(2), 759–765.

[5] 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* **5**(9), art. no. e12862, pp. 1-8

[6] Hjort, N.L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates. *The Annals of Statistics* **18**, 1221–1258.

[7] Hollander & Pena (1992). A Chi-Squared Goodness-of-Fit Test for Randomly Censored Data. *Journal of the American Statistical Association* **87**(418), 458-463.

[8] O'Brien, S.J., Hendrickson, S.L. (2013). Host genomic influences on HIV/AIDS. *Genome Biology* **14**, in Press.

[9] 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* **34**, 563–591.

[10] 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* **97**(26), 14467-14472.

[11] Troyer, J.L., Nelson, G.W., Lautenberger, J.A., Chinn, L, McIntosh, C., Johnson, R.C., Sezgin, E., Kessing, B., Malasky, M., Hendrickson, S.L., Li, G., Pontius, J., Tang, M., An,P., Winkler, C.A., Limou, S., Le Clerc, S., Delaneau, O., Zagury, J.-F., Schuitemaker, H., van Manen, D., Bream, J.H., Gomperts, E.D., Buchbinder, S., Goedert, J.J., Kirk, G.D. and O'Brien, S.J. (2011) Genome-Wide Association Study Implicates PARD3B-Based AIDS Restriction, *The Journal of Infectious Diseases* **203**, 1491502.

[12] Turnbull & Weiss (1978). A Likelihood Ratio Statistic for Testing Goodness of Fit with Randomly Censored Data. *Biometrics* **34**(3), 367–375.

Estimation of Unconditional Distributions from Data obeying Conditional Distributions

Shigeru Mase[1] and Yoshitaka Yamaguchi[2]
[1] *Tokyo Institute of Technology, Tokyo, Japan*
[2] *Tokyo Institute of Technology, Tokyo, Japan*
e-mail: mase@is.titech.ac.jp

Abstract

Inspired by the multiple decrement problem in actuarial mathematics, we consider how to estimate of unconditional distributions from data obeying conditional distributions. We show three such methods. The first one is nonparametric and according to a standard idea of actuarial mathematics. The second one is based on nonparametric density estimation. The last one, the main part of this paper, is parametric based on maximum likelihood estimations. Various examples are shown.

Keywords: multiple decrements, actuarial mathematics, estimation, conditional distribution, unconditional distribution.

Introduction

One of main topics in actuary mathematics is "multiple decrements". For example, life insurance contracts may be finished by two random causes, death of contractants or cancellation of contracts. Let X and Y be the time of cancellation and the time of death of contractants from a certain stating period respectively. Life insurance companies are interested in the probability distribution $\boldsymbol{P}\{X \leq x\}$, but X is observed only if $X \leq Y$. That is, what is actually observable is a cancellation data obeying conditional distribution $\boldsymbol{P}\{X \leq x \mid X \leq Y\}$. How can we estimate the unconditional probability distribution function $\boldsymbol{P}\{X \leq x\}$ from data obeying the conditional distribution function $\boldsymbol{P}\{X \leq x \mid X \leq Y\}$? Similar problems sometimes occur. For example, authors studied previously "life tables of marriages", that is, rates of Japanese married couples who will divorce eventually. In this problem, what can be observed are cases both of couple are alive. A widow might divorce if his or her partner would live longer.

In this paper, we consider three methods of estimating the unconditional distribution from data obeying the conditional distribution.

1 Basic formulations

Let X, Y be two mutually independent random variables with $\boldsymbol{P}\{X < Y\} > 0$. We have the following relation:

$$\boldsymbol{P}(X \leq x) = \boldsymbol{P}(X \leq x \mid X \leq Y) + \boldsymbol{P}(Y < X)[\boldsymbol{P}(X \leq x \mid Y < X) - \boldsymbol{P}(X \leq x \mid X \leq Y)].$$

Let $f(s)$ and $F(s)$ (resp. $g(s)$ and $G(s)$) be the density and the CDF of X (resp. Y). If we let $M = \int f(s)[1 - G(s)]ds$, then

$$\begin{aligned}
\boldsymbol{P}(X \le x \ \&\ X \le Y) &= \iint_{s\le x, s\le t} f(s)g(t)dsdt = \int_{-\infty}^{x} f(s)[1 - G(s)]ds, \\
\boldsymbol{P}(X \le Y) &= \int f(s)[1 - G(s)]ds = M, \\
\boldsymbol{P}(X \le x \ \&\ Y < X) &= \iint_{s\le x, t\le s} f(s)g(t)dsdt = \int_{-\infty}^{s} f(s)G(s)ds, \\
\boldsymbol{P}(Y < X) &= 1 - M.
\end{aligned}$$

Therefore the conditional CDF $H(x) = \boldsymbol{P}(X \le x \mid X \le Y)$ has the density function

$$h(s) = \frac{f(s)[1 - G(s)]}{M}. \tag{1}$$

In the following, we consider three methods of estimating the unconditional distribution F form data obeying the conditional distribution H. Another distribution G is assumed to be known in the first two methods. It may be unknown and is estimated together with F in the last method.

Remark 1. Although it is not always necessary, we assume $X, Y \ge 0$ in the sequel unless otherwise stated. F and G may be discrete distributions with appropriate modifications.
Also it is seen that we can recover f from h only if $G(s) < 1$ whenever $f(s) > 0$.

Remark 2. C. R. Rao already noticed (1) and called it "Calcutta blackout distribution" (probably a joke), see [5]. If X and Y be the time necessary to complete a laboratory experiment and the the time where electric power blackout happens from a certain starting time respectively, X can be observed only if $X < Y$.

Remark 3. We can also consider the conditional distribution $\boldsymbol{P}\{X \le x \mid X > Y\}$. It is a kind of truncated distributions and its density function is

$$h(s) = \frac{f(s)G(s)}{1 - M}.$$

2 Nonparametric method (1) – Piecewise linear interpolations

The idea of interpolating F and G piecewise linearly is standard in actuarial mathematics, see [3]. If $G(x)$ is known and we can approximate $F(x)$ and $G(x)$ linearly in

an interval $x \in (a, b]$:

$$F(s) \approx F(a) + \frac{F(x) - F(a)}{x - a}(s - a) \quad a < s < x \ (\leq b),$$
$$G(s) \approx G(a) + \frac{G(x) - G(a)}{x - a}(s - a) \quad a < s < x \ (\leq b),$$

then, from integrations in parts,

$$\begin{aligned}
\int_a^x sf(s)ds &= xF(x) - aF(a) - \int_a^x F(s)ds \\
&\approx xF(x) - aF(a) - \int_a^x \left[F(a) + \frac{F(x) - F(a)}{x - a}(s - x)\right]ds \\
&= \frac{x + a}{2}[F(x) - F(a)], \\
\int_a^x f(s)G(s)ds &= \int_a^x f(s)G(s)ds \\
&\approx \int_a^x f(s)\left[G(a) + \frac{G(x) - G(a)}{x - a}(s - x)\right]ds = \frac{G(x) + G(a)}{2}[F(x) - F(a)].
\end{aligned}$$

Therefore, for $a < x \leq b$,

$$\begin{aligned}
F(x) - F(a) - \boldsymbol{P}(a < X \leq x \mid X \leq Y) &= \int_a^x \left[\frac{M - 1}{M}f(s) + \frac{1}{M}f(s)G(s)\right]ds \\
&\approx \frac{M - 1}{M}[F(x) - F(a)] + \frac{1}{M}\frac{G(x) + G(a)}{2}[F(x) - F(a)] \\
&= \frac{2(M - 1) + G(x) + G(a)}{2M}[F(x) - F(a)]
\end{aligned}$$

and, finally, we can get an approximation

$$F(x) \approx F(a) + \frac{2M}{2 - G(x) - G(a)}\boldsymbol{P}(a < X \leq x \mid X \leq Y), \quad a < x \leq b.$$

Let $\hat{F}(a)$ be an estimate of $F(a)$ (of course, $\hat{F}(0) = 0$), $\hat{\boldsymbol{P}}(a < X \leq x \mid X \leq Y)$ and $\hat{M}$ be estimates of the conditional probability $\boldsymbol{P}(a < X \leq x \mid X \leq Y)$ and M respectively, we can estimate $F(x)$ for $x \in (a, b]$ by

$$\hat{F}(x) = \hat{F}(a) + \frac{2\hat{M}}{2 - G(x) - G(a)}\hat{\boldsymbol{P}}(a < X \leq x \mid X \leq Y), \quad a < x \leq b.$$

In practice, let $x_0 = 0 < x_1 < x_2 < \cdots$ be a fine division of $[0, \infty)$ and let

$$\hat{F}(x) = \hat{F}(x_{n-1}) + \frac{2\hat{M}}{2 - G(x) - G(x_{n-1})}\hat{\boldsymbol{P}}(x_{n-1} < X \leq x \mid X \leq Y) \quad x_{n-1} < x \leq x_n.$$

If $a = \min\{x; G(x) = 1\} < \infty$, the preceding recurrence relation can be applied till the interval with $x_{n-1} < a \leq x_n$. $\hat{F}(x)$ is monotone increasing if $\hat{\boldsymbol{P}}(X \leq x \mid X \leq Y)$ is monotone increasing. If $M' = \lim_{x \uparrow \infty} \hat{F}(x) \neq 1$, then we should replace $\hat{M}$ by $\hat{M}/M'$.

3 Nonparametric method (2) – Density estimation

The relation (1) can be rewritten as

$$f(s) = M\frac{h(s)}{1-G(s)} \propto \frac{h(s)}{1-G(s)}. \tag{2}$$

Therefore, if we have an estimate $\hat{h}(s)$ of the density function h from data and if G is known, we can get the estimate $\hat{f}(s)$ of f by

$$\hat{f}(s) = C\frac{\hat{h}(s)}{1-G(s)}.$$

The normalizing constant C can be determined by the condition

$$C\int \frac{\hat{h}(s)}{1-G(s)}ds = 1.$$

As to nonparametric density estimation, kernel-type density estimators are well-known, see [6].

4 Parametric method – Maximum likelihood estimation

Consider parametric families $f_\theta(s)$ and $g_\xi(s)$. The conditional density is

$$h(s) = h_{\theta,\xi}(s) = \frac{f_\theta(s)[1-G_\xi(s)]}{M(\theta,\xi)},$$

where

$$G(s) = G_\xi(s) = \int_{-\infty}^{s} g_\xi(t)dt, \quad M = M(\theta,\xi) = \int f_\theta(s)[1-G_\xi(s)]ds.$$

Its log-likelihood is

$$LL(\theta,\xi) = \sum_{i=1}^{n}\log h_{\theta,\xi}(X_i) = \sum_{i=1}^{n}\log f_\theta(X_i) + \sum_{i=1}^{n}\log[1-G_\xi(X_i)] - n\log M(\theta,\xi)$$

and its log-likelihood equations are

$$\begin{aligned}
\frac{\partial LL}{\partial\theta_j} &= \sum_{i=1}^{n}\frac{1}{f_\theta(X_i)}\frac{\partial f_\theta(X_i)}{\partial\theta_j} - \frac{n}{M(\theta,\xi)}\frac{\partial M(\theta,\xi)}{\partial\theta_j} \\
&= \sum_{i=1}^{n}\frac{1}{f_\theta(X_i)}\frac{\partial f_\theta(X_i)}{\partial\theta_j} - \frac{n}{M(\theta,\xi)}\int\frac{\partial f_\theta(s)}{\partial\theta_j}[1-G_\xi(s)]ds = 0, \\
\frac{\partial LL}{\partial\xi_k} &= -\sum_{i=1}^{n}\frac{1}{1-G_\xi(X_i)}\frac{\partial G_\xi(X_i)}{\partial\xi_k} - \frac{n}{M(\theta,\xi)}\frac{\partial M(\theta,\xi)}{\partial\xi_k} \\
&= -\sum_{i=1}^{n}\frac{1}{1-G_\xi(X_i)}\frac{\partial G_\xi(X_i)}{\partial\xi_k} + \frac{n}{M(\theta,\xi)}\int f_\theta(s)\frac{\partial G_\xi(s)}{\partial\xi_k}ds = 0.
\end{aligned} \tag{3}$$

ξ is fixed if g is known. In order to solve log-likelihood equations (3), it is necessary to be able to compute functions M and their derivatives. Usually functions M are fairly complex, but can be sometimes expressed in terms of higher transcendental functions. In the sequel, we list seven such examples among others. As to higher transcendental functions and their basic properties, see, e.g., [1, 2]. For various univariate distributions, we refer to [4].

In the following, *const* stands for a term which is independent of relevant parameters. Notations such as $\overline{r(X)}$ mean $1/n \cdot \sum_{i=1}^{n} r(X_i)$.

(1) Exponential and Exponential distributions. Let $f_\theta(x) = \theta e^{-\theta x}$ and $g_\xi(x) = \xi e^{-\xi x}$ $(x > 0)$. Then $G_\xi(x) = 1 - e^{-\xi x}$, $LL(\theta, \xi) = n \log(\theta + \xi) - n(\theta + \xi)\bar{X}$ and, therefore, we can only estimate $\theta + \xi$. On the other hand, if ξ is known,

$$\frac{1}{n}\frac{\partial LL(\theta, \xi)}{\partial \theta} = \frac{1}{\theta + \xi} - \overline{X} = 0$$

and the MLE of θ is $\hat{\theta} = 1/\bar{X} - \xi$.

(2) Gamma and Exponential distributions. Fix $k > 0$.

$$f_\alpha(s) = \frac{\alpha^k}{\Gamma(k)} s^{k-1} e^{-\alpha s},\ g_\lambda(s) = \lambda e^{-\lambda s}, \quad M = (\alpha/(\alpha + \lambda))^k.$$

Hence

$$\frac{1}{n} LL(\alpha, \lambda) = const + k \log(\alpha + \lambda) - (\alpha + \lambda)\overline{X},$$
$$\frac{1}{n}\frac{\partial LL}{\partial \alpha} = \frac{1}{n}\frac{\partial LL}{\partial \lambda} = \frac{k}{\alpha + \lambda} - \overline{X} = 0.$$

It is seen that we can estimate only $\alpha + \lambda$.

(3) Exponential and Gamma distributions. Fix $k > 0$. Let $\gamma(a, x)$ is the incomplete gamma function of the first kind.

$$f_\lambda(s) = \lambda e^{-\lambda s}, \quad g_\beta(s) = \frac{\beta^k}{\Gamma(k)} s^{k-1} e^{-\beta s},$$
$$G_\beta(s) = \frac{\gamma(k, \beta s)}{\Gamma(k)}, \quad M(\beta, \lambda) = 1 - (1 + \lambda/\beta)^{-k}.$$

Hence

$$\begin{aligned}
\frac{1}{n}LL(\beta,\lambda) &= const + k\log(\beta+\lambda) + \log\lambda - \lambda\overline{X} - \log((\beta+\lambda)^k - \beta^k) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\log(\Gamma(k) - \gamma(k,\beta X_i)), \\
\frac{1}{n}\frac{\partial LL}{\partial\lambda} &= \frac{k}{\beta+\lambda} + \lambda^{-1} - \overline{X} - \frac{k(\beta+\lambda)^{k-1}}{(\beta+\lambda)^k - \beta^k} = 0, \\
\frac{1}{n}\frac{\partial LL}{\partial\beta} &= \frac{k}{\beta+\lambda} - k\frac{(\beta+\lambda)^{k-1} - \beta^{k-1}}{(\beta+\lambda)^k - \beta^k} - \frac{1}{n}\sum_{i=1}^{n}\frac{X_i(\beta X_i)^{k-1}e^{-\beta X_i}}{\Gamma(k) - \gamma(k,\beta X_i)} = 0.
\end{aligned}$$

(4) Gomperz and Exponential distributions. Let $\Gamma(z,p)$ be the incomplete gamma functions of the second kind.

$$f_{\alpha,\beta}(s) = \beta\alpha\cdot\exp(\beta s - \alpha(e^{\beta s}-1)), \quad g_\lambda(s) = \lambda e^{-\lambda s}, \quad M = \alpha^{\lambda/\beta}e^{\alpha}\Gamma(1-\lambda/\beta,\alpha).$$

Hence

$$\begin{aligned}
\frac{1}{n}LL &= (\beta-\lambda)\overline{X} - \frac{\alpha}{n}\sum_{i=1}^{n}e^{\beta X_i} - \log\Gamma(1-\lambda/\beta,\alpha) + (\log\alpha + \log\beta - (\lambda/\beta)\log\alpha), \\
\frac{1}{n}\frac{\partial LL}{\partial\alpha} &= -\overline{e^{\beta X}} + \frac{\alpha^{-\lambda/\beta}e^{-\alpha}}{\Gamma(1-\lambda/\beta,\alpha)} + (1-\lambda/\beta)/\alpha = 0, \\
\frac{1}{n}\frac{\partial LL}{\partial\beta} &= \overline{X} - \alpha\overline{Xe^{\beta X}} + \beta^{-1} - \frac{\lambda\int_\alpha^\infty t^{-\lambda/\beta}e^{-t}\log t\ dt}{\beta^2\Gamma(1-\lambda/\beta,\alpha)} = 0, \\
\frac{1}{n}\frac{\partial LL}{\partial\lambda} &= -\overline{X} - \frac{\log\alpha}{\beta} + \frac{\int_\alpha^\infty t^{-\frac{\lambda}{\beta}}e^{-t}\log t\ dt}{\beta\Gamma(1-\lambda/\beta,\alpha)} = 0.
\end{aligned}$$

(5) Beta and Exponential distributions. Let $B(x,y)$, ${}_1F_1(\alpha,\gamma;x)$ and $\psi(x)$ be the beta, Kummer's confluent hyper-geometric and the digamma function respectively.

$$\begin{aligned}
f_{\alpha,\beta}(x) &= \frac{1}{bB(\alpha,\beta)}\left(\frac{x}{b}\right)^{\alpha-1}\left(1-\frac{x}{b}\right)^{\beta-1} \quad 0\le x\le b \\
g_\xi(x) &= \xi e^{-\xi x}, \quad M = {}_1F_1(\alpha,\alpha+\beta;-b\xi).
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{1}{n}LL &= const - \xi\bar{X} + \alpha\overline{\log X} + \beta\overline{\log(1-X/b)} - \log B(\alpha,\beta) \\
&\quad - \log {}_1F_1(\alpha, \alpha+\beta; -b\xi), \\
\frac{1}{n}\frac{\partial LL(\theta,\xi)}{\partial\alpha} &= \bar{X} - \psi(\alpha) + \psi(\alpha+\beta) - (\psi(\alpha+\beta) - \psi(\alpha)) \times {}_1F_1(\alpha, \alpha+\beta; -b\xi) \\
&\quad - (\alpha-1){}_1F_1(\alpha-1, \alpha+\beta-1; -b\xi) = 0, \\
\frac{1}{n}\frac{\partial LL(\theta,\xi)}{\partial\beta} &= \bar{Y} - \psi(\beta) + \psi(\alpha+\beta) - (\psi(\alpha+\beta) - \psi(\beta)) \times {}_1F_1(\alpha, \alpha+\beta; -b\xi) \\
&\quad - (\beta-1) \times {}_1F_1(\alpha-1, \alpha+\beta-1; -b\xi) = 0, \\
\frac{1}{n}\frac{\partial LL(\theta,\xi)}{\partial\xi} &= -\bar{X} + \frac{b\alpha}{\alpha+\beta} \times {}_1F_1(\alpha+1, \alpha+\beta+1; -b\xi) = 0.
\end{aligned}$$

(6) Pareto and Exponential distributions Fix $b > 0$. Let $W_{k,\mu}(z)$ be Whittaker function.

$$f_a(s) = \frac{a/b}{(s/b)^{a+1}}, \quad g_\lambda(s) = \lambda e^{-\lambda x}, \quad M = a(b\lambda)^a(\lambda b)^{-(a+1)/2}e^{-(\lambda b)/2}W_{-(a+1)/2,-a/2}(\lambda b).$$

Hence

$$\begin{aligned}
\frac{1}{n}LL &= const + \frac{(1-a)\log\lambda}{2} + \frac{a\log b}{2} - \log W_{-(a+1)/2,-a/2}(\lambda b) - \lambda\overline{X} + \frac{b\lambda}{2} - a\overline{\log X}, \\
\frac{1}{n}\frac{\partial LL}{\partial a} &= -\frac{\log\lambda}{2} + \frac{\log b}{2} - \overline{\log X} + \frac{e^{\lambda b/2}(\lambda b)^{(a+1)/2}\int_{\lambda b}^{\infty}\frac{e^{-t}}{t^{a+1}}\log t\ dt}{W_{-(a+1)/2,-a/2}(\lambda b)} = 0, \\
\frac{1}{n}\frac{\partial LL}{\partial\lambda} &= \frac{(1-a)}{2\lambda} - \overline{X} + \frac{b}{2} + \frac{\lambda^{-(a+1)/2}b^{-(a-1)/2}e^{-\lambda b/2}}{W_{-(a+1)/2,-a/2}(\lambda b)} = 0.
\end{aligned}$$

(7) Exponential and Normal distributions. Let $\mathrm{erf}(z)$ and $\mathrm{erfc}(z)$ be the error function and the complementary error function respectively.

$$\begin{aligned}
&f_\lambda(s) = \lambda e^{-\lambda s}, \quad g_{\sigma^2}(s) = (2\pi\sigma^2)^{-1/2}e^{-s^2/(2\sigma^2)}, \ -\infty < s < \infty, \\
&G(s) = \frac{1}{2}\left(1 + \mathrm{erf}(s/\sqrt{2\sigma^2})\right), \quad M = \frac{1}{2}\left[1 - e^{\lambda^2\sigma^2/2}\mathrm{erfc}(\lambda\sqrt{2\sigma^2}/2)\right].
\end{aligned}$$

Hence

$$\frac{1}{n}LL = \log\lambda - \log\left(1 - e^{\lambda^2\sigma^2/2}\mathrm{erfc}(\lambda\sqrt{2\sigma^2}/2)\right) - \lambda\overline{X} + \frac{1}{n}\sum_{i=1}^{n}\log\left(1 - \mathrm{erf}(X_i/\sqrt{2\sigma^2})\right),$$

$$\frac{1}{n}\frac{\partial LL}{\partial\lambda} = \lambda^{-1} - \frac{1}{1 - 2e^{\lambda^2\sigma^2/2}(1 - \Phi(\lambda\sigma)}\left[(2\sigma^2/\pi)^{-1/2} - \lambda\sigma^2 e^{\lambda^2\sigma^2/2}\mathrm{erfc}(\lambda\sqrt{2\sigma^2}/2)\right] - \overline{X}$$

$$= 0,$$

$$\frac{1}{n}\frac{\partial LL}{\partial\sigma^2} = \frac{1}{1 - 2e^{\lambda^2\sigma^2/2}(1 - \Phi(\lambda\sigma))}\left[\lambda^2 e^{\lambda^2\sigma^2/2}(1 - \Phi(\lambda\sigma)) - \lambda/\sqrt{2\pi\sigma^2}\right]$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\left[\frac{X_i e^{-X_i^2/(2\sigma^2)}}{\sqrt{\pi}(2\sigma^2)^{3/2}(1 - \Phi(X_i/\sigma))}\right] = 0$$

Conclusions

There are many data which obey some conditional distributions. Although we consider one particular conditioning and and a few examples, the used ideas will be useful to discuss such data.

References

[1] Abramowitz M., Stegun I.A. (1965). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Courier Dover Publications.

[2] Frank W.J. et al. (2010). *HIST Handbook of Mathematical Functions*. NIST, Cambridge University Press, Cambridge.

[3] Gerber H.U. (1997). *Life Insurance Mathematics*, 3rd ed. Springer, Berlin Heidelberg.

[4] Johnson N.L., Kemp A.W., Kotz S. (2005). *Univariate discrete distributions*. Wiley Interscience, New York.

[5] Rao C.R. (1997). *Statistics and Truth: Putting Chance to Work*, 2nd ed. World Scientific Publishing Co., 1997.

[6] Silverman B.W. (1986). *Density Estimation*. Chapman and Hall, London.

Remarks about Algorithms of Statistical Simulation

Gennady A. Mikhailov
Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
prospect Akademika Lavrentjeva, 6, Novosibirsk,
630090, Russia
e-mail: gam@osmf.sscc.ru

Abstract

A lot of numerical algorithms to simulation of random variables and vectors and besides to functional numerically-statistical analysis, in which elaboration the author has took the part, are considered. Practically important specifications and explanations of the algorithms formulation and substantiations are given. ***Keywords:*** base random number, probability density function, discrete superposition method, Gaussian vector, generalized exponential distribution , branching of trajectories, histogram.

Introduction

Now there are a lot of computing programs complexes for solving the applied problems by the Monte Carlo method. But sometimes new actual problems appear so that it is expedient to obtain rapidly the first numerical results by statistical simulation. For this purpose there are convenient, may be computationally expensive, but technically simple and easy programing algorithms. In the paper a lot of those algorithms are presented with practically important remarks related to their substantiation, specification and explanation.

1 Simulation of uniformly distributed pairs of numbers

Let

$$\xi, \eta \in \overline{1,n}, \;\; \xi \neq \eta.$$

Those equally probable pairs of numbers are simulated, for example, when solving non-linear kinetic equations by the Monte Carlo method (see, for example, [1]).

The next algorithm is known:
$\alpha_1 := rand; \;\; \alpha_2 := rand;$
$\xi := entier(\alpha_1 \times n) + 1; \quad \eta := entier(\alpha_2 \times (n-1)) + 1;$
$if \quad \eta \geq \xi \;\; then \quad \eta := \eta + 1.$

The substantiation of this algorithm followes from the relation

$$P(\xi = i, \eta = j) = P(\xi = i)P(\eta = j|\xi = i) =$$
$$= \frac{1}{n-1}P(\xi = i), \quad j = 1, ..., i-1, i+1, ..., n,$$

which is related to the fact that the conditional distribution on any fixed subset of considering pairs is so uniform. It is interesting to remark, that the change: $\alpha_2 = n\alpha_1 - \xi + 1$ is admissible here (see further sect. 5).

Here and further symbol *rand* denotes the procedure of generating the next, i. e. every time new, pseudo-random number. Symbol α with different indexes denotes independent realizations of the based random number uniformly distributed in (0,1).

The additional remark: the random numbers

$$entier[n(n^{k-1}\alpha - entier(n^{k-1}\alpha))] + 1, \;\; k = 1, 2, 3, ...,$$

are independent and uniformly distributed in $\overline{1, n}$.

2 Simulation of "gamma" and "beta"-distributions

2.1. Here we consider the distribution with the density

$$f_{\nu,\lambda}(x) = \frac{x^{\nu-1}e^{-\lambda x}}{\Gamma(\nu)}, \quad x > 0, \quad \lambda > 0$$

for $0 < \nu < 1$. It is known that the case of arbitrary ν is reduced to considered above on the base of the "gamma"-distributions composition rule.

The representation holds:

$$\xi_{\nu,\lambda} = \frac{\xi_\nu}{\lambda},$$

where ξ_ν is distributed with the density

$$f_\nu(x) = \frac{x^{\nu-1}e^{-x}}{\Gamma(\nu)}, \quad x > 0,$$

because $P(\frac{\xi_\nu}{\lambda} < x) = F_\nu(\lambda x) = F_{\nu,\lambda}(x)$. Simulation of ξ_ν is realized by "majorant rejection method" [2] based on the relation

$$g(x) = x^{\nu-1}e^{-x} \le g_1(x) = \begin{cases} x^{\nu-1}, & x \le 1, \\ e^{-x}, & x > 1, \end{cases}$$

Here the random variable ξ_1 with the density $(\nu^{-1} + e^{-1})^{-1}g_1(x)$ is simulated by the "inverse function method", i. e. by solving the equation

$$\frac{1}{\nu^{-1} + e^{-1}} \int_0^{\xi_1} g_1(x)\, dx = \alpha.$$

If $\alpha < \frac{\nu^{-1}}{\nu^{-1}+e^{-1}} = \frac{1}{1+\nu e^{-1}}$, *then*

$$\frac{1}{\nu^{-1}+e^{-1}} \int_0^{\xi_1} x^{\nu-1}\, dx = \alpha, \quad \xi_1 = [\alpha(1+\nu e^{-1})]^{1/\nu}, \tag{3.1}$$

else

$$\frac{1}{\nu^{-1}+e^{-1}} \int_1^{\xi_1} e^{-x}\, dx = \alpha - \frac{\nu^{-1}}{\nu^{-1}+e^{-1}}, \quad \xi_1 = -\ln[(e^{-1}+\nu^{-1})(1-\alpha)]. \tag{3.2}$$

In the rejection algorithm, which is realized by verification of the inequality $\alpha_1 g_1(\xi_1) < g(\xi_1)$, it is possible to decrease the computational cost by cancelling one of factors in the expression of $g(\xi_1)$, i. e. in the case (3.1) $\xi = \xi_1$, if $\alpha_1 < e^{-\xi_1}$, and in the case (3.2) $\xi = \xi_1$, if $\alpha_1 < \xi_1^{\nu-1}$.

It is known that the average number of cycles in this rejection algorithm is proportional to $S(\nu) = (\nu^{-1}+e^{-1})/\Gamma(\nu)$, while $S(0) = 1$, $S(1) = 1 + e^{-1}$ and $S'_\nu(\nu) > 0$, $0 < \nu < 1$.

Remark, that simulation of "gamma"-distribution is widely used when solving stochastic problems of meteorology and financian mathematics.

2.2. The density of "beta"-distribution is determined by the formula:

$$f_{p,m}(x) = \frac{x^{p-1}(1-x)^{m-1}}{B(p,m)}, \quad 0 < x < 1,$$

where $p, m > 0$ are parameters.

For the case of integer m the following simulating formula is known [2]:

$$\xi_{p,m} = \exp\left(\sum_{k=1}^{m} \frac{\ln \alpha_k}{p+k-1}\right). \tag{3.3}$$

The derivation of this formula given in textbooks is nonstandard and technically complex. That derivation is difficult and in this connection formula (3.3) is used wrongly seldom. But it is possible to verify this formula by transition to the random variable - $\ln \xi_{p,m}$. The corresponding density function is obtained by induction over m with using the convolution formula:

$$C_0 \int_0^x e^{-zp}(1-e^{-z})^{m-1} e^{-(p+m)(x-z)} dz =$$

$$= C_0 e^{-(p+m)x} \int_0^x e^z (e^z-1)^{m-1} dz =$$

$$= Ce^{-(p+m)x}(e^x-1)^m = Ce^{-px}(1-e^{-x})^m.$$

For noninteger p, m it is possible to use the rejection method using as majorant for $x^{p-1}(1-x)^{m-1}$ one of the functions:

$$x^{[p]-1}(1-x)^{m-1}, \quad x^{p-1}(1-x)^{[m]-1},$$

exept of the case: $p,\ m < 1$, for which the superposition algorithm is premilinarilly used (see section 5) on the base of the representation:

$$f_{p,m}(x) = \frac{p}{p+m} f_{p+1,m}(x) + \frac{m}{p+m} f_{p,m+1}(x).$$

3 The "majorant (maximal) cross-section method" for simulating the generalized exponential distribution

The Poisson random point flux $\{\tau_i\}\ i = 0, 1, ...$ with intensity $\sigma(t)$ is characterized by the fact that random variables $\tau_i - \tau_{i-1} (i = 1, 2, ...)$ are distributed accordingly with the densities

$$f_i(t|\tau_{i-1}) = \sigma(t + \tau_{i-1}) \exp(-T(t; \tau_{i-1})), \quad t > 0,$$

where $T(t; \tau_{i-1}) = \int\limits_0^t \sigma(t' + \tau_{i-1})\ dt';\ \ \tau_0 = 0.$

It is supposed that $T(\infty; \tau_{i-1}) = +\infty$. The distribution of the random variable $\xi = \tau_1$ with the density

$$f(t) \equiv f_1(t|0) = \sigma(t) \exp\left(- \int\limits_0^t \sigma(t')\ dt' \right), \quad t > 0,$$

is named generalized (or nonhomogeneous) exponential distribution.

Let $\sigma(t) \le \sigma_m(t)$, while Poisson flux $\{\tau_i^{(m)}\}$ with intensity $\sigma_m(t)$ is enough simple simulated. The majorant cross-section method followes directly from the rejection property of the flux $\{\tau_i^{(m)}\}$ (see, for example, [3]):
if the points $\tau_i^{(m)} (i = 1, 2, ...)$ are conditionally independly rejected with the probabilities $1 - \sigma(\tau_i)/\sigma^{(m)}(\tau_i)$, i. e. are selected with probabilities $p(\tau_i) = \sigma(\tau_i)/\sigma^{(m)}(\tau_i)$, then the selected points $\{\tau_j\}$ realize the point flux with intensity $\sigma(t)$.

On the base of above-stated it is possible to formulate the algorithm of majorant cross-section:
the flux $\{\tau_i^{(m)}\}$ is realized, while $\nu = \min\{i : \alpha_i < \sigma(\tau_i^{(m)})/\sigma_m(\tau_i^{(m)})\}$ is determined; in result $\xi = \tau_\nu^{(m)}$.

Corresponding to the variant $\sigma_m(t) \equiv \sigma_m$ the Coleman algorithm of "maximal cross-section" can be formulated in especially simple way:

$$\xi := 0;\ \ L : \xi := \xi - \ln(rand)/\sigma_m;\ \ if\ rand > \sigma(\xi)/\sigma_m(\xi)\ \ then\ \ go\ \ to\ \ L.$$

Here $\mathrm{E}(\nu) < +\infty$, in particular, if $\sigma(t) \geq \varepsilon > 0$. More real criteria for unequality $\mathrm{E}(\nu) < +\infty$ is obtained by considering the probability of "selection after one point" of the flux $\tau_i^{(m)}$ [3].

The majorant cross-section method is widely used while solving kinetic equations for geometrically complex radiation models. It is known also the randomized (with respect to items of the cross-section) variant of this method under the name "method of majorant frequency" (see, for example, [1]).

4 The modified method of discrete superposition

4.1. Let $\{p_i\}$ are probabilities, $\{f_i(x)\}$ are probability densities, $\{\psi_i(\alpha)\}$ are corresponding simulating functions and

$$f(x) = \sum_i p_i f_i(x), \quad i = 1, 2, ...$$

According to the discrete superposition method the random variable ξ with the probability density $f(x)$ is simulated so:

$$\textit{if} \quad \alpha_1 \in \Delta_k = \left[\sum_{i=1}^{k-1} p_i, \sum_{i=1}^{k} p_i \right) \quad \textit{then} \quad \xi = \psi_k(\alpha_2).$$

The value α_2 *can be here changed on* $\left(\alpha_1 - \sum\limits_{i=1}^{k-1} p_i\right)/p_k$, *since under the condition* $\alpha_1 \in \Delta_k$ *the value* $\alpha_1 - \sum\limits_{i=1}^{k-1} p_i$ *is uniformly distributed in the interval* $[0, p_k)$.

So the modified algorithm is obtained.

4.2. Further the practically important example of using the modified algorithm is considered.

Let $f(x) \equiv f(x; s)$, $\xi \equiv \xi_s$, where s - is a parameter and

$$f(x; s) = \frac{s - s_1}{s_2 - s_1} f(x; s_2) + \frac{s_2 - s}{s_2 - s_1} f(x; s_1), \quad s_1 \leq s \leq s_2,$$

i. e. $f(x; s)$ in the interval $s_1 \leq s \leq s_2$ is determined by the linear interpolation with respect to the parameter. Let $(s - s_1)/(s_2 - s_1) = p_2$, $f(x; s_1) = f_1(x)$, $f(x; s_2) = f_2(x)$, then the corresponding modified algorithm of discrete superposition is so:

$$\textit{if} \quad \alpha < p_1 \quad \textit{then} \quad \xi := \psi_1(\alpha/p_1) \quad \textit{else} \quad \xi := \psi_2\big((\alpha - p_1)/p_2\big).$$

It is expedient to use that algorithm for simulation of a particle scattering if the indicatrix linearly depends on the parameter. If $0 \leq \xi_s \leq s$, then here it is expedient to simulate the random variable ξ_s/s.

5 Simulating the Gaussian random vector

The Gaussian vector $\eta = (\eta_i, ..., \eta_l)^T$ with zero average ($\mathrm{E}\eta = 0$) and the correlation matrix $K = \{K_{ij}\}$ $i, j = 1, ..., l$ is simulated by the formula: $\eta = A\xi$, where $\xi = (\xi_1, ..., \xi_l)^T$ is the vector with independent standard normal components and $A = \{a_{ij}\}$ $(1 \le j \le i \le l)$ is the lower-triangular matrix . Let matrix K is positive and its principal minors are positive respectively. Since $\mathrm{E}\xi\xi^T = diag(1, ..., 1)$, then the relation holds:

$$\mathrm{E}A\xi(A\xi)^T = \mathrm{E}A\xi\xi^T A^T = A\mathrm{E}\xi\xi^T A^T = AA^T = R.$$

From the last equality it is easy to obtain the known (see, for example, [2]) recurrent representation of the values a_{ij}:

$$a_{jj} = \sqrt{R_{jj} - \sum_{k=1}^{j-1} a_{jk}^2}, \quad a_{ij} = \frac{R_{ij} - \sum\limits_{k=1}^{j-1} a_{ik}a_{jk}}{a_{jj}}.$$

Entier the square root here is the j-th principal minor of matrix K. If the values K_{ij} are approximated statistically then the principal minors may be nonpositive and matrix K must be modified. The simplest modification is obtained by the change: $K_{ij} \to K_{ij} + \varepsilon$. More exact modification relates to reducing to zero the nonpositive eigenvalues of matrix K by using rotation of the coordinat sistem.

Now let us consider the Wiener formulae:

$$\xi_1 = \sqrt{-2\ln\alpha_1}\cos(2\pi\alpha_2), \quad \xi_2 = \sqrt{-2\ln\alpha_2}\sin(2\pi\alpha_2). \tag{6.1}$$

The random variables ξ_1, ξ_2 are independent and standard normal. But, when using pseudo-random numbers , here the two-dimensional distribution of the point (α_1, α_2) must be sufficiently uniform in the unit square.

The vector determined by the formulae (6.1) is isotropic on the plane. It is known the more general statement:
the random vector $(\xi_1, \ldots, \xi_n)$ *with independent equally normal distributed components is isotropic in* R^n.

On the other hand the following expression of the l-th component of the n-dimensional isotropic vector ω_n is obtained in [4]:

$$F_l(t) = P(x_l < t) = \frac{S_{l-1}}{S_l} \int\limits_{-\frac{\pi}{2}}^{arc\sin t} \cos^{l-2}(\Theta)\, d\Theta,$$

where S_l is the surface area of the unit sphere in R^l. The corresponding algorithm is especially useful for realization only a subset of vector ω_n components.

6 Algorithms with trajectories branching and the problem of minimizing the variance of an integer random variable with given averge

The effective coefficient k_{ef} of particle multiplication over generations of branching trajectories can be estimated by direct simulation which gives numbers $\{n_i\}$ of particles appearing in the corresponding generations. In result the following statistical estimate is constructed:

$$k_{ef} \approx \tilde{k} = \frac{n_1 + n_2 + \ldots + n_m}{n_2 + n_3 + \ldots + n_{m+1}} = \frac{L(2,m)}{L(1,m-1)}.$$

To improve the estimate some number N of initial generations is neglected, i. e. the change $n_N \to n_1$ is performed.

It is known the following asymptotically exact (while $N, n_1 \to \infty$) estimate of the variance (see, example, [5]):

$$D\tilde{k} \approx \frac{k_{ef}(1 - \frac{k_{ef}}{\tilde{\mathrm{E}}\nu}) + \frac{k_{ef}}{\mathrm{E}\nu} D\nu}{L(1,m-1)}, \tag{7.1}$$

where ν is the random number of particles per one branching. Remark that previously the item with $D\nu$ was neglected, possibly, because the models with the fixed ν were usually considered. The quantity k_{ef} according to given particle transfer model is totally determined by the values $q = \mathrm{E}\nu$ independently to the distribution of ν. So according to (7.1), the problem of minimizing the value $D\nu$ arises.

Lemma 6.1. *The value $D\nu$ is minimal in the class Σ_q of integer random variables ν with given value $\mathrm{E}\nu = q$, while $D\nu = (q - [q])(1 - q + [q])$, if*

$$P(\nu = [q]) = 1 - (q - [q]), \quad P(\nu = [q] + 1) = q - [q]. \tag{7.2}$$

Proof of the lemma followes directly from well-known "centre" representation of the variance:

$$D\nu = \mathrm{E}(\nu - ([q] + 1/2))^2 - (\mathrm{E}\nu - ([q] + 1/2))^2,$$

since the first item here for the distribution (7.2) equals to $1/4$, and for any another distribution from Σ_q it is less than $1/4$.

7 The elementary histogram-type functional estimate

In the particle transfer problems with branching and intersection of trajectories (see sect. 7, 4) statistical simulation gives the realization of the point field of N particles (or collision) arrangement in the l-dimensional phase space X. Further the

corresponding density $\varphi(x)$ is estimated usually as histogram by determining the frequencies $\{n_i/N\}$, where n_i is the random points number in the cell Δ_i (of the domain $D \subset X$) with the phase volume h^l, $i = 1, 2, \ldots$. Consider the problem of the uniform (with respect to i) minimizing the histogram -estimate error by choosing of h relativily to N. It is clear that the practically sufficient solution of this problem is obtained in the case when the probabilistic error approximately equals to the deterministic one:

$$D\big(\frac{n_i}{Nh^l}\big) \approx \sup_{x\in\Delta_i} \big(\varphi(x) - \mathrm{E}(\frac{n_i}{Nh^l})\big)^2. \tag{8.1}$$

Under the assumption about weak the values n_i dependency it is possible for optimizing the histogram to consider the above-mentioned point field as the Poisson one with the intensity $\varphi(x)$. Because of that

$$\mathrm{E}(n_i) = D(n_i) = N \int\limits_{\Delta_i} \varphi(x)\ dx,$$

$$D(\frac{n_i}{Nh^l}) \approx \frac{\int\limits_{\Delta_i} \varphi(x)\ dx}{Nh^{2l}} \approx \frac{\tilde{\varphi}_i}{Nh^l},$$

where $\tilde{\varphi}_i = h^{-l} \int\limits_{\Delta_i} \varphi(x)\ dx$ is the middle-integral value of $\varphi(x)$ in Δ_i. Therefore it is possible to rewrite (8.1) in the form:

$$\frac{\tilde{\varphi}_i}{Nh^l} \approx c_i h^2,$$

where $c_i \approx \sup\limits_{x\in\Delta_i} |grad\varphi(x)|^2$.

So, the approximately optimal (for the i-th cell) value h is expressed by the formula

$$h \approx \big(\frac{\tilde{\varphi}_i}{Nc_i}\big)^{\frac{1}{l+2}}.$$

For uniform optimizing the histogram it is necessary to average this value with respect to i. Here the weak fluctuating the value $\tilde{\varphi}_i/c_i$ is desirable. In particular it is fulfilled for exponential densities $\varphi(x)$.

Remark, that the above-mentioned approach was used in [1] for optimizating the global statistical estimate of the solution of a nonlinear kinetic equation.

8 The distributive method of using the pseudo-random numbers

It is expedient to determine different pseudo-random numbers subsequences for realizing the corresponding statistical trials, i. e. trajectories of the investigating random process. It is possible to name this method as distributive one, because it is especially

convenient for distributed computating (see, example, [6]). To realize this method it is possible to use an auxiliary generator, which gives initial numbers for above basic subsequences totally determining them [2]. If the basic generator is multiplicative congruent one with the factor M then it is expedient to use as auxiliary one the similar generator with the factor M^{μ}, where μ is the necessary length of a basic subsequence. So the partition of the basic congruent pseudo-random sequence is realized.

The distributive method correlates statistical estimates for different versions of computations, improving the parametrical analysis of results.

Acknowledgements

This work was supported by the Russian Foundation, projects nos. 12-01-00034, 13-01-00441, 13-01-00746 and by Multidisciplinar Integral projects SB RAS nos. 47, 126.

References

[1] *Mikhailov G.A., Rogazinskii S.V.* Probabilistic model of many-particle evolution and estimation of solutions to a nonlinear kinetic equation // Russ. J. Numer. Anal. Math. Modelling. - 2012. - Vol. 27, No. 3. - P. 229-242.

[2] *Mikhailov G.A.* Some Questions on Theory of Monte Carlo Methods. Nauka, Novosibirsk, 1974.

[3] *Averina T.A., Mikhailov G.A.* Algorithms for exact and approximate Statistical Simulation of Poisson Ensembles // Comp. Math. and Math. Phys.- 2010.- Vol. 50, No. 6, pp. 951-962.

[4] *Ambos A. Ju., Mikhailov G.A.* Statistical simulation of an exponentially correlated many-dimensional random field // Russ. J. Num. Anal. Math. Model. - 2011. - Vol. 26, No. 3. - P. 263-273.

[5] *Brednikhin S.A., Medvedev I.N., Mikhailov G.A.* Estimation of the criticality parameters of branching progesses by the Monte Carlo method // Comp. Math. and Math. Phys.- 2010.- Vol. 50, No. 2, pp. 345-356.

[6] *Marchenko M. A., Mikhailov G.A.* Distributed Computing by the Monte Carlo Method // Automation and Remote Control, 2007, Vol. 68, No. 5, pp. 888-900.

Some Problems of Statistical Simulation of the Polarized Radiation Transfer

Gennady A. Mikhailov[1,2] and Sergey A. Ukhinov[1,2]
[1] *Novosibirsk State University, Russia*
[2] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia*
e-mail: gam@sscc.ru, sau@sscc.ru

Abstract

Some aspects of usage and substantiation of standard vectorial algorithm of statistical modeling of polarized radiation transfer are considered. Due to the fact that the appropriate statistical estimates can have the infinite variance, the method of "ℓ-fold polarization", in which recalculation of a Stokes vector on a "scalar" trajectory is carried out no more, than ℓ times, is offered deprived of this deficiency. Dual representation of mean squares of the Monte Carlo estimates of studied functionals and an evaluation of vector estimates variances are considered also.

Keywords: transfer of polarized radiation, statistical simulation, variance of a standard vector Monte Carlo estimate

Introduction

Light propagation can be treated as a random Markov chain of photon-substance collisions that lead to either photon scattering or photon absorption. In the Monte Carlo method, the trajectories of this chain are simulated on a computer and statistical estimates for the desired functionals are computed. The construction of random trajectories for a physical model of the process is known as direct simulation. No weights are used, and the variances of Monte Carlo estimates are always finite (see [2]). In the case of considered polarized radiation, a general matrix-weighted algorithms for solving systems of radiative transfer integral equations with allowance for polarization were constructed and preliminarily studied in [2, 5].

This paper is devoted to additional researches of the variant of the matrix-weight algorithm based on direct simulation of scalar transfer process. Due to the fact that the appropriate statistical estimates can have the infinite variance, the method of "ℓ-fold polarization", in which recalculation of a Stokes vector on a scalar trajectory is carried out no more, than ℓ times, is offered deprived of this deficiency. Thus polarization is not exactly taken into account, but errors of required estimates can be quite small.

Also this paper examines the finiteness of the variance of corresponding standard vector Monte Carlo estimates, which is required for constructing the correct confidence intervals. To this end, in [5] is considered the system of integral equations defining the covariance matrix of a weighted vector estimate. Numerical estimates based on the iteration of the resolvent showed that the spectral radius of the corresponding matrix-integral operator is fairly close to the product of the spectral radius

for an infinite medium, which is calculated analytically, and the easy-to-estimate spectral radius of the scalar integral operator associated with an identity scattering matrix. In the purpose of enhancement of analytical study of this practically important factorization, in this paper is given obtained at [4] dual (to the one which is considered in [3]) representation of the mean square error of the estimates of considered functionals.

1 General information

Various methods are available for describing the polarization properties of light. The most widespread and convenient method is that proposed by Stokes in 1852, who introduced four parameters I, Q, U, V with the dimension of intensity, which determine the intensity, degree of polarization, polarization plane, and degree of ellipticity of radiation. In what follows, we consider the corresponding components of the Stokes vector function of light intensity:

$$\mathbf{I}(\mathbf{r},\omega) = \big(I_1(\mathbf{r},\omega), I_2(\mathbf{r},\omega), I_3(\mathbf{r},\omega), I_4(\mathbf{r},\omega)\big)^T.$$

The simplest phenomenological Markov model of polarized radiative transfer arises when the medium is assumed to be isotropic. The only difference from the standard scalar model is that the scattering phase function is replaced with a scattering matrix, which transforms the Stokes vector associated with a given photon at a scattering point (see [2, 6]).

We used the following notations: $x = (\mathbf{r},\omega)$ is a point of the phase space, $\mathbf{r}$ is a point of R^3 space, $\omega = (a,b,c)$ is a unit direction vector aligned with the run of the particle ($a^2+b^2+c^2=1$); $\mu = (\omega,\omega')$ is the cosine of the scattering angle, φ is the azimuthal scattering angle, $r_{11}(\mu)$ is the scattering phase function, $\sigma(\mathbf{r})$ is the extinction coefficient, $q(\mathbf{r})$ is the probability of scattering, l is the free path, $p_\chi(l;\mathbf{r}',\omega)$ is the sub-stochastic distribution density of the free path l from the point $\mathbf{r}'$ in the direction ω: $p_\chi\left(l;\mathbf{r}',\omega\right) = \sigma\left(\mathbf{r}'+\omega l\right)\exp\left(-\tau_{op}\left(l;\mathbf{r}',\omega\right)\right), \quad l \le l^*\left(\mathbf{r}',\omega\right);\ \tau_{op}(l;\mathbf{r}',\omega) = \tau_{op}(\mathbf{r}',\mathbf{r}) = \int\limits_0^l \sigma\left(\mathbf{r}'+s\omega\right)ds$ is the optical length of the interval $\left[\mathbf{r}',\mathbf{r}'+l\omega=\mathbf{r}\right]$, and $l^*\left(\mathbf{r}',\omega\right)$ is the distance from the point $\mathbf{r}'$ in the direction ω up to the boundary of the medium, which may be assumed to be convex. Here, the trajectory can terminate since the particle escapes from the medium.

Let $F(x), H(x)$ are the column vectors of the functions $f_1(x),...,f_4(x)$ and $h_1(x),...,h_4(x)$, respectively, and

$$\Phi(x) = (\varphi_1(x),\varphi_2(x),\varphi_3(x),\varphi_4(x))^T = \sigma(\mathbf{r})\mathbf{I}(x)$$

is the vector density of collisions.

The system of integral equations describing radiative transfer with allowance for polarization has the following matrix kernel:

$$K(x',x) = \frac{q(\mathbf{r}')e^{-\tau_{op}(\mathbf{r}',\mathbf{r})}\sigma(\mathbf{r})P(\omega',\omega,\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^2} \times \delta\Big(\omega - \frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}\Big).$$

Thus, we have a vector-integral equation of transfer with allowance for polarization with respect to the vector function Φ:

$$\Phi(x) = \int\limits_X K(x', x)\Phi(x')dx' + F(x), \quad \Phi = \mathbf{K}\Phi + F. \tag{1}$$

Lets call the operator $\mathbf{K}$ the matrix-integral transfer operator. Monte Carlo algorithms are based on a representation of the solution of the Eq. (1) in the form of a Neumann series. Such a representation holds if the norm of the operator $\mathbf{K}$ (or of its power $\mathbf{K}^{n_0}$) is less than unity [2, 5].

Linear functionals of the solution of the integral equation are usually estimated by applying Monte Carlo methods. In the case of a system of second-kind integral equations, the general Monte Carlo algorithm for estimating such functionals can be described as follows.

Suppose that we want to calculate the functional

$$\mathrm{I}_H = (\Phi, H) = \sum_{i=1}^{m} \int\limits_X \varphi_i(x)h_i(x)dx = \sum_{n=0}^{\infty} (\mathbf{K}^n F, H) \ .$$

Here, H is a vector function with absolutely bounded components; i.e., $H \in L_\infty$. A homogeneous Markov chain $\{x_n\}$ in the phase space X is defined by the probability density $\pi(x)$ of the initial state x_0, by the transition probability density $r(x', x)$ from x' to x, and by the probability $p(x')$ that the trajectory terminates in the transition from the state x'. The function $p(x', x) = r(x', x)[1 - p(x')])$ is called the transition density.

An auxiliary random vector $\mathbf{Q}$ of weights is defined by the formulas

$$\mathbf{Q}_0 = \frac{F(x_0)}{\pi(x_0)}, \quad \mathbf{Q}_n = [K(x_{n-1}, x_n)/p(x_{n-1}, x_n)]\mathbf{Q}_{n-1}, \quad Q_n^{(i)} = \sum_{j=1}^{4} Q_{n-1}^{(j)} \frac{k_{ij}(x_{n-1}, x_n)}{p(x_{n-1}, x_n)}.$$

We assume that the weight factors are limited, i.e.

$$\frac{k_{ij}(x', x)}{p(x', x)} < C < +\infty \qquad \forall \quad x, x' \in X; \qquad i, j \in \overline{1, 4}.$$

By analogy with a single integral equation, it is shown (see [2, 3]) that $I_H = (\Phi, H) = \mathsf{E}\zeta$, where

$$\zeta = \sum_{n=0}^{N} \mathbf{Q}_n^T H(x_n) = \sum_{n=0}^{N} \sum_{i=1}^{4} Q_n^{(i)} H_i(x_n). \tag{2}$$

Here, N is the random index of the last state of the chain. Relation (2) describes the Monte Carlo algorithm for estimating I_H. The substantiation of this relation essentially relies on the expansion the solutions of equations in the Neumann series (see [2]). Since the first component in (2) is nonnegative, it can be averaged term by term. The remaining components can be averaged because of the majorant property of the first component (see [2]).

Consider the Monte Carlo algorithm for computing the intensity and polarization of multiply scattered light. The simplest part in this problem is the transition probability density $r(x', x)$, which is defined by the kernel $k_{11}(x', x)$ corresponding to radiative transfer without allowance for polarization. Obviously, in the simulation of such process, the vector of "weights" after scattering has to be transformed by a matrix with the elements $k_{ij}(x', x)/k_{11}(x', x)$.

As was mentioned above, a light ray is characterized by the Stokes vector $\mathbf{I} = (I, Q, U, V)$. The unscattered solar light $\mathbf{I}_0$ is assumed to be natural; i.e., $\mathbf{I}_0 = (I_0, 0, 0, 0)^{\mathrm{T}}$.

After scattering, the Stokes vector $\mathbf{I}$ is transformed according to the formula

$$\mathbf{I}(\mathbf{r}, \omega) = P(\omega', \omega, \mathbf{r}) \cdot \mathbf{I}(\mathbf{r}, \omega'),$$

where $P(\omega', \omega, \mathbf{r}) = L(\pi - i_2)R(\omega', \omega, \mathbf{r})L(-i_1)/2\pi$,

$$L(i) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 2i & \sin 2i & 0 \\ 0 & -\sin 2i & \cos 2i & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Here, i_1 is the angle between the plane ω', s and the scattering plane ω', ω; i_2 is the angle between the scattering plane ω', ω and the plane ω, s; and s is a vector of the local spherical system of coordinates [2].

For an anisotropic medium, all 16 components of the scattering matrix $R(\omega', \omega, \mathbf{r})$ are generally different. For an isotropic medium, the scattering matrix simplifies to

$$R(\omega', \omega, \mathbf{r}) = \begin{pmatrix} r_{11} & r_{12} & 0 & 0 \\ r_{21} & r_{22} & 0 & 0 \\ 0 & 0 & r_{33} & r_{34} \\ 0 & 0 & -r_{43} & r_{44} \end{pmatrix}, \quad r_{ij} \equiv r_{ij}(\mu, \mathbf{r}).$$

If the scattering particles are homogeneous spheres, then $r_{11} = r_{22}$, $r_{12} = r_{21}$, $r_{33} = r_{44}$, $r_{34} = r_{43}$. The matrix R is normalized so that $\int\limits_{-1}^{1} r_{11}(\mu)\, d\mu = 1$.

New photon's direction ω after scattering is defined by the scattering angle θ and the azimuthal angle φ. The cosine μ of the angle θ is simulated according to the r_{11}, i.e., according to the scattering phase function. The angle $\varphi \in (0, 2\pi)$ is assumed to be isotropic and is equal to that between the planes ω', s and ω, ω' measured counterclockwise when viewed against the incident ray ω'. Thus, the azimuthal angle is equal to i_1. After the new direction was chosen, i_1 and i_2 can be found using spherical trigonometry formulas.

The procedure for updating the Stokes vector after scattering includes the formulas

$$\begin{gathered} I(\mathbf{r}, \omega) = r_{11} \cdot I(\mathbf{r}, \omega') + r_{12} \cdot A, \\ Q(\mathbf{r}, \omega) = (r_{21} I(\mathbf{r}, \omega') + A r_{22}) \cos 2i_2 - (r_{33} B - r_{34} V(\mathbf{r}, \omega')) \sin 2i_2, \\ U(\mathbf{r}, \omega) = (r_{21} I(\mathbf{r}, \omega') + A r_{22}) \sin 2i_2 + (r_{33} B - r_{34} V(\mathbf{r}, \omega')) \cos 2i_2, \\ V(\mathbf{r}, \omega) = r_{43} B + r_{44} V(\mathbf{r}, \omega'), \end{gathered} \tag{3}$$

where $A = Q(\mathbf{r}, \omega') \cos 2i_1 - U(\mathbf{r}, \omega') \sin 2i_1$, $\quad B = Q(\mathbf{r}, \omega') \sin 2i_1 + U(\mathbf{r}, \omega') \cos 2i_1$.

2 Method of ℓ-fold polarization

The "scalar" integral equation $\varphi = K\varphi + f$ [2] corresponding to the base scalar model of radiation transfer can be written in the vector form:

$$\Phi_0 = \mathbf{K}_0\Phi_0 + F_0,$$

where $\Phi_0 = (\varphi, 0, 0, 0)^{\mathrm{T}}, F_0 = (f, 0, 0, 0)^{\mathrm{T}}$ and $\mathbf{K}_0$ is matrix-integral operator corresponding to the diagonal scattering matrix: $R = diag(r_{11}, r_{11}, r_{11}, r_{11})$.

After ℓ iterations of the equation (1) beginning with Φ_0, we get such approximation to the solution Φ:

$$\Phi_\ell = \mathbf{K}^\ell\Phi_0 + \sum_{n=0}^{\ell-1}\mathbf{K}^n F = \sum_{n=0}^{\infty}\mathbf{K}^\ell\mathbf{K}_0^n F_0 + \sum_{n=0}^{\ell-1}\mathbf{K}^n F. \tag{4}$$

We designate the usage of formula (4) for the approximate computation as "the method of ℓ-fold polarization" (see, also, [1]).

For constructing the corresponding estimate we should use instead of ζ from (2) the following random variable:

$$\zeta_\ell = \sum_{n=0}^{\infty}\delta_n q_n \tilde{Q}_n^{\mathrm{T}} H(x_{n+\ell}) + \sum_{n=0}^{\min(\ell-1,N)} Q_n^{\mathrm{T}} H(x_n).$$

Here q_n are scalar weights, i.e.

$$q_0 = \frac{f(x_0)}{\pi(x_0)}, \quad q_n = q_{n-1}\frac{k_{11}(x_{n-1}, x_n)}{p(x_{n-1}, x_n)},$$

and the vector weight $\tilde{Q}_n$ corresponding to ℓ-fold polarization is calculated by the formula

$$\delta_{n+\ell}\frac{K(x_{n+\ell-1}, x_{n+\ell})}{p(x_{n+\ell-1}, x_{n+\ell})} \cdot \ldots \cdot \delta_{n+1}\frac{K(x_n, x_{n+1})}{p(x_n, x_{n+1})} I_0,$$

where $I_0 = (1, 0, 0, 0)^T$; δ_n is an indicator that a trajectory doesn't terminate before the state x_n.

The point of special interest for the solution of atmospheric optics problems (see, for example, [7]) is the estimate of an influence of polarization on the intensity of radiation, i.e. the difference $\Delta_p(x) = \varphi_1(x) - \varphi_1^{(0)}(x)$, where $\varphi_1^{(0)}(x)$ corresponds to approximate scalar model.

Quantity $\Delta_p(x)$ is an error in intensity estimate $\varphi_1(x)$ caused by non-account of polarization. We denote the value $\Delta_p(x)$ produced by ℓ-fold polarization as $\Delta_p^{(\ell)}(x)$.

If a source of radiation is non-polarized, i.e. $F_0 = (f, 0, 0, 0)^{\mathrm{T}}$, then we have, due to (3), $\Delta_p^{(1)}(x) = 0$. Hence, "in first approximation" for estimate of $\Delta_p(x)$ we should use value $\Delta_p^{(2)}(x)$, whose statistical estimate is easy to find from formulas (3).

Let's denote x_n, x_{n+1}, x_{n+2} as x'', x', x and let I_D be an indicator of domain $D \subset X$. In case $F \equiv F_0$ and $H = (I_D, 0, 0, 0)^{\mathrm{T}}$, which corresponds to the estimate of the integral $\int\limits_D \varphi_1(x)dx$ for non-polarized source, we have from (3):

$$q_n\tilde{Q}_n^{\mathrm{T}} = q_n\Delta'\Delta\mathbf{I}_0'(r_{11}r_{11}' + r_{21}(\mu')r_{12}(\mu)\cos 2i_2'\cos 2i_1 - r_{22}(\mu')r_{12}(\mu)\sin 2i_2'\sin 2i_1),$$

where Δ', Δ are indicators that trajectory doesn't terminate with transition to points x', x. Due to finiteness of weight multipliers the vector norm $\|\tilde{Q}\|$ of auxiliary weight is uniformly bounded and $\mathsf{D}\zeta_\ell < \infty$ if $\mathsf{D}\zeta_0 < \infty$. The last inequality holds in case of direct simulation for basic scalar model and also when absorption or escape from medium are not simulated and instead are accounted by weight multipliers, which are equal to probabilities of these events.

3 Criterion for the finiteness of the $\mathsf{E}\zeta^2$

Consider the space $\mathbf{L}_1$ of matrix functions $\Psi(x)$ with norm $||\Psi|| = \int\limits_X \sum\limits_{i,j=1}^{m} |\Psi_{i,j}(x)|dx$ and define the linear functional

$$(\Psi, \Psi^*) = \int\limits_X \mathrm{tr}[\Psi(x)\Psi^{*\mathrm{T}}(x)]dx = \int\limits_X \sum\limits_{i,j=1}^{m} \Psi_{i,j}(x)\Psi^*_{i,j}(x)dx,$$

$\Psi^* \in \mathbf{L}_\infty, \|\Psi^*\|_{\mathbf{L}_\infty} = \mathrm{vrai}\sup\limits_{i,x} \|\Psi^*_i(x)\|$ [4].

Define also a linear operator $\mathbf{K}_p$ by

$$[\mathbf{K}_p\Psi](x) = \int\limits_X \frac{K^{\mathrm{T}}(y,x)\Psi(y)K(y,x)}{p(y,x)}dy$$

and, accordingly to [3], linear operator $\mathbf{K}^*_p$:

$$[\mathbf{K}^*_p\Psi^*](x) = \int\limits_X \frac{K(x,y)\Psi^*(y)K^{\mathrm{T}}(x,y)}{p(x,y)}dy.$$

Since $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, then $\mathrm{tr}(\Psi K\Psi^{*\mathrm{T}}K^{\mathrm{T}}) = \mathrm{tr}(K^{\mathrm{T}}\Psi K\Psi^{*\mathrm{T}})$, and therefore $(\Psi, \mathbf{K}^*_p\Psi^*) = (\mathbf{K}_p\Psi, \Psi^*)$. Moreover, we have $|(\Psi, \Psi^*)| \le ||\Psi||_{\mathbf{L}_1}||\Psi^*||_{\mathbf{L}_\infty}$.

Hence $||\mathbf{K}_p||_{\mathbf{L}_1} = ||\mathbf{K}^*_p||_{\mathbf{L}_\infty}$ and $\rho(\mathbf{K}_p) = \rho(\mathbf{K}^*_p)$.

Note also that the operator $\mathbf{K}_p$ leaves invariant the cone $\mathbf{L}^+_1 \subset \mathbf{L}_1$ of symmetric nonnegative definite matrix functions, because the transformation $K^{\mathrm{T}}\Psi K$ preserves the nonnegative definiteness of matrices Ψ.

The following statement is proved in [4].

Theorem. Suppose that $\rho(\mathbf{K}_p) < 1$, $FF^{\mathrm{T}}/\pi_0 \in \mathbf{L}_1$, $H \in L_\infty$.

Then

$$\mathrm{E}\zeta^2 = (\Psi, H[2\Phi^* - H]^{\mathrm{T}}),$$

where $\Phi^* = \mathbf{K}^*\Phi^* + H$, $\Psi = \mathbf{K}_p\Psi + FF^{\mathrm{T}}/\pi_0$, and $\Psi \in \mathbf{L}^+_1$.

Note that in [3] dual presentation of $\mathrm{E}\zeta^2$ was constructed:

$$\mathrm{E}\zeta^2 = \int\limits_X \frac{F^{\mathrm{T}}(x)\Psi^*(x)F(x)}{\pi(x)}dx = (\frac{FF^{\mathrm{T}}}{\pi}, \Psi^*),$$

where $\Psi^* = H\Phi^{*\mathrm{T}} + \Phi^*H^{\mathrm{T}} - HH^{\mathrm{T}} + \mathbf{K}^*_p\Psi^*$.

In [5] the spectral radius $\rho(\mathbf{K}_p)$ of the operator $\mathbf{K}_p$ was estimated by resolvent iterations on the basis of the limit relation of the form:

$$\frac{F^T[\lambda\mathbf{I}-\mathbf{K}]^{-(m+1)}H}{F^T[\lambda\mathbf{I}-\mathbf{K}]^{-m}H} \to \frac{1}{\lambda-\rho(\mathbf{K})}, \quad \lambda > \rho(\mathbf{K}), \quad \mathbf{I} = diag(1,1,1,1).$$

In order to improve the convergence of the algorithm in place of $H(x_n)H^{\mathrm{T}}(x_n)$ was taken $\psi^{(0)}$, i.e. the first eigenfunction of the operator S_p, which represents the realization of $\mathbf{K}_p$ for the case of the infinite medium where $F^{\mathrm{T}} = (1,0,0,0)$.

It appeared that even for optically thin layers approximate equality $\rho(\mathbf{K}_p) \approx \rho(S_p)\rho(L_p)$ is correct, here L_p is scalar integral operator with the kernel $k_{11}^2(x',x)/p(x',x)$.

In [5] is shown more detailed then in [3], that value $\lambda_0 = \rho(S_p)$ is the solution of the system of equations:

$$\begin{gathered} c_{11} + c_{21}a_1 = \lambda_0 \\ c_{12} + (c_{22}+c_{33})a_1 + c_{43}a_2 = 2\lambda_0 a_1 \\ c_{34}a_1 + c_{44}a_2 = \lambda_0 a_2 \end{gathered}$$

where $c_{ij} = \int\limits_{-1}^{1} \frac{r_{ij}^2(\mu)}{p_2(\mu)} d\mu$ and $p_2(\mu)$ is simulated distribution density of $\mu = (\omega,\omega')$.

It was found that for the aerosol scattering the value λ_0 is majorated with the value $\lambda_m = 1,178$, corresponded to the molecular scattering. On the other hand, for the real atmosphere layers value $\rho(L_p)$ is small, therefore $\rho(\mathbf{K}_p^*) < 1$ and $\mathrm{D}\zeta < +\infty$.

In [5] the results of calculations of the spectral radii of the operators $\mathbf{K}_p$ and L_p for the molecular and the aerosol scattering are presented. Obtained values of $\rho(\mathbf{K}_p)/\rho(L_p)$ statistically insignificant differ from the analytically found values $\rho(S_p)$ and are estimated with sufficient accuracy using even only the first iteration of the resolvent.

On basis of the dual representation obtained in [4] new approximate estimate of the $\rho(\mathbf{K}_p)$ is constructed [1]:

$$\rho(\mathbf{K}_p) \approx \tilde{\rho}(\mathbf{K}_p) = \frac{(\mathbf{K}_p\Psi_0, \mathbf{I})}{(\Psi_0, \mathbf{I})} \approx C\rho(\tilde{L}_p)\rho(S_p), \tag{5}$$

and a value C is not significantly different from 1. Here $\mathbf{I} = diag(1,1,1,1)$, $\Psi_0 = \tilde{\Psi}^*\tilde{\psi}(x)$, $\tilde{\Psi}^*$ is considered above eigenmatrix of the operator S_p and $\tilde{\psi}(x)$ is the main eigenfunction of the scalar operator $\tilde{L}_p$, which corresponds to the radiation model with the replacement of anisotropic scattering on an isotropic, i.e. with $r_{11} \equiv 1/2$ and $\tilde{p}_{11}(\mu) \equiv 1/2$.

This estimate (5) isn't contrary to the numerical results given in [5], because for corresponding flat layers with the isotropic scattering it was obtained that $\rho(\tilde{L}_p|\tau = 1) \approx 0.62$, $\rho(\tilde{L}_p|\tau = 2) \approx 0.78$, $\rho(\tilde{L}_p|\tau = 4) \approx 0.9$, and this values are sufficiently close to the values of $\rho(L_p)$ from [5]. The estimate (5) can be recommended for practical use taking into account that for the optically thick media the substitution

of the essentially anisotropic scattering with the isotropic scattering slightly increases the value $\rho(L_p)$.

Also the value $\rho_0(\tilde{L}_p|\tau = 10) \approx 0.974$ was obtained. Hence, we have for the flat layer with the optical thickness 10 and the molecular scattering: $\rho(\mathbf{K}_p) \approx 0.974 \cdot 1.178 = 1.15$, and with the aerosol scattering $\rho(\mathbf{K}_p) \approx 0.974 \cdot 1.02077 = 0.994$.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (13-01-00441, 12-01-00034), and by MIP SB RAS (A-47, A-52).

References

[1] Korda A.S., Mikhailov G.A., Ukhinov S.A. (2013). Mathematical problems of statistical modelling of polarized radiation transfer. *Russ. J. Numer. Anal. Math. Modelling*. Vol. **28**. No. 3, pp. 213-230.

[2] Marchuk G. I., Mikhailov G. A. , Nazaraliev M. A. , et al. (1976). *Monte Carlo Methods in Atmospheric Optics*. Nauka, Novosibirsk [in Russian].

[3] Mikhailov G. A. (1987). *Optimization of Weighted Monte Carlo Methods*. Nauka, Moscow [in Russian].

[4] Mikhailov G.A., Ukhinov S.A. (2011). Dual Representation of the Mean Square of the Monte Carlo Vector Estimator. *Doklady Mathematics*. Vol. **83**. No. 3, pp. 386-388.

[5] Mikhailov G.A., Ukhinov S.A., Chimaeva A.S. (2006). Variance of a Standard Vector Monte Carlo Estimate in the Theory of Polarized Radiative Transfer. *Computational Mathematics and Mathematical Physics*. Vol. **46**. No. 11, pp. 20062019.

[6] Sushkevich T. A. (2005). *Mathematical Models of Radiation Transfer*. BINOM, Moscow [in Russian].

[7] Ukhinov S.A., Chimaeva A.S. (2011). Convergence of Monte Carlo Algorithms for Reconstructing the Scattering Phase Function with Polarization. *Numerical Analysis and Applications*. Vol. **4**. No. 1, pp. 8192.

Modeling in Random Environments through Two-Sex Branching Processes

Molina M.[1], Mota M.[1] and Ramos A.[2]
[1] *University of Extremadura, Badajoz, Spain*
[2] *University of Extremadura, Cáceres, Spain*
e-mail: mmolina@unex.es, mota@unex.es, aramos@unex.es

Abstract

This work deals with statistical modeling through branching processes. The main motivation is the study of stochastic models to describe the demographic dynamics of sexually reproducing populations. We introduce a class of two-sex branching models where the reproduction phase is affected, in each generation, by the current numbers of females and males in the population. Under a nonparametric setting, we provide some inferential results.

Keywords: Branching models, nonparametric methods, statistical modeling.

Introduction

Branching processes are appropriate probabilistic models to describe the evolution of systems whose components (cells, particles, individuals, etc.) after certain life period reproduce and die. They have especially played a major role in modeling general population dynamics, see Jagers (1975), Kimmel and Axelrod (2002), Guttorp (1991) or Haccou *et al.* (2005). We are particularly interested in stochastic models for description of populations where females and males coexist and form couples (female-male mating units). This research line was initially considered in Daley (1968) where the called bisexual Galton-Watson process was introduced. From Daley's model, several classes of two-sex branching processes have been studied, see Hull (2003) and Molina (2010). In particular, significant efforts have been made to develop models based on the assumption that the reproduction phase is influenced by the number of couples in the population, see e.g. Molina *et al.* (2002, 2004) and Xing and Wang (2005). However, in many biological populations, it is more realistic to assume that the reproduction phase is affected by the numbers of females and males in the population. Such situations have not been studied in the literature about two-sex processes. In an attempt to contribute some solution, we introduce a class of two-sex branching processes where, in each generation, the offspring probability distribution is determined taking into account the current numbers of females and males in the population.

The paper is structured as follows: In Section 2, the new class of two-sex branching models is formally described and intuitively interpreted. Section 3 is devoted to investigating some inferential questions. Assuming a general nonparametric setting, Bayesian estimators for the offspring probability distributions involved in the model and for their main moments are determined. In order to derive highest posterior

density credibility sets, a computational algorithm is suggested. In Section 4 we present our concluding remarks.

1 The two-sex branching model

On a probability space $(\Omega, \mathcal{F}, P)$, we define the discrete-time two-sex process $\{\mathbf{X}_n\}_{n=0}^{\infty}$, $\mathbf{X}_n = (F_n, M_n)$, as follows:

$$\mathbf{X}_{n+1} = \sum_{i=1}^{Z_n} \left(f_{n,i}^{\varphi(\mathbf{X}_n)}, m_{n,i}^{\varphi(\mathbf{X}_n)} \right), \quad Z_{n+1} = L(\mathbf{X}_{n+1}), \quad n \in \mathbb{N} \tag{1}$$

where the empty sum is considered to be $\mathbf{0} = (0, 0)$ and $\mathbb{N}$ denotes the set of non-negative integers.

1. The process starts with $f_0 \geq 1$ females and $m_0 \geq 1$ males in the population, namely $\mathbf{X}_0 = (f_0, m_0)$.

2. $\mathbf{X}_{n+1} = (F_{n+1}, M_{n+1})$ represents the numbers of females and males in the $(n+1)$st generation. These females and males form Z_{n+1} couples according to the mating function L assumed to be non-decreasing in each argument, integer-valued on the integers and satisfying that $L(f, 0) = L(0, m) = 0$, $f, m \in \mathbb{N}$.

3. φ is a function defined on $\mathbb{N}^2$ and taking values in the set $\mathbb{N}_g = \{1, \ldots, g\}$, being g a positive integer which represents the number of reproductive situations that may occur in the specie. The reproductive situation depends, in each generation, of the current numbers of females and males in the population.

4. Given that $\mathbf{X}_n = \mathbf{x} = (f, m) \in \mathbb{N}^2$ then, irrespectively of $n \geq 1$,

   $$\left(f_{n,i}^{\varphi(\mathbf{x})}, m_{n,i}^{\varphi(\mathbf{x})} \right), \ i = 1, \ldots, L(\mathbf{x})$$

   are independent and identically distributed random vectors with probability distribution:

   $$\left\{ p_{k,l}^{\varphi(\mathbf{x})} \right\}_{(k,l) \in S_{\varphi(\mathbf{x})}}, \ p_{k,l}^{\varphi(\mathbf{x})} = P\left(f_{1,1}^{\varphi(\mathbf{x})} = k, m_{1,1}^{\varphi(\mathbf{x})} = l \right), \ S_{\varphi(\mathbf{x})} \subset \mathbb{N}^2.$$

   which is referred to as the offspring probability distribution when there are f females and m males in the population. The variables $f_{n,i}^{\varphi(\mathbf{x})}$ and $m_{n,i}^{\varphi(\mathbf{x})}$ represent, respectively, the number of females and males descending from the ith couple in the nth generation.

$\{\mathbf{X}_n\}_{n=0}^{\infty}$ is a process developing in an environment which changes, stochastically in time, influenced by the current number of females and males in each generation. Initially, $\mathbf{x}_0 = (f_0, m_0)$, so the number of couples is $z_0 = L(\mathbf{x}_0)$. Each of these couples produces a random number of females and males according to the offspring

probability distribution $\{p_{k,l}^{\varphi(\mathbf{x}_0)}\}_{(k,l)\in S_{\varphi(\mathbf{x}_0)}}$ and then disappears. Hence, the number of females and males in the first generation, $\mathbf{x}_1 = (f_1, m_1)$, is determined by summing the corresponding numbers of female and male descendants. These descendants produce $z_1 = L(\mathbf{x}_1)$ couples which reproduce according to the offspring probability distribution $\{p_{k,l}^{\varphi(\mathbf{x}_1)}\}_{(k,l)\in S_{\varphi(\mathbf{x}_1)}}$. Then, we obtain the next generation of females and males, $\mathbf{x}_2 = (f_2, m_2)$, and so on.

From (1) we deduce, almost surely, that:

$$E[\mathbf{X}_{n+1} \mid \mathbf{X}_0, \ldots, \mathbf{X}_n] = E[\mathbf{X}_{n+1} \mid \mathbf{X}_n],\ n \geq 1.$$

On the other hand,

$$P(\mathbf{X}_{n+1} = \mathbf{y} \mid \mathbf{X}_n = \mathbf{x}) = P\left(\sum_{i=1}^{L(\mathbf{x})} \left(f_{n,i}^{\varphi(\mathbf{x})}, m_{n,i}^{\varphi(\mathbf{x})}\right) = \mathbf{y}\right) = \left(p_{\mathbf{y}}^{\varphi(\mathbf{x})}\right)^{*\mathbf{L}(\mathbf{x})}$$

with * denoting the convolution of distributions.

Note that, if for some $n \geq 1$, $\mathbf{X}_n = \mathbf{0}$, then we have that $\mathbf{X}_{n+m} = \mathbf{0}$, $m \geq 1$.

Consequently, $\{\mathbf{X}_n\}_{n=0}^{\infty}$ is a homogeneous Markov chain, being $\mathbf{0}$ an absorbing state. For the particular case when $g = 1$, the model (1) is reduced to the classical bisexual Galton-Watson process studied in Daley (1968).

2 Some inferential results

Let $\{\mathbf{X}_n\}_{n=0}^{\infty}$ be a two-sex branching process (1). We assume that g is known and that, for each $h \in \mathbb{N}_g$, S_h is a finite set. By simplicity, let us write:

$$\mathbf{p}_h = \left(p_{k,l}^h;\ (k,l) \in S_h\right),\ p_{k,l}^h = P\left(f_{1,1}^h = k, m_{1,1}^h = l\right) > 0,\ h \in \mathbb{N}_g.$$

We are considering a nonparametric setting. Therefore, no assumption is made about the functional form of the offspring probability distributions involved in the two-sex branching model.

a Bayesian estimation

We shall determine Bayesian estimators for $\mathbf{p}_1, \ldots, \mathbf{p}_g$. To this end, we consider the observation of the entire family tree, up to the nth generation is reached:

$$\left\{\mathbf{X}_i, \left(f_{i,j}^{\varphi(\mathbf{X}_i)}, m_{i,j}^{\varphi(\mathbf{X}_i)}\right),\ i = 0, \ldots, n;\ j = 1, \ldots, L(\mathbf{X}_i)\right\}$$

For $(k,l) \in S_{\varphi(\mathbf{X}_i)}$, let

$$Z_{i,(k,l)} = \sum_{j=1}^{L(\mathbf{X}_i)} 1_{\{(f_{i,j}^{\varphi(\mathbf{X}_i)}, m_{i,j}^{\varphi(\mathbf{X}_i)}) = (k,l)\}},\ i = 0, \ldots, n$$

be the number of couples in the ith generation giving rise to exactly k females and l males. Clearly,

$$Z_i = \sum_{(k,l)\in S_{\varphi(\mathbf{X}_i)}} Z_{i,(k,l)}, \quad \mathbf{X}_{i+1} = \sum_{(k,l)\in S_{\varphi(\mathbf{X}_i)}} (k,l) Z_{i,(k,l)}, \quad i = 0, \ldots, n.$$

Let us denote by $\mathcal{L}$ the corresponding likelihood function. We deduce that:

$$\mathcal{L}\left(\mathbf{p}_1, \ldots, \mathbf{p}_g \mid \left\{\mathbf{X}_i, \left(f_{i,j}^{\varphi(\mathbf{X}_i)}, m_{i,j}^{\varphi(\mathbf{X}_i)}\right),\ i = 0, \ldots, n;\ j = 1, \ldots, L(\mathbf{X}_i)\right\}\right) \propto$$

$$\prod_{h=1}^{g} \prod_{(k,l)\in S_h} (p_{k,l}^h)^{Y_{n,(k,l)}^h} \tag{2}$$

where, for each $(k,l) \in S_h$,

$$Y_{n,(k,l)}^h = \sum_{\{i\in\{0,\ldots,n\}:\ \varphi(\mathbf{X}_i)=h\}} Z_{i,(k,l)}$$

represents the total number of couples in the first n generations which have produced exactly k females and l males.

From (2), see e.g. Bernardo and Smith (1994) or Mendoza and Gutierrez-Peña (2000), we derive that an appropriate conjugate class of prior distributions on $(\mathbf{p}_1, \ldots, \mathbf{p}_g)$ is given by the following product of Dirichlet distributions:

$$\pi(\mathbf{p}_1, \ldots, \mathbf{p}_g) = \prod_{h=1}^{g} C_{\boldsymbol{\tau}_h} \prod_{(k,l)\in S_h} (p_{k,l}^h)^{\tau_{k,l}^h - 1} \tag{3}$$

where, for $h \in \mathbf{N}_g$,

$$\boldsymbol{\tau}_h = \left(\tau_{k,l}^h;\ (k,l) \in S_h\right),\ \tau_{k,l}^h > 0$$

is a parameter vector and

$$C_{\boldsymbol{\tau}_h} = \prod_{(k,l)\in S_h} (\Gamma(\tau_{k,l}^h))^{-1} \Gamma(\tau_*^h), \quad \tau_*^h = \sum_{(k,l)\in S_h} \tau_{k,l}^h.$$

Let the σ-algebra:

$$\mathcal{A}_n = \sigma\left(\mathbf{X}_i, \left(f_{i,j}^{\varphi(\mathbf{X}_i)}, m_{i,j}^{\varphi(\mathbf{X}_i)}\right),\ i = 0, \ldots, n;\ j = 1, \ldots, L(\mathbf{X}_i)\right).$$

From (2) and (3), conditioned to $\mathcal{A}_n$, we obtain as posterior distribution of $(\mathbf{p}_1, \ldots, \mathbf{p}_g)$ the product of Dirichlet distributions:

$$\pi(\mathbf{p}_1, \ldots, \mathbf{p}_g \mid \mathcal{A}_n) = \prod_{h=1}^{g} C_{\boldsymbol{\gamma}_h} \prod_{(k,l)\in S_h} (p_{k,l}^h)^{\gamma_{k,l}^h - 1} \tag{4}$$

$$C_{\boldsymbol{\gamma}_h} = \prod_{(k,l)\in S_h} (\Gamma(\gamma_{k,l}^h))^{-1} \Gamma(\gamma_*^h), \quad \gamma_*^h = \sum_{(k,l)\in S_h} \gamma_{k,l}^h$$

$$\boldsymbol{\gamma}_h = \left(\gamma_{k,l}^h;\ (k,l) \in S_h\right),\ h = 1, \ldots, g$$

where

$$\gamma_{k,l}^h = \tau_{k,l}^h + Y_{n,(k,l)}^h.$$

Result 3.1 Considering squared error loss function, the Bayesian estimator of $p_{k,l}^h$ is given by:

$$\widehat{p_{k,l}^h} = (\gamma_*^h)^{-1}\gamma_{k,l}^h,\ (k,l) \in S_h,\ h \in \mathbb{N}_g.$$

Proof. From (4), we deduce as marginal posterior distribution corresponding to $p_{k,l}^h$ the Beta distribution with parameters $\gamma_{k,l}^h$ and $\gamma_*^h - \gamma_{k,l}^h$. Thus, assuming squared error loss function, the Bayesian estimator of $p_{k,l}^h$, $(k,l) \in S_h$ is given by:

$$\widehat{p_{k,l}^h} = E\left[p_{k,l}^h \mid \mathcal{A}_n\right] = (\gamma_*^h)^{-1}\gamma_{k,l}^h.$$

For $h \in \mathbb{N}_g$, let us denote by:

$$\mu_1^h = E[f_{1,1}^h], \quad \mu_2^h = E[m_{1,1}^h]$$
$$\sigma_{11}^h = Var[f_{1,1}^h], \quad \sigma_{22}^h = Var[m_{1,1}^h], \quad \sigma_{12}^h = Cov[f_{1,1}^h, m_{1,1}^h].$$

assumed to be finite.

Result 3.2 Considering squared error loss function, the Bayesian estimators of μ_i^h and σ_{ij}^h, $i,j = 1,2$; $h \in \mathbb{N}_g$, are given by:

$$\widehat{\mu_i^h} = (\gamma_*^h)^{-1} \sum_{(k_1,k_2)\in S_h} k_i\gamma_{k_1,k_2}^h,$$

$$\widehat{\sigma_{ij}^h} = (\gamma_*^h(1+\gamma_*^h))^{-1}(\gamma_*^h \sum_{(k_1,k_2)\in S_h} k_ik_j\gamma_{k_1,k_2}^h - \sum_{(k_1,k_2),(l_1,l_2)\in S_h} k_il_j\gamma_{k_1,k_2}^h\gamma_{l_1,l_2}^h).$$

Proof. Taking into account Result 3.1, we deduce that:

$$\begin{aligned}\widehat{\mu_i^h} &= E[\sum_{(k_1,k_2)\in S_h} k_ip_{k_1,k_2}^h \mid \mathcal{A}_n] = \sum_{(k_1,k_2)\in S_h} k_iE[p_{k_1,k_2}^h \mid \mathcal{A}_n] \\ &= (\gamma_*^h)^{-1} \sum_{(k_1,k_2)\in S_h} k_i\gamma_{k_1,k_2}^h,\ i = 1,2.\end{aligned}$$

$$\begin{aligned}\widehat{\sigma_{ij}^h} &= E[\sum_{(k_1,k_2)\in S_h} (k_i - \mu_i^h)(k_j - \mu_j^h)p_{k_1,k_2}^h \mid \mathcal{A}_n] \\ &= \sum_{(k_1,k_2)\in S_h} k_ik_jE\left[p_{k_1,k_2}^h \mid \mathcal{A}_n\right] - \sum_{(k_1,k_2)\in S_h} k_ik_jE\left[(p_{k_1,k_2}^h)^2 \mid \mathcal{A}_n\right] \\ &- \sum_{(k_1,k_2)\neq(l_1,l_2)} k_il_jE\left[p_{k_1,k_2}^h p_{l_1,l_2}^h \mid \mathcal{A}_n\right],\ i,j = 1,2.\end{aligned}$$

The proof is completed using that:

$$E\left[(p^h_{k_1,k_2})^2 \mid \mathcal{A}_n\right] = \left(\gamma^h_*(\gamma^h_*+1)\right)^{-1}\gamma^h_{k_1,k_2}(\gamma^h_{k_1,k_2}+1).$$

$$E\left[p^h_{k_1,k_2}p^h_{l_1,l_2} \mid \mathcal{A}_n\right] = \left(\gamma^h_*(\gamma^h_*+1)\right)^{-1}\gamma^h_{k_1,k_2}\gamma^h_{l_1,l_2},\ (k_1,k_2)\neq(l_1,l_2).$$

b Highest posterior density credibility sets

From $\pi(p^h_{k,l} \mid \mathcal{A}_n)$, we can determine sets of probable values of $p^h_{k,l}$. The most common procedure is based on looking at the points where the posterior density takes the highest values:

$$I(c) = \{p^h_{k,l} : \pi(p^h_{k,l} \mid \mathcal{A}_n) \geq c\}$$

where the constant c is chosen such that, given a credibility coefficient $1-\alpha$,

$$\int_{I(c)} \pi(p^h_{k,l} \mid \mathcal{A}_n)dp^h_{k,l} = 1-\alpha.$$

We say that $I(c)$ is a high posterior density credibility set.

From the posterior densities of μ^h_i and σ^h_{ij}, $i,j=1,2$, we could derive the corresponding highest posterior density credibility sets. It is not easy to compute such posterior densities. By applying Monte Carlo method we can derive very accurate approximations. We suggest the following procedure:

1. We generate a sufficiently large number of values for $p^h_{k,l}$ according to the posterior density $\pi(p^h_{k,l} \mid \mathcal{A}_n)$.

2. Using that:

$$\mu^h_i = \sum_{(k_1,k_2)\in S_h} k_i p^h_{k_1,k_2}, \quad \sigma^h_{ij} = \sum_{(k_1,k_2)^h\in S} k_i k_j p^h_{k_1,k_2} - \mu^h_i\mu^h_j,\ i,j=1,2$$

   we calculate the corresponding values of μ^h_i and σ^h_{ij}.

3. By applying a Gaussian kernel method, see e.g. Silverman (1986), we estimate the posterior densities of μ^h_i and σ^h_{ij}, $i,j=1,2$.

Conclusions

Inside the general context of statistical modeling, we have focused our interest on stochastic models describing the probabilistic evolution of two-sex populations with sexual reproduction. To this end, several classes of branching processes have been studied. However, the range of models investigated is not large enough to get an optimum modeling in some populations. It might seem conceivable that by several

factors, in many biological populations the reproduction phase is influenced by the numbers of females and males in the population. These situations have not been considered in the literature about two-sex processes. The motivation behind this work has been the interest in developing appropriate stochastic models to describe the demographic dynamics in such populations. To sum up:

- We have introduced a new class of two-sex branching models which considers several reproductive situations. In each generation, the offspring probability distribution which governs the reproduction phase, is determined according to the numbers of females and males in the population.

- We have developed some inferential methods. By considering a general non-parametric setting, we have determined Bayesian estimators for the offspring probability distributions involved in the probability model (Result 3.1) and for their corresponding mean vectors and covariance matrices (Result 3.2).

- We have suggested a computational algorithm to determine the corresponding highest posterior density credibility sets.

It is worth pointing out that in addition to its theoretical interest, the class of two-sex branching processes introduced has also practical implications, especially in population dynamics. In this respect, we have developed a specific software for both the simulation of the model and the practical application of the inferential methods studied.

Acknowledgements

This research has been supported by the Gobierno de Extremadura (grant GR10118), the Ministerio de Economía y Competitividad of Spain (grant MTM2012-31235) and the FEDER.

References

[1] Bernardo J.M., Smith A.F. (1994). *Bayesian Theory*. Wiley, Chichester.

[2] Daley D.J. (1968). Extinction conditions for certain bisexual Galton-Watson branching processes. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*. Vol. 9, pp. 315-322.

[3] Guttorp P. (1991). *Statistical Inference for Branching Processes*. Wiley, New York.

[4] Haccou P., Jagers P., Vatutin V. (2005). *Branching Processes: Variation, Growth, and Extinction of Populations*. Cambridge University Press, New York.

[5] Hull D.M. (2003). A survey of the literature associated with the Bisexual Galton-Watson branching process. *Extracta Mathematicae*. Vol. 18, pp. 321343.

[6] Jagers P. (1975). *Branching Processes with Biological Applications*. Wiley, London.

[7] Kimmel M., Axelrod D.E. (2002). *Branching Processes in Biology*. Springer-Verlag, New York.

[8] Mendoza M., Gutiérrez-Peña E. (2000). Bayesian conjugate analysis of the Galton-Watson process. *Test*. Vol. 9, pp. 149-172.

[9] Molina M. (2010). Two-sex branching process literature. *Lectures Notes in Statistics*. Vol. 197, pp. 279-293.

[10] Molina M., Mota M., Ramos A. (2002). Bisexual Galton-Watson branching process with population-size dependent mating. *Journal of Applied Probability*. Vol. 39, pp. 479-490.

[11] Molina M., Mota M. Ramos, A. (2004). Limit behaviour for a supercritical bisexual Galton-Watson branching process with population-size dependent mating. *Stochastic Processes and their Applications*. Vol. 112, pp. 309-317.

[12] Silverman B.W. *Density Estimation for Statisitics and Data Analysis*. Chapman and Hall, London.

[13] Xing Y., and Wang Y. (2005). On the extinction of one class of population-size-dependent bisexual branching processes. *Journal of Applied Probability*. Vol. 42, pp. 175-184.

Parametric Method for Observational Results Processing based on Right Censored and Missing Data

Dmitry Nikolaev
Institute of Nuclear Power Engineering,
National Research Nuclear University "MEPhI",
Obninsk, Russia
e-mail: dafanday@gmail.com

Abstract

This paper reviews the current state of the problem of statistical analysis with missing data and methods of its solution. It is proposed to use a resampling-method for solving this problem. The work is aimed at demonstration of resampling possibilities for this task and investigation of the effectiveness of a resampling method.

Keywords: missing data, maximum-likelihood estimation, resampling-method.

Introduction

Evaluation of the nuclear power plant equipment reliability is an important issue in the development of nuclear power engineering. These estimates are important for planning maintenance, repairs and determination of the spare elements composition. NPP equipment falls into the category of highly reliable equipment, the failures of such facilities are rare events. Therefore, statistical methods are used to predict the equipment lifetime. To achieve this task, you should use all available information on the functioning of one-type equipment, including information on NPP operation without failure and information about equipment operation when the failure fixation has not occurred. Thus, the objective is to develop a parametric approach to restoration of the density distribution of time to failure based on censored data and missing data.

Today this objective is particularly relevant, as far as many nuclear power plants of Russia were built in the 60s and 70s, the equipment during that time was heavily worn, the cracks began to appear. Data about failures have not always been collected systematically. Requirements to data gathering were formulated in the middle of 80s. Therefore, there is a amount of facilities with samples of missing data.

Knowing all the information about the research object, the most appropriate method for estimating the density distribution of time to failure is parametric. This is caused by fact that the distribution function is known with the accuracy within a set of parameters measured for the sample given.

For evaluation of parameters distribution laws we use the maximum-likelihood estimation method (MLE). This method is widely used in estimating the parameters

of complex systems. It also serves as the basis for procedures of testing statistical hypotheses and confidence interval estimation. Performance estimates obtained produced by the method of maximum likelihood have a amount of important properties, such as unbiasedness, asymptotic efficiency, consistency. The likelihood function has only single maximum, which considerably simplifies the analysis of the distribution parameter estimates.

Since this paper is aimed at analyzing the problem of reliability characteristics taking into account missing data, consequently, before using MLE, it is essential to restore missing data. A resampling-method will be used to restore the missing data. This method combines four basic approaches that differ in the algorithm, but they are similar in effect: randomization or permutation test, bootstrap, the method of jackknife and the method of cross-validation. They belong to the new class of a computer-intensive technology and are widely used to verify the statistical hypothesis or to obtain unbiased characteristics of the parameter required, such as of expectation, variance and confidence interval estimates. This method is implemented in a higher-level language C/C++.

1 Data description

When solving the issues of statistical evaluation of reliability elements and systems, the task of collecting and presenting the original information about the analyzed object behavior is of special relevance. The accuracy of estimates and calculations of reliability characteristics depends on the accuracy and validity of initial information.

Therefore, in statistical analysis one has to face with a situation when at certain time intervals the information on object behavior is missing. There is a situation with missing data (Fig. 1), which significantly complicates the mathematical processing and leads to a shift in the basic statistical characteristics such as mathematical expectation and dispersion.

Fig. 1 shows the situation where data collection on object functioning has begun at a specific time moment T_1. Data collection has been performed within the $[0, T_1]$ interval. Data represent the operating time which can be both the time to failure and the operating time to censoring.

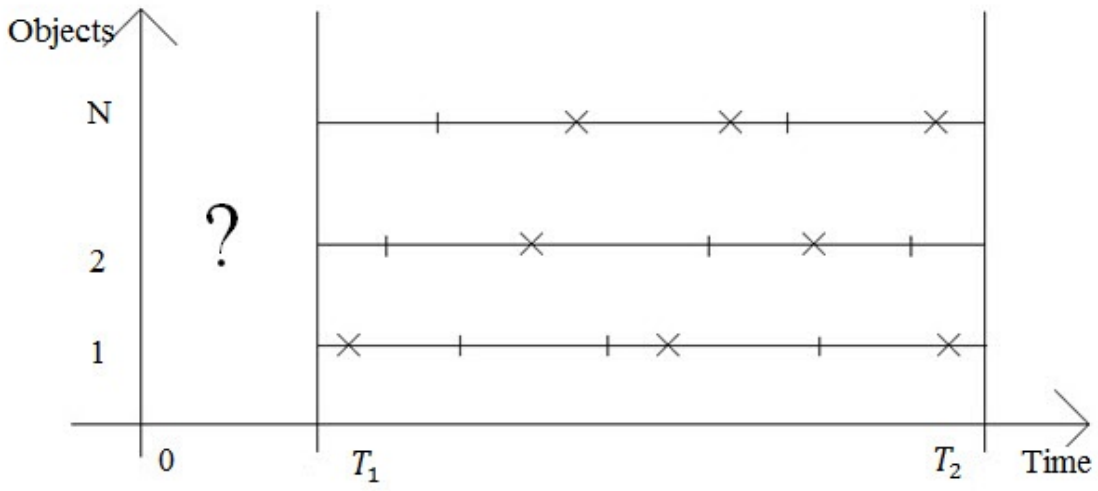

Figure 1: Data description

Fig. 1 gives censored data (dashed lines) and missing data (stars).

Let N facilities are under observation. There is a set of data at a given time for each object. During the object operation in the interval from 0 to T_1 the information on object behavior is not fixed. Data are collected from T_1 to T_2, have been: equipment failures, censored data (when the tested equipment has no faults). Based on observations results in the interval $[T_1, T_2]$ can be obtained time to failure, operating time to censoring for each object. The problem is to restore data in the interval $[T_1, T_2]$ and to estimate the distribution parameter, taking into account missing data, which can be failures as well as censored data for the whole operating time.

The reason to restore missing data is the fact that in practice statistical analysis is limited by the analysis of not total population as a whole, but only a selective amount of observations. The analyzed sample should meet high-quality and completeness criteria. In reality, we are often faced with a situation where some of the properties of one or more objects are absent, i.e. there is a situation with missing data, which significantly complicates mathematical processing, since the displacement of main statistical characteristics, such as mathematical expectation or variance, for example, increases the proportion ally to the amount of missing.

2 Development of the method for accounting the missing data by resampling

There are missing data at the time interval $[0, T_1]$, which must also be taken into account when calculating the system reliability.

The problem of reconstructing missing data in this paper is solved with a procedure based on a resampling-method.

The method consists in the following:

1. From observation results in the interval $[T_1, T_2]$ the distribution parameter estimate is recovered from data taking into account both the complete and the censored operating time.

2. The amount of operating time is predicted n in the interval $[0, T_1]$ for each group:
$$n = \frac{mT_1}{T_2 - T_1}, \tag{1}$$
where m is the amount of operating time in the interval $[T_1, T_2]$. Such a ratio is possible in the case of a uniform failure flow.

3. In parametric approach, the distribution function is being known in advance. The amount of operating time n is simulated from the distribution function.

4. Samples are combined.

5. The distribution parameter is estimated with allowance for censored and missing data.

6. The mean time and standard deviation are calculated for the original sample and the recovered data.

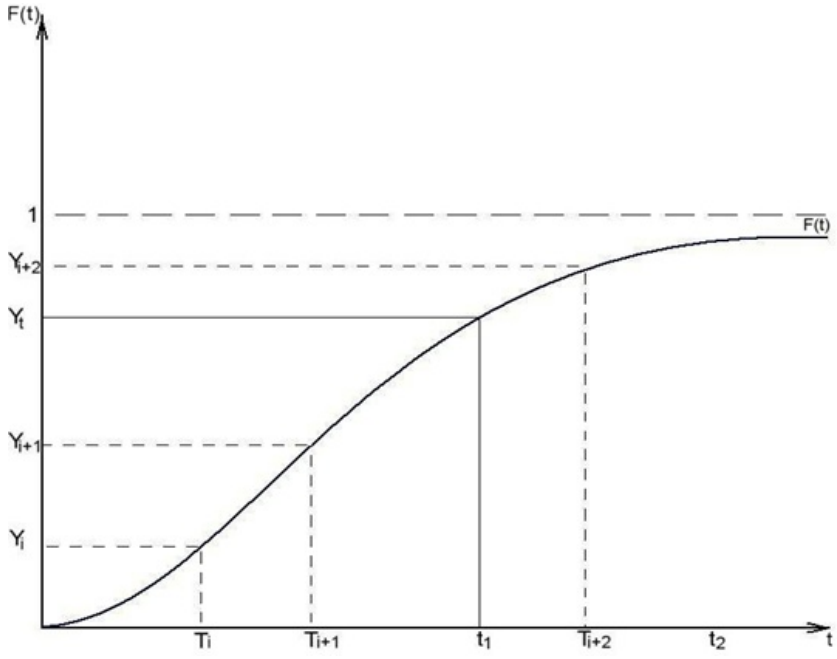

Figure 2: Method for recovering the missing data from distribution function

3 A test example

As the original information used are the data obtained from random variable simulation in accordance with the given distribution (exponential distribution law). The time to a failure is simulated with λ_1= 0.002 and censored data with $\lambda_2 = 0.003$. The amount of failures and censored data determined randomly in a given interval $[0, T_2]$, and are 242 and 54, respectively.

Having taken a simulated sample 30% of data are removed artificially. After this operation the amount of failures and censored data is 170 and 39. The distribution law is assessed for each of the resulting data sets ($\lambda_1 = 0.00180$ for the time to failure and $\lambda_2 = 0.00358$ for the operating time to censored data). The whole totality of data is also evaluated, where $\lambda = 0.00162$ and root-mean is found $\sigma = 1.211{*}10^{-4}$. Distribution law parameters and variances were assessed using the method of maximum likelihood.

Then an amount of operating time n is predicted in the interval $[0, T_1]$ for each of data groups ($n_1 = 35$ failure and $n_2 = 8$ censored).

Simulation has been performed on the basis of (λ_1 and λ_2) estimates and the predicted amount of operating time (n_1 and n_2) (Fig. 2). The simulated sample was combined with the existing full operating time for which distribution law parameters were evaluated by MMPs. Next, the pooled sample is broken down into failure and censored data and re-evaluation was done for each of groups (λ_1 and λ_2). The resulting estimates of distribution law parameter are used for new modeling of missing data. The process is repeated a several times. Then the mean value of the estimated distribution law λ = 0,00162 and root-mean σ = 1,038e-004 are found. Results for one object are presented in Table 1.

Table 1: Results

	Failure	Censored
$\lambda_{modeling}$	0.002	0.003
The amount of each operating time type simulated in the interval $[T_1, T_2]$	170	39
λ for the obtained operating time flow	0.00162	
Root-mean deviation, excluding missing data	$1.211{*}10^{-4}$	
λ for grouped data in the interval $[T_1, T_2]$	0.00180	0.00358
The predicted amount of operating time for each data type in the interval $[0, T_1]$	64	15
λ with missing data	0.00162	
Root-mean deviation with missing data	$1.038{*}10^{-4}$	

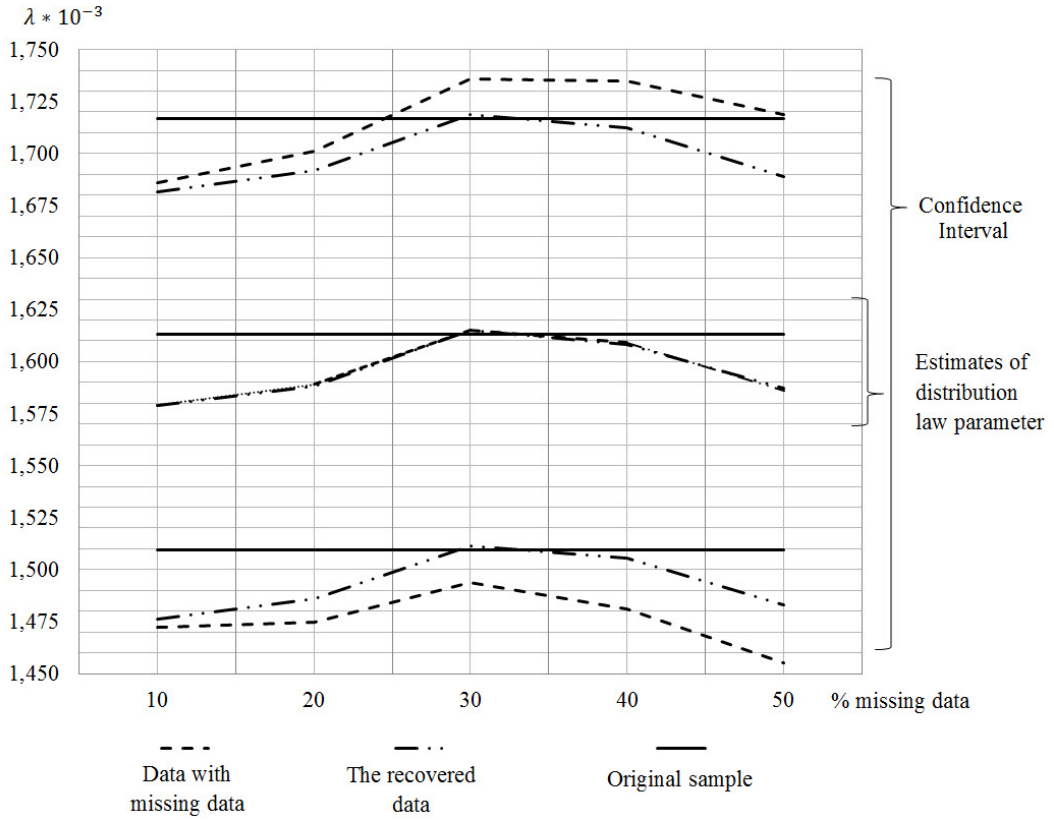

Figure 3: Estimated results and confidence intervals for different amount of missing data

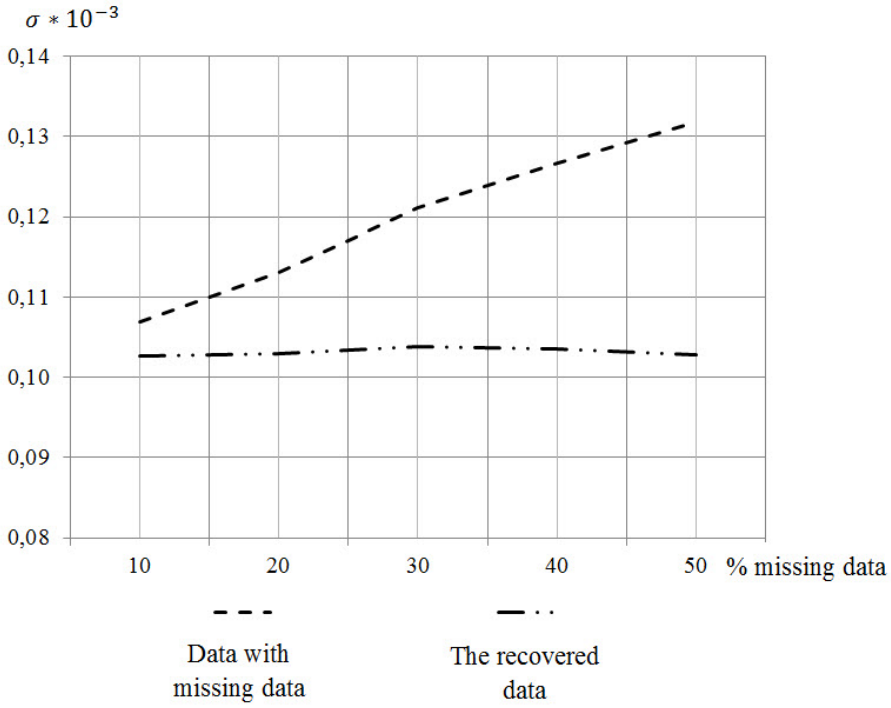

Figure 4: The variance values

Conclusions

Now it is clear that the recovered data influence the estimate. Fig. 3 shows plots of the estimations with confidence intervals. Although the assessment does not change, data recovery significantly reduces the variance when a high percentage of missing data. As a result, root-mean deviation (Fig. 4) shows the value dispersion of a random variable with respect to its mathematical expectation. Accounting missing data can improve the accuracy of produced estimates. Optimization problems controlling the operation of nuclear power plants are solved with a high degree of certainty.

References

[1] Antonov A.V. (2004). *System analysis*. Higher School, Moscow.

[2] Zloba E., Yaitsky I. (2004). Statistical methods for recovery of missing data. *Computer Modeling & New Technologies*. Vol. **6**, pp. 55-56.

[3] Little R.J.A., Rubin D.B. (1991). *Statistical analysis with missing data*. Finance and Statistics, Moscow.

Numerical Stochastic Model of Indicator Fields of Daily Sums of Liquid Precipitation

Vasily A. Ogorodnikov[1] and Olga V. Sereseva[1]
[1] *Institute of Computational Mathematics and Mathematical Geophysics, Siberian Brunch of Russian Academy of Sciences, Novosibirsk State University, Novosibirsk, Russia*
e-mail: ova@osmf.sscc.ru, seresseva@mail.ru

Abstract

The numerical stochastic model of spatio-temporal indicator fields of daily sums of liquid precipitation at the stations, as well as on regular grid on the basis of observation data is considered in this paper. The approach to modeling is based on a threshold transformation specially selected Gaussian field and stochastic interpolation of indicator fields from weather stations to a grid points. The specificity consists in the method of approximate accounting of the field spatial inhomogeneity.

Keywords: numerical stochastic model, indicator field, precipitation, inhomogeneity.

Introduction

In the given paper the numerical stochastic models of spatio-temporal indicator fields of the daily sums of liquid precipitation on a regular and irregular grids are considered. The approach to modeling is close to the approaches considered in [1], [4], [6], [8] and is based on threshold transformation of Gaussian field with the specially given correlation matrix. The method of the approaches account of heterogeneity of a field consists in modeling a non-homogeneous field at the stations with the subsequent stochastic interpolation of indicators from stations to a grid points. The results of the verification models are given in this paper. To construct the models the data of 15-year precipitation observations at 47 stations of the Novosibirsk region for the warm half of the year were used.

The spatio-temporal field can be represented as a series of spatial fields in which the temporal and spatio-temporal correlations are given on the basis of the real information. In case, when the field is homogeneous in space and stationary in time, the correlation function can be given as product of spatial and temporary correlations, which corresponds to direct product of the spatial and temporary correlation matrices constructed on these correlations for the field on a given grid. Modeling methods of Gaussian fields with the such correlation structure are known [6]. In this paper we will consider the method of modeling of heterogeneous indicator precipitation fields in which Gaussian fields at the stations are modeled by using a stationary vector autoregression process with the given matrix covariance function. The elements of this matrix are estimated by real samples. The principle of constructing indicator

fields of precipitation based on Gaussian spatio-temporal fields will be discussed in the first section.

1 Numerical stochastic model of spatial fields of daily sums of liquid precipitation at the stations

The joint stationary time series $\vec{\omega}_1, \ldots, \vec{\omega}_n = \vec{\omega}_{(n)}$ of indicators of daily precipitation at the stations (or spatio-temporal field $\{\omega_{i,t}\},\ i = 1, \ldots, m,\ \ t = 1, 2, \ldots, n,\ (m = 47)$ at the stations) are constructed as a threshold transformation of a Gaussian spatio-temporal field $\{\xi_{i,t}\}$with mean zero in the form [1], [4], [6], [8]

$$\omega_{i,t} = \begin{cases} 1, & \xi_{i,t} \le c_i \\ 0, & \xi_{i,t} > c_i \end{cases} . \tag{1}$$

Here, the quantities c_i (we will use the approach, according to which c_i do not depend on t) are determined from the equations

$$p_i = P(\omega_i = 1) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{c_i} e^{-\frac{1}{2}u^2} du \tag{2}$$

for the given probability of precipitation p_i at the stations. The elements of the block Toeplitz correlation matrix $G_{(n)} = (G_{k-l}) = (G_h),\ k, l = 1, \ldots, n$ with the blocks $G_h = (g_{ik,jk+h}),\quad i, j = 1, \ldots, m,\quad h = 0, \ldots, n-1$ or, for simplicity $G_{(n)} = G = \{g_{ij}\}$, of the Gaussian field $\{\xi_{i,t}\}$ associated with the corresponding elements of the correlation matrix $S_{(n)} = (S_h) = S = \{s_{ij}\}$of the field $\{\omega_{i,t}\}$ by the following relation [1]

$$\begin{array}{l} s_{ij} = \frac{1}{2}\left(\sqrt{\frac{p_i q_j}{p_j q_i}} + \sqrt{\frac{p_j q_i}{p_i q_j}}\right) - \frac{(T(c_i, a_{ij}) + T(c_j, a_{ij}))}{\sqrt{p_i q_i p_j q_j}}, \\ T(c_i, a_{ij}) = \frac{1}{2\pi} \int\limits_0^{a_{ij}} e^{-\frac{c_i^2(1+u^2)}{2}} \frac{du}{1+u^2},\ \ a_{ij} = \frac{c_i - c_j g_{ij}}{c_j\sqrt{1-g_{ij}^2}}, \end{array} , \tag{3}$$

where $q_i = 1 - p_i = P(\omega_i = 0)$, and $T(c_i, a_{ij})$ is the Owen function [2].

Gaussian field $\{\xi_{i,t}\}$ at the stations or stationary sequence of Gaussian vectors

$$\vec{\xi}_1, \ldots, \vec{\xi}_t, \ldots, \vec{\xi}_n = \vec{\xi}_{(n)},$$

where $\vec{\xi}_k = (\xi_{1k}, \ldots, \xi_{mk})^T$ are the vectors of dimension m with the given matrix covariance functions (or block Toepliz covariance matrices) required for the construction of the field $\{\omega_{i,t}\}$ were simulated by using a vector autoregression process [6], [5]:

$$\vec{\xi}_t = \vec{B}_1^T[q]\vec{\xi}_{t-1} + \ldots + \vec{B}_q^T[q]\vec{\xi}_{t-q} + C\vec{\varphi}_t,\ \ t = 1, \ldots, n.$$

The initial vectors are calculated recursively by the following scheme

$$\vec{\xi}_1 = C_0\vec{\varphi}_1,\quad \vec{\xi}_k = \vec{B}^T[k-1]J_{(k-1)}\vec{\xi}_{(k-1)} + C_{k-1}\vec{\varphi}_k,\quad k = 2, \ldots, q$$

where $\vec{\varphi}_k$ are independent Gaussian vectors of the dimension m,

$$J_{(k)}G_{(k)}^{-1}J_{(k)}\vec{G}_k = \vec{B}[k], \quad \vec{B}[k] = \left(B_1^T[k], \ldots, \vec{B}_r^T[k]\right)^T,$$

$$\vec{G}_k = (G_1^T, \ldots, G_k^T)^T, \quad C_k C_k^T = Q_k, \;\; Q_k = R_0 - \vec{B}^T[k]\tilde{R}_{(k)}\vec{B}[k],$$

$$J_{(k)} = \left\| \begin{array}{ccc} 0 & \cdots & I \\ \cdots & \cdots & \cdots \\ I & \cdots & 0 \end{array} \right\| \quad -$$

is a block permutation matrix, where 0 and I are the zero and identity matrices of the dimension $m \times m$ respectively. As the covariance matrix $G_{(n)}$ the matrix is obtained by the solution of the equations (3) for the given values of sampling block Toeplitz matrix $S_{(n)} = S = \{s_{ij}\}$ was used.

Probabilities of precipitations p_i and block-Toeplitz covariance matrix $S_{(n)}$ were estimated by the observation data in the stationary approximation using standard formulas to estimate probability and matrix covariance functions. In this paper, all calculations were performed for May, where the averaging was taken according to the sample size of $n \times L$ =465, where the n =31 - the number of days in May, and L =15 - the number of years of observations.

2 Estimation of the degree of inhomogeneity of the indicator fields of precipitation at the stations

In the construction of inhomogeneous precipitation fields it is necessary to know information about their structure as a function of spatial coordinates. An extent of field inhomogeneity by the one-point characteristics, such as the probabilities of rainfall p_i which, though weak, but depend on the spatial coordinates [3], is determined by choosing values c_i from the given values p_i in the transformations (1),(2).

To study the inhomogeneity of the real field with respect to correlations one can use the property of a homogeneous field, which consists in the fact that the correlation between its values at the two points do not depend on their parallel shift. Since the stations are located irregularly this condition can be checked only approximately. Consider two pairs of stations. Let's displace in parallel the second pair so that the first point of this pair coincides with the first point of the first pair. We denote the ratio of the length of the segment connecting the free points to the average length of the intervals between points within each of the pairs through $\nu = \Delta s/\bar{s}$. If ν is small, approximately we will assume these pairs are parallel, and the points in them - equidistant. As a measure of a deviation of the considered field from homogeneous one let us consider the value $\Delta r_f/\bar{r}_f = \lambda$, where Δr_f - the modulus of the difference between the correlation coefficients of values of the field at the points relating to each of pair in the considered double pairs, and $\bar{r}_f$ - the correlation coefficient obtained by averaging of these coefficients. Of all the possible binary pairs of stations were

selected 1868, for which $\nu < 0.04$. Note that the systematic error in λ for selected ν calculated for the exponential correlation function of a homogeneous field, on one order of magnitude smaller than the relative standard deviation $\sigma(\bar{r}_f)/\bar{r}_f = \gamma$ of the estimates of the correlation coefficients $\bar{r}_f$. In this case, for this purpose the sample size is 465 values.

To estimate the degree of inhomogeneity of the spatial indicator field of precipitation following numerical experiments were carried out. The block S_0 of matrix $S_{(n)}$ is the correlation matrix of the spatial indicator field of precipitation at the stations, which determines the corresponding correlation function $s(x_s, y_s; x_h, y_h)$. Within the framework of the given experiment it is necessary to compare the values λ and $\tilde{\lambda}$ for inhomogeneous and homogeneous fields, respectively. For this purpose, the real correlation function $s(x_s, y_s; x_h, y_h)$ of the indicator fields at the stations was approximated by the correlation function

$$s(x_l, y_l; x_h, y_h) = s(x_l - x_h, y_l - y_h) = s(x, y) = \exp(-[ax^2 + bxy + cy^2]^\theta) \qquad (4)$$

with a choice of parameters a, b, c and θ by minimizing the mean square difference between the real and the approximating functions. With the help of this function the correlation matrix $\tilde{S}_0$ of a homogeneous field at the stations was constructed. Further for the matrix S_0 and $\tilde{S}_0$ was built two ensembles of samples of 465 indicator fields. Each ensemble was contained 1000 elements. For each sample were estimated the corresponding correlation matrix. With the help of these matrixes values λ and $\tilde{\lambda}$ were calculated for each of the 1868 pairs and further on sample of volume 1868000 values were estimated probabilities $P(-\gamma \le \lambda < \gamma) = p_\lambda$ and $P(-\tilde{\gamma} \le \tilde{\lambda} < \tilde{\gamma}) = p_{\tilde{\lambda}}$. Results were showed that $p_\lambda = 43\%$, $p_{\tilde{\lambda}} = 55\%$. Thus, in 12% of 55% cases inequalities $-\gamma \le \lambda < \gamma$ are carried out because of inhomogeneity. Let's notice, that the experiment gives an estimation of a degree of heterogeneity only approximately, as the considered field is inhomogeneous and we have to compare it with the modeling homogeneous field with the correlation function (4).

3 Simulation of indicator fields of precipitation on a regular grid

For the construction of the spatio-temporal indicator fields of daily sums of precipitation an approach based on the simulation of joint time-series of indicators at the stations, discussed in the first section with the following stochastic interpolation values of the field from stations to the mesh point of the regular grid was used. Some of the algorithms of modeling of conditional Gaussian fields necessary for the stochastic interpolation of Gaussian fields are described, for example, in [7]. Approximate approaches to modeling the conditional non-Gaussian fields in this paper are based on the method of inverse distribution functions and algorithms for modeling conditional Gaussian fields.

In this paper, a simple algorithm for stochastic interpolation of indicators precipitation from stations to the grid points was used. It can be represented as the

following transformation. Let $\{\omega_i\}$ - a spatial indicator field of precipitation at the stations, where ω_i - the value of the field in the point with coordinates $\{x_i, y_i\}$, and $\{\omega_{ik}\}$, $i = 1, \ldots, n, \quad k = 1, \ldots, m$ - a field on the regular grid, where ω_{ik} - the value of the field in the mesh point with coordinates $\{x_{ik}, y_{ik}\}$. Consider the normalized weights $w_{lk,1}, \ldots, w_{lk,n}$ that are built by using the following equations

$$\bar{w}_{lk,i} = \frac{1}{\sqrt{(x_{lk} - x_i)^2 + (y_{lk} - y_i)^2}}, \quad w = \sum_{i=1}^{n} \bar{w}_{lk,i}, \quad w_{lk,i} = \frac{\bar{w}_{lk,i}}{w}.$$

Stochastic interpolation of field values at the stations $\{\omega_i\}$ to the mesh points was made by using the formula

$$\omega_{lk} = \omega_i, \tag{5}$$

where i- a random number of the station with the distribution $w_{lk,1}, \ldots, w_{lk,n}$. Usually in practical modeling do not use all the available stations, and only a few nearby stations. In this paper we used the 5 nearby stations.

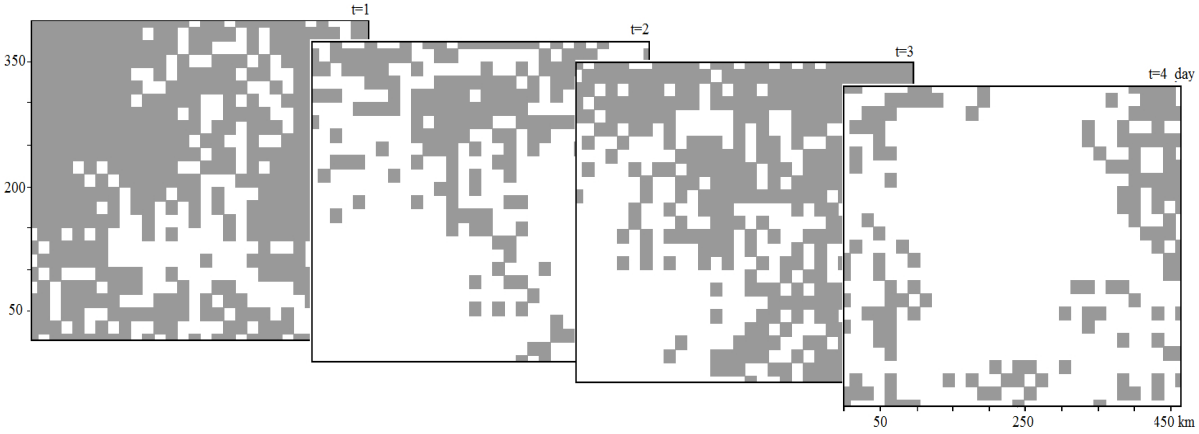

Figure 1: The example of a realization of the spatio-temporal indicator field of daily sums of liquid precipitation. Novosibirsk, May.

The example of realization of the model spatio-temporal indicator field of daily sums of liquid precipitation is presented in Fig.1. The field is built on the grid of 30×25 mesh points, time interval is equal to 1 month (spatial step is 16 km, time step - 1 day, the first 4 fields of 31 are shown in the figure). Selection of the step of grid depends on the application. Calculating the value p_λ for field $\{\omega_{ik}\}$ on the grid after interpolation (5) is showed that p_λ= 38%, i.e. the proportion of correlations satisfying $-\gamma \leq \lambda < \gamma$ due to inhomogeneity of the field increased by 5%. It can be explained, in particular, by the fact that the interpolation was performed with non-uniform network of stations, which introduces a systematic error in the correlations, increasing inhomogeneity.

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (11-01-00641, 12-01-00727-, 12-05-00169-)).

References

[1] Anisimova A.V.(1997). Numerical simulation of indicator random fields of liquid precipitation. *Proceedings of the Conference of Young Scientists, Inst. of Comp. Math. and Math. Geoph.* Novosibirsk, pp. 3-15 [in Russian].

[2] Bolshev L.N. and Smirnov N.V. (1983) *Tables of Mathematical Statistics*. Nauka, Moscow [in Russian].

[3] Drobyshev A.D., Marchenko A.S., Ogorodnikov V.A. and Chizhykov V.D.(1989). Statistical structure of time series for daily sums of liquid precipitations in the plane part of Novosibirsk region. *Proc. West Sib. Research Inst. Goskomgidromet. 86*,pp. 44-66 [in Russian].

[4] Kliber W., Katz R. W., Rajagopalan B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour. Res. 48*, W01523, doi:10.1029/2011WR011105.

[5] Marpl S.L. (1987). *Digital Spectral Analysis with Applications*. Prentice-Hall.

[6] Ogorodnikov V.A. and Prigarin S.M. (1996). *Numerical Modelling of Random Processes and Fields: Algorithms and Applications*. VSP. Utrecht. The Netherlands.

[7] Ogorodnikov V.A., Kargapolova N.A., Sereseva O.V. Numerical stochastic model of spatial fields of daily sums of liquid precipitation. (2013). Russian Journal of Numerical Analysis and Mathematical Modelling. Vol. 28, No. 2. pp.187-200.

[8] Ukhinova O.S., Ogorodnikov V.A.(2009). Stochastic models of spatial-time fields of precipitation sums. *Proc. of the 6^{th} St. Peterburg WorkShop on simulation*, pp.193-197.

A Power Comparison of Homogeneity Tests for Randomly Censored Data

Petr Philonenko[1] and Sergey Postovalov[1]
[1] *Novosibirsk State Technical University, Novosibirsk, Russia*
e-mail: petr-filonenko@mail.ru, postovalov@ngs.ru

Abstract

In this paper the behavior of homogeneity test power is presented for different cases (different censoring rates, different sample sizes and different alternative hypotheses). The simulation results show that the power of tests depends on the censoring distribution while survival functions are intersected. If the survival curves do not intersect, then the choice of censoring distribution does not affect on the power of the test. If the survival curves have intersection, the Bagdonavičus-Nikulin tests are the most powerful, but their power decreases while the censoring rate increases. If the survival curves are not intersected, the logrank test, the Cox-Mantel test, the Gehan test and the Peto test are more powerful than the Bagdonavičus-Nikulin test.

Keywords: randomly censored data, hypothesis of homogeneity, Peto test, Gehan test, logrank test, Cox-Mantel test, Q-test, Bagdonavičius-Nikulin test (single crossing), Bagdonavičius-Nikulin test (multiple crossing).

Introduction

In survival data analysis it is often impossible to observe every subject under study until the end, thus we deal with a *right censored data* . For such cases special statistical methods for comparison two survival curves are required, e.g. Gehan's Generalized Wilcoxon test, the Cox-Mantel test etc [5][8][9]. Recently a number of new homogeneity tests were developed. Bagdonavičius and Nikulin developed tests for homogeneity against crossing survival functions alternatives in the case of randomly censored data [1][2][3][4]. Martnez and Naranjo suggested pretest to choose more powerful test [7]. The purpose of this work is to compare the power of homogeneity tests for different alternative hypotheses, especially with a different number of intersections of survival function. For randomly right censored data we consider two different distribution law for censoring time and then we compare how this affects on the test power.

1 Homogeneity tests for right censored data

Suppose that we have two samples of continues random variables ξ_1 and ξ_2 respectively, $X_1 = \{t_{11}, t_{12}, ..., t_{1n_1}\}$ and $X_2 = \{t_{21}, t_{22}, ..., t_{2n_2}\}$, with survival distributions $S_1(t)$ and $S_2(t)$, and

$$t_{ij} = \min\left(T_{ij}, C_{ij}\right),$$

where T_{ij} and C_{ij} are the failure and censoring times for the jth object of the ith group. T_{ij} and C_{ij} are i.i.d. with CDF $F_i(t)$ and $F_i^C(t)$ respectively. Survival curve $S_i(t)$ means the probability of survival in the time interval $[0, t]$

$$S_i\left(t\right) = P\{\xi_i > t\} = 1 - F_i\left(t\right).$$

The main hypothesis is

$$H_0 : S_1(t) = S_2(t).$$

Further we will suppose that the elements of samples are ordered: $t_{11} < ... < t_{1n_1}$ and $t_{21} < ... < t_{2n_2}$. Also we pool these samples and we sort their elements by ascending order $T = X_1 \cup X_2 = \left\{t_1,\ t_2,\ \ldots,\ t_n\right\}$, where $t_1 < t_2 < ... < t_n,\ n = n_1 + n_2$.

Denote the sample indicator v_i and the censoring indicators c_{ij} and c_i as

$$v_i = \begin{cases} 0,\ if\ t_i \in X_1, \\ 1,\ if\ t_i \in X_2 \end{cases}$$

$$c_{ij} = \begin{cases} 1,\ if\ t_{ij} \text{ is censoring time}, \\ 0, if\ t_{ij} \text{ is failure time}; \end{cases}$$

$$c_i = \begin{cases} 1,\ if\ t_i \text{ is censoring time}, \\ 0, if\ t_i \text{ is failure time}. \end{cases}$$

Let us consider tests that is used for randomly right censored data. The following two criteria are based on the Wilcoxon test.

Gehan's generalized Wilcoxon test

The test statistic is [6, p.103]

$$S_G = \frac{\sum\limits_{i=1}^{n}(1 - v_i)h_i}{\sqrt{\frac{n_1 n_2}{n(n-1)}\sum\limits_{i=1}^{n}(1 - v_i)h_i^2}}, \tag{1}$$

where

$$h_i = \sum_{j=1}^{n} v_j h_{ij}, \quad h_{ij} = \begin{cases} +1,\ if\ t_i > t_j\ \&\ c_j = 0\ \&\ v_i = 0\ \&\ v_j = 1; \\ -1,\ if\ t_i < t_j\ \&\ c_i = 0\ \&\ v_i = 0\ \&\ v_j = 1\ ; \\ 0, \quad otherwise\ . \end{cases}$$

The null hypothesis is rejected with the significance level α, if $|S_G| > z_{1-\frac{\alpha}{2}}$, where $z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$-quantile of the standard normal distribution.

Peto and Peto's generalized Wilcoxon test

The test statistic is [6, p.116]

$$S_P = \frac{n(n-1)\sum\limits_{i=1}^{n} u_i(1 - v_i)}{n_1 n_2 \sum\limits_{i=1}^{n} u_i^2}, u_i = \begin{cases} \hat{s}(t_i) + \hat{s}(t_{k_i}) - 1,\ if\ c_i = 0 \\ \hat{s}(t_{k_i}) - 1,\ if\ c_i = 1, \end{cases} \tag{2}$$

where $k_i = \max\{j | j \in \{0, ..., i-1\}, c_j = 0\}, c_0 = 0$, and

$$\hat{S}(t) = \prod_{i:t_i<t, c_i=0} \frac{n-i}{n-i+1}$$

is the Kaplan-Meier estimator of $S(t)$ by the pooled sample T. The null hypothesis is rejected with the significance level α, if $|S_P| > z_{1-\frac{\alpha}{2}}$.

The following two criteria are based on the Logrank test

Logrank test

The test statistic is [6, p.111]

$$S_L = \frac{\sum_{i=1}^{n}(1-c_i) - \sum_{j=1}^{i}\frac{1}{n-j+1}}{\sqrt{\left[\sum_{i=1}^{n}(1-c_i)\frac{n-i}{n-i+1}\right]\frac{n_1 n_2}{n(n-1)}}}, \tag{3}$$

The null hypothesis is rejected with the significance level α, if $|S_L| > z_{1-\frac{\alpha}{2}}$.

Cox-Mantel test

The test statistic is [6, p.109]

$$S_{CM} = \frac{r_2 - \sum_{i=1}^{n}(1-c_i)A_{(i)}}{\sqrt{\sum_{i=1}^{n}(1-c_i)A_{(i)}\left(1-A_{(i)}\right)}}, \tag{4}$$

where

$$r_2 = \sum_{i=1}^{n} v_i(1-c_i),\ A_{(i)} = \frac{1}{n-i}\sum_{j=i}^{n} v_j.$$

The null hypothesis is rejected with the significance level α, if $|S_C M| > z_{1-\frac{\alpha}{2}}$.

Q-test

Martnez and Naranjo suggested pretest to choose what test is better to use: Wilcoxon type or logrank type [7]. The Q-test statistic is

$$S_Q = \begin{cases} S_L \text{ or } S_{CM}, \text{ if } Q < 0 \\ S_P \text{ or } S_G, \text{ otherwise.} \end{cases} \tag{5}$$

where $Q = \left[\hat{S}_2(q_{0.6}) - \hat{S}_1(q_{0.6})\right] - \left[\hat{S}_2(q_{0.2}) - \hat{S}_1(q_{0.2})\right]$, $q_p = \hat{S}_1^{-1}(p)$, and

$$\hat{S}_i(t) = \prod_{j:t_{ij}<t, c_{ij}=0} \frac{n-j}{n-j+1}$$

is the Kaplan-Meier estimator of $S_i(t)$. For unambiguity we use the pair S_L and S_P further.

The following two tests are developed against alternative with intersections of survival functions.

Bagdonavičius-Nikulin test 1 (single crossing)

The test statistic is [2]

$$S_{BN1} = (U_1, U_2)\,\Sigma^{-1}(U_1, U_2)^T, \tag{6}$$

where

$$U_1 = \sum_{j:c_{1j}=0} \frac{Y_2(t_{1j})}{Y(t_{1j})} - \sum_{j:c_{2j}=0} \frac{Y_1(t_{2j})}{Y(t_{2j})},$$

$$U_2 = -\sum_{j:c_{1j}=0} \frac{Y_2(t_{1j})}{Y(t_{1j})} \ln\left(1+\Lambda(t_{1j})\right) + \sum_{j:c_{2j}=0} \frac{Y_1(t_{2j})}{Y(t_{2j})} \ln\left(1+\Lambda(t_{2j})\right),$$

$$Y(t) = Y_1(t) + Y_2(t),\ Y_i(t) = \sum_{j=1}^{n_i} Y_{ij}(t),\ Y_{ij}(t) = 1_{\{t_{ij}\geq t\}}, \Lambda(t) = \sum_{i=1}^{2} \sum_{j:c_{ij}=0, t_{ij}\leq t} \frac{1}{Y(t_{ij})}.$$

The elements of the matrix Σ are

$$\sigma_{11} = \sum_{i=1}^{2} \sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})};\ \sigma_{22} = \sum_{i=1}^{2} \sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})} \ln^2(1+\Lambda(t_{ij})).$$

$$\sigma_{12} = \sigma_{21} = \sum_{i=1}^{2} \sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})} \ln(1+\Lambda(t_{ij}));$$

The H_0 is rejected with the significance level α if $S_{BN1} > \chi^2_{1-\alpha}(2)$, where $\chi^2_{1-\alpha}(2)$ is the $(1-\alpha)$-quantile of the χ^2-distribution with 2 degrees of freedom.

Bagdonavičius-Nikulin test 2 (multiple crossing)

The test statistic is [2]

$$S_{BN2} = (U_1, U_2, U_3)\,\Sigma^{-1}(U_1, U_2, U_3)^T, \tag{7}$$

where

$$U_1 = \sum_{j:c_{1j}=0} \frac{Y_2(t_{1j})}{Y(t_{1j})} - \sum_{j:c_{2j}=0} \frac{Y_1(t_{2j})}{Y(t_{2j})},\ U_2 = \sum_{j:c_{1j}=0} \frac{Y_2(t_{1j})}{Y(t_{1j})}\Lambda(t_{1j}) - \sum_{j:c_{2j}=0} \frac{Y_1(t_{2j})}{Y(t_{2j})}\Lambda(t_{2j}),$$

$$U_3 = \sum_{j:c_{1j}=0} \frac{Y_2(t_{1j})}{Y(t_{1j})}\Lambda^2(t_{1j}) - \sum_{j:c_{2j}=0} \frac{Y_1(t_{2j})}{Y(t_{2j})}\Lambda^2(t_{2j}),$$

The elements of the matrix Σ are

$$\sigma_{11} = \sum_{i=1}^{2} \sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})};\ \sigma_{12} = \sigma_{21} = \sum_{i=1}^{2} \sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})}\Lambda(t_{ij})$$

$$\sigma_{13} = \sigma_{31} = \sigma_{22} = \sum_{i=1}^{2} \sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})} \Lambda^2(t_{ij})$$

$$\sigma_{23} = \sigma_{32} = \sum_{i=1}^{2} \sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})} \Lambda^3(t_{ij});\ \sigma_{33} = \sum_{i=1}^{2} \sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})} \Lambda^4(t_{ij})$$

The H_0 is rejected with the significance level α if $S_{BN2} > \chi^2_{1-\alpha}(3)$, where $\chi^2_{1-\alpha}(3)$ is the $(1-\alpha)$-quantile of the χ^2-distribution with 3 degrees of freedom.

2 Alternative hypotheses

Since there are a lot of homogeneity tests, a researcher wants to know what the test should be used in practice. Unfortunately the most powerful test is not exist in general case. Looking at empirical survival functions (or Kaplan-Meier estimators for randomly censored data) we could choose what kind of alternative we will have. In general there are two main situation of survival curves differences: whether they have intersections or not, and where the difference is lager, for early or for late times. Therefore we construct 10 different alternative hypotheses.

The alternative hypotheses are shown in Table 1 and in Figure 1, they are based on Exponential (8), Gamma (9), Lognormal (10), Weibull (11), and Generalized Weibull (12) families of distribution laws.

$$f_{Exp}(t;\theta_1) = \theta_1 e^{-\theta_1 t},\ t \geq 0 \tag{8}$$

$$f_{\Gamma}(t;\theta_1,\theta_2) = \frac{\left(\frac{t}{\theta_1}\right)^{\theta_2 - 1} e^{-\frac{t}{\theta_1}}}{\theta_1 \Gamma(\theta_2)},\ t \geq 0 \tag{9}$$

$$f_{LgN}(t;\theta_1,\theta_2) = \frac{1}{t\sqrt{2\pi\theta_2^2}} e^{-\frac{(\ln t - \theta_1)^2}{2\theta_2^2}},\ t \geq 0 \tag{10}$$

$$f_{We}(t;\theta_1,\theta_2) = \frac{\theta_2 t^{\theta_2 - 1} e^{-\left(\frac{t}{\theta_1}\right)^{\theta_2}}}{{\theta_1}^{\theta_2}},\ t \geq 0 \tag{11}$$

$$f_{GWe}(t;\theta_1,\theta_2) = \frac{\frac{\theta_1}{\theta_2}\left(1 + t^{\theta_1}\right)^{\frac{1}{\theta_2} - 1} e^{1-(1+t^{\theta_1})^{\frac{1}{\theta_2}}}}{t^{1-\theta_1}},\ t \geq 0 \tag{12}$$

3 Power simulation

The test power was calculated by 83000 repetitions (five series by 16600 repetitions to calculate confidence interval). For all tests both samples X_1 and X_2 had an equal size. The significance level is equal 0.05.

Table 1: Alternative hypotheses

H_i	$\mathbf{1}^{st}$ **sample**	$\mathbf{2}^{nd}$ **sample**	**Number of intersections**	**Intersections**
H_1	We(2;2)	Γ(0.56; 3.12)	2	$\approx$0.7 and $\approx$2.8
H_2	We(2;2)	LgN(0.41; 0.62)	2	$\approx$0.75 and $\approx$2.5
H_3	Exp(1)	Exp(1.1)	0	-
H_4	LgN(0;1)	LgN(0.1; 1.0)	0	-
H_5	GWe(0.95;0.85)	Exp(1)	0	-
H_6	We(1.1; 1.4)	Γ(0.9; 1.0)	1	$\approx$1.81
H_7	We(1.0; 1.1)	We(1.1; 0.9)	1	$\approx$0.65
H_8	LgN(0.1;1.0)	LgN(0; 1.2)	1	$\approx$1.82
H_9	We(1.0; 1.1)	We(0.7; 0.9)	1	$\approx$4.97
H_{10}	Γ(1; 1)	LgN (0; 0.65)	2	$\approx$2.28 and $\approx$5.1

In Tables 4, 5, 6 the power of 7 tests, including Gehan's generalized Wilcoxon test ('Gehan'), Peto and Peto's generalized Wilcoxon test ('Peto'), the logrank test (Lg), the Cox-Mantel test (CM), the Q-test (Q), the Bagdonavičius-Nikulin tests (single crossing – BN_1, multiple crossing – BN_2), are presented for sample sizes $n_1 = n_2 = 200$ with censoring rates 10%, 20%, 30%, 40%, and 50%. We investigated how the censoring rate and censoring distribution affect on the power of tests.

a How does the distribution of censoring time affect on the power of the test?

We consider two ways of modeling censoring times using Weibull distribution law (Table 2) or Gamma distribution law (Table 3). The results are shown in Tables 4, 5, 6 for $n_1 = n_2 = 200$.

When survival curves are not crossed (see Tables 6), test powers with different F^C are almost equal. The differences between power values can be explained by the simulation error. However, when survival curves are crossed, test powers are significantly different for different distributions of censoring times (see Tables 4 and 5). The most dependence is appeared for the Bagdonavičius-Nikulin tests, because these tests use information about possibility of the intersections and early censoring times can hide this information.

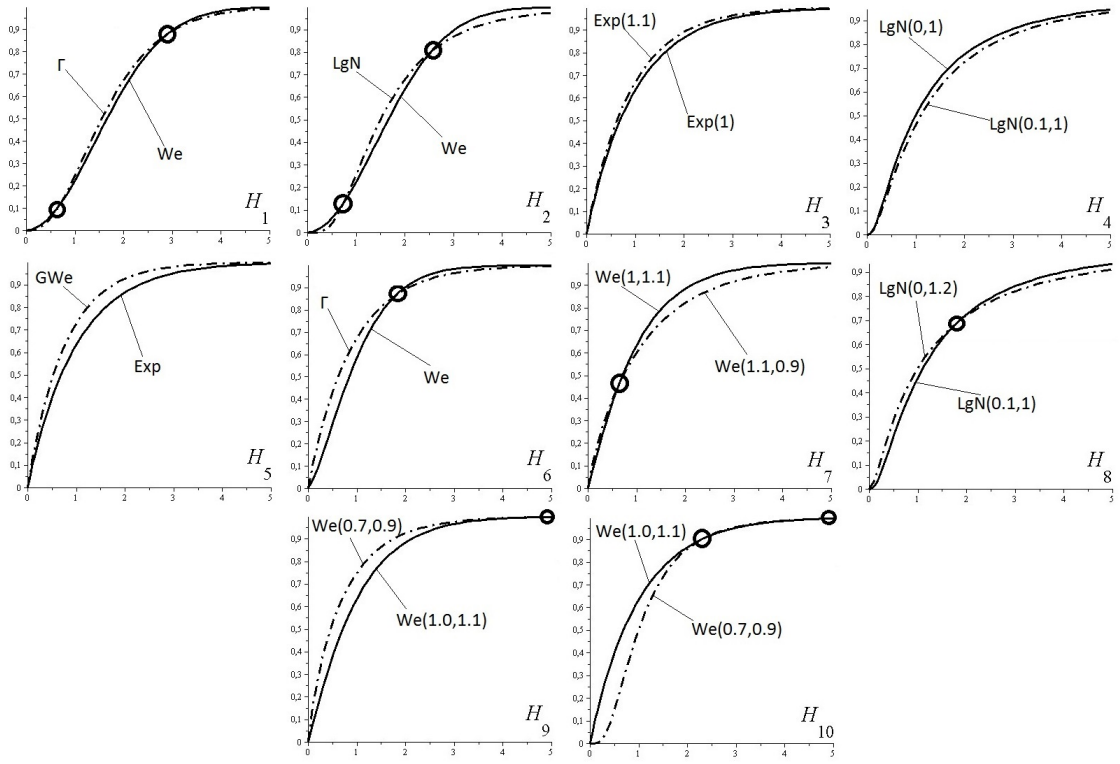

Figure 1: CDFs of alternative hypotheses $H_1 - H_{10}$. The circle means the CDFs crossing

b How does the censoring rate affect on the power of the test?

Table 7 gives the power of tests against alternatives $H_1 - H_{10}$ for censoring rates $0\% - 50\%$ and the Weibull distribution of censoring times. We found that the power of tests can increase or decrease while the censoring rate increses.

When survival curves are not crossing, the most powerful tests are the Logrank and the Cox-Mantel against alternatives H_3, H_5, or Gehan's generalized Wilcoxon test and Peto and Peto's generalized Wilcoxon test against alternative H_4. The lowest tests power are both tests of Bagdonavičius-Nikulin in this case.

The alternative hypothesis H_9 has one intersection on the right tail of survival curves. In this case Gehan's generalized Wilcoxon test, Peto and Peto's generalized Wilcoxon are the most powerful tests.

For others alternatives H_6, H_7, H_8 with one intersection of survival curves the most powerful test is the Bagdonavičius-Nikulin test BN1. When survival curves have two intersections (alternatives H_1, H_2), the most powerful test is the Bagdonavičius-Nikulin test BN2. If survival curves have late intersection (alternative H_{10}) four criteria (Gehan's generalized Wilcoxon test, Peto and Peto's generalized Wilcoxon and both the Bagdonavičius-Nikulin tests) are high power.

Table 2: First way of modeling censoring time using Weibull distribution as $F_i^C(t)$

H	i	10%	20%	30%	40%	50%
H_1	1	We$(3.44, 6.88)$	We$(2.87, 5.74)$	We$(2.48, 4.96)$	We$(2.16, 4.32)$	We$(1.87, 3.74)$
H_1	2	We$(4.00, 4.00)$	We$(4.00, 2)$	We$(4.00, 1.25)$	We$(2.25, 2.10)$	We$(3.74, 0.40)$
H_2	1	We$(3.44, 6.88)$	We$(2.87, 5.74)$	We$(2.48, 4.96)$	We$(2.16, 4.32)$	We$(1.87, 3.74)$
H_2	2	We$(4.16, 4.20)$	We$(3.35, 3.11)$	We$(3.00, 2.49)$	We$(2.35, 2.10)$	We$(3.47, 0.45)$
H_3	1	We$(3.4, 2.47)$	We$(2.00, 3.25)$	We$(2, 1.25)$	We$(1.00, 6.50)$	We$(1.00, 1.00)$
H_3	2	We$(3.11, 2.47)$	We$(2.00, 2.14)$	We$(2.00, 1.09)$	We$(1.00, 3.00)$	We$(1.00, 0.75)$

Table 3: Second way of modeling censoring time using Gamma distribution as $F_i^C(t)$

H	i	10%	20%	30%	40%	50%
H_1	1	$\Gamma(2.00, 3.555)$	$\Gamma(2.50, 1.580)$	$\Gamma(2.50, 1.185)$	$\Gamma(2.65, 0.869)$	$\Gamma((2.73, 0.678)$
H_1	2	$\Gamma(2.9, 1.870)$	$\Gamma(2.6, 1.460)$	$\Gamma(2.5, 1.150)$	$\Gamma(2.2, 1.050)$	$\Gamma(2.0, 0.927)$
H_2	1	$\Gamma(2.00, 3.555)$	$\Gamma(2.50, 1.580)$	$\Gamma(2.50, 1.185)$	$\Gamma(2.65, 0.869)$	$\Gamma((2.73, 0.678)$
H_2	2	$\Gamma(2.1, 3.35)$	$\Gamma(3.2, 1.13)$	$\Gamma(2.0, 1.59)$	$\Gamma(2.2, 1.06)$	$\Gamma(4.0, 0.42)$
H_3	1	$\Gamma(2.3, 1.71)$	$\Gamma(3.7, 0.55)$	$\Gamma(3.5, 0.41)$	$\Gamma(3.3, 0.32)$	$\Gamma(3.7, 0.209)$
H_3	2	$\Gamma(7.71, 0.316)$	$\Gamma(7.00, 0.235)$	$\Gamma(7.20, 0.166)$	$\Gamma(7.60, 0.117)$	$\Gamma(7.70, 0.086)$

Conclusions

Let us summarize received results.In the case of censoring data test power depends on distribution of censoring time, if survival curves are crossed, especially for the Bagdonavičius-Nikulin tests. In contrast, if survival curves are not crossed, the choice of censoring time distribution is not so important. If survival curves have intersections, then the Bagdonavičius-Nikulin tests have the most power, but it decreases sharply while censoring rate increases. If survival curves do not have intersections, then the Logrank test, the Cox-Mantel test, Gehan's generalized Wilcoxon test and Peto & Peto's generalized Wilcoxon test are more powerful than the Bagdonavičius-Nikulin tests.

Acknowledgements

The authors are grateful to Vilijandas Bagdonavičius and Mikhail Nikulin for their valuable advices in an interpretation of the achieved results.

Table 4: The power of tests against H_1

Test	Censoring times from Weibull						Censoring times from Gamma					
	0%	10%	20%	30%	40%	50%	0%	10%	20%	30%	40%	50%
Gehan	0.111	0.121	0.121	0.110	0.105	0.080	0.111	0.117	0.116	0.103	0.092	0.080
Peto	0.106	0.116	0.116	0.111	0.113	0.088	0.106	0.110	0.108	0.104	0.102	0.093
Lg	0.053	0.072	0.083	0.090	0.106	0.082	0.053	0.062	0.069	0.074	0.085	0.086
CM	0.054	0.073	0.090	0.103	0.113	0.096	0.054	0.061	0.070	0.082	0.085	0.086
Q-test	0.062	0.082	0.094	0.098	0.102	0.082	0.062	0.071	0.077	0.085	0.091	0.088
BN1	0.184	0.130	0.105	0.098	0.096	0.086	0.184	0.147	0.123	0.103	0.091	0.080
BN2	0.186	0.134	0.123	0.114	0.106	0.107	0.186	0.150	0.122	0.100	0.097	0.090

Table 5: The power of tests against H_2

Test	Censoring times from Weibull						Censoring times from Gamma					
	0%	10%	20%	30%	40%	50%	0%	10%	20%	30%	40%	50%
Gehan	0.120	0.139	0.160	0.173	0.161	0.106	0.120	0.128	0.138	0.132	0.121	0.105
Peto	0.120	0.132	0.147	0.161	0.157	0.121	0.120	0.124	0.136	0.129	0.127	0.132
Lg	0.061	0.049	0.076	0.103	0.132	0.102	0.061	0.046	0.050	0.058	0.074	0.107
CM	0.070	0.049	0.076	0.108	0.138	0.126	0.070	0.050	0.050	0.058	0.080	0.111
Q-test	0.050	0.052	0.092	0.120	0.131	0.102	0.050	0.043	0.055	0.064	0.084	0.122
BN1	0.729	0.378	0.236	0.170	0.133	0.109	0.729	0.562	0.375	0.271	0.171	0.111
BN2	0.749	0.412	0.292	0.238	0.199	0.192	0.749	0.604	0.405	0.323	0.223	0.130

References

[1] Bagdonavičius V.B., Levuliene R.J., Nikulin M.S., Zdorova-Cheminade O. (2004). Tests for equality of survival distributions against non-location alternatives *Lifetime Data Analysis*. Vol. **10**, pp. 445–460.

[2] Bagdonavičius V.B., Nikulin M. (2006). On goodness-of-fit tests for homogeneity and proportional hazards *Applied Stochastic Models in Business and Industry*. Vol. **22**, pp. 607–619.

[3] Bagdonavičius, Malov, Nikulin (2008). Testing ofthe homogeneity of marginal distrinbution in copula models *Comptes Rendus de l'Academie de Paris*. Vol. **346**, pp. 445-450.

[4] Bagdonavičius, Malov, Nikulin (2012). Homogeneity Tests for Related Samples Under Censoring *Communications in statistics*. Vol. **41, 6**, pp. 764–775.

[5] Gehan E.A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples *Biometrika*. Vol. **52**, pp. 203–223.

[6] Lee E.T., Wang J.W. (2003). *Statistical Methods for Survival Data Analysis*. Wiley, Hoboken.

[7] Martinez Ruvie Lou Maria C., Naranjo Joshua D. (2010). A pretest for choosing between logrank and wilcoxon tests in the two-sample problem *International Journal of Statistics*. Vol. **LXVIII**, pp. 111–125.

Table 6: The power of tests against H_3

Test	Censoring times from Weibull						Censoring times from Gamma					
	0%	10%	20%	30%	40%	50%	0%	10%	20%	30%	40%	50%
Gehan	0.128	0.121	0.122	0.104	0.105	0.086	0.128	0.122	0.121	0.110	0.098	0.093
Peto	0.135	0.131	0.120	0.116	0.108	0.091	0.135	0.129	0.123	0.120	0.107	0.094
Lg	0.159	0.136	0.132	0.121	0.107	0.103	0.159	0.136	0.131	0.122	0.110	0.098
CM	0.160	0.142	0.133	0.126	0.114	0.100	0.160	0.141	0.134	0.123	0.110	0.101
Q-test	0.145	0.136	0.134	0.122	0.110	0.103	0.145	0.142	0.132	0.127	0.112	0.100
BN1	0.119	0.117	0.107	0.100	0.095	0.085	0.119	0.113	0.110	0.100	0.091	0.082
BN2	0.107	0.101	0.097	0.091	0.092	0.078	0.107	0.096	0.095	0.090	0.091	0.085

[8] Mantel N. (1967). Evaluation of survival data and two new rank order statistics arising in its consideration *Cancer Chemotherapy Rep.*. Vol. **50**, pp. 163-170.

[9] Peto R., Peto J. (1972). Asymptotically efficient rank invariant test procedures *Journal of the Royal Statistical Society, Series A (General)*. Vol. **135**, pp. 303-307.

Table 7: The power of tests against H_1–H_{10}

	0%	10%	20%	30%	40%	50%	0%	10%	20%	30%	40%	50%
Test	H_1						H_6					
Gehan	0.105	0.118	0.116	0.110	0.107	0.081	0.799	0.826	0.869	0.886	0.907	0.90
Peto	0.106	0.116	0.116	0.111	0.113	0.089	0.800	0.814	0.841	0.835	0.852	0.846
Lg	0.054	0.073	0.086	0.092	0.107	0.084	0.176	0.306	0.426	0.441	0.631	0.658
CM	0.054	0.071	0.087	0.096	0.102	0.093	0.180	0.297	0.420	0.457	0.623	0.672
Q-test	0.058	0.085	0.094	0.100	0.109	0.084	0.213	0.357	0.484	0.516	0.669	0.717
BN1	0.183	0.130	0.109	0.098	0.093	0.087	0.944	0.926	0.915	0.904	0.886	0.882
BN2	0.181	0.134	0.120	0.116	0.110	0.104	0.896	0.868	0.864	0.857	0.829	0.842
Test	H_2						H_7					
Gehan	0.120	0.139	0.160	0.168	0.159	0.109	0.066	0.057	0.051	0.053	0.064	0.088
Peto	0.120	0.132	0.147	0.161	0.157	0.121	0.066	0.059	0.053	0.050	0.053	0.061
Lg	0.065	0.049	0.076	0.105	0.130	0.100	0.372	0.222	0.157	0.087	0.044	0.052
CM	0.067	0.051	0.078	0.101	0.139	0.122	0.378	0.231	0.165	0.093	0.054	0.053
Q-test	0.051	0.054	0.083	0.119	0.141	0.113	0.312	0.182	0.133	0.078	0.059	0.050
BN1	0.728	0.397	0.238	0.168	0.134	0.108	0.724	0.548	0.472	0.365	0.262	0.259
BN2	0.743	0.423	0.297	0.242	0.208	0.202	0.603	0.493	0.396	0.307	0.238	0.203
Test	H_3						H_8					
Gehan	0.135	0.129	0.118	0.108	0.105	0.084	0.146	0.157	0.179	0.224	0.269	0.278
Peto	0.135	0.131	0.120	0.116	0.108	0.091	0.147	0.154	0.169	0.184	0.211	0.242
Lg	0.158	0.147	0.133	0.124	0.109	0.100	0.052	0.049	0.059	0.074	0.097	0.140
CM	0.155	0.149	0.135	0.126	0.115	0.098	0.055	0.050	0.057	0.067	0.095	0.184
Q-test	0.150	0.144	0.130	0.125	0.114	0.102	0.051	0.052	0.066	0.082	0.110	0.269
BN1	0.126	0.113	0.109	0.094	0.091	0.086	0.433	0.418	0.394	0.367	0.356	0.373
BN2	0.105	0.101	0.100	0.088	0.090	0.078	0.331	0.363	0.335	0.323	0.314	0.395
Test	H_4						H_9					
Gehan	0.169	0.165	0.168	0.150	0.141	0.147	0.940	0.941	0.940	0.940	0.919	0.889
Peto	0.168	0.165	0.168	0.149	0.140	0.144	0.941	0.940	0.938	0.934	0.907	0.869
Lg	0.146	0.150	0.151	0.140	0.128	0.126	0.771	0.829	0.853	0.843	0.815	0.766
CM	0.153	0.153	0.151	0.143	0.131	0.136	0.770	0.833	0.868	0.853	0.848	0.779
Q-test	0.158	0.157	0.156	0.144	0.134	0.133	0.816	0.868	0.893	0.883	0.874	0.782
BN1	0.123	0.129	0.123	0.118	0.111	0.113	0.891	0.892	0.896	0.884	0.865	0.807
BN2	0.108	0.111	0.109	0.107	0.097	0.121	0.830	0.856	0.859	0.843	0.825	0.744
Test	H_5						H_{10}					
Gehan	0.510	0.472	0.449	0.392	0.334	0.329	0.995	0.995	0.996	1.000	1.000	1.000
Peto	0.511	0.484	0.462	0.422	0.343	0.341	0.995	0.994	0.995	0.998	0.999	0.998
Lg	0.641	0.566	0.510	0.471	0.292	0.334	0.586	0.673	0.784	0.851	0.975	0.957
CM	0.637	0.570	0.509	0.466	0.359	0.355	0.591	0.670	0.780	0.856	0.978	0.958
Q-test	0.575	0.521	0.493	0.462	0.364	0.353	0.992	0.993	0.995	0.997	0.998	0.998
BN1	0.523	0.454	0.417	0.364	0.261	0.283	0.996	0.999	1.000	1.000	1.000	1.000
BN2	0.435	0.384	0.360	0.316	0.206	0.247	0.997	0.999	1.000	1.000	1.000	1.000

Optimal Discrete Two-Stage Study Design for Genome-Wide Association Studies

Sergey Postovalov[1], Andreas Ziegler[2] and Eugenia Konomanina[1]
[1] *Novosibirsk State Technical University, Novosibirsk, Russia*
[2] *Institute of Medical Biometry and Statistics, Luebeck, Germany*
e-mail: postovalov@ngs.ru

Abstract

Genome-wide Association Studies (GWAS) require large phenotyping and genotyping costs. Two-stage design can be efficient to reduce genotyping costs: on the first stage some disease associated SNP are detected and these associations are checked on the second stage with reliable significance level. This procedure decreases the number of genotyped SNP on the second stage, thus the genotyping costs will be less than genotyping costs of one-stage design. Modern genotyping technologies allow using 96 and 384 well plates. Thus the number of individuals should be proportional to well plate size. Monte Carlo simulation was used to find optimal number of well plates and critical values on the first and second stages. We also found that the costs have inverse relationship to Kullback-Leibler divergence between cases and controls distributions under alternative hypothesis.

Keywords: Sequential study design, genetic association, robust statistical test, genetic model selection, computer simulation.

Introduction

Two-stage case-control designs were considered for the epidemiologic purpose by White [1] and have been developed by Breslow and colleagues [2]. Approach of minimization costs in two-stage design was proposed by Elston et al. [3] for linkage analysis. Later this approach was transferred to association analysis by Satagopan et al. [12][13][14]. Optimization of the design consists in choosing the proportion of samples between two stages and critical values in such a manner as to minimize the total cost for specified genome-wide significance level and power [9][15][21][5] [8][16][10].

The start point of present work was a paper of Nguyen et al [10], where an optimal robust two-stage design using the MAX3 test were considered. In this paper we improved their results. First we found an optimal design for considered genetic models more accurately. Then we discovered the relationship between optimal sample size (and costs) and the Kullback-Leibler divergence between cases and controls distributions under alternative hypothesis. Also we considered the discrete design when the sample size on the first stage is proportional to a well plate size.

1 Statistical hypotheses

Suppose that there are three genotypes AA, Aa, and aa for SNP with two alleles, A and a. Let us there are r cases and s controls in the study. In the r cases, there

are r_0, r_1 and r_2 affected people with genotypes AA, Aa, and aa respectively. In the s controls, there are s_0, s_1 and s_2 affected people with genotypes AA, Aa, and aa respectively (Table 1).

Table 1: Case-control association studies

Genotype	AA	Aa	aa	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

The null hypothesis H_0 is that there are not an association between the disease and the genotyped SNP. There are three common alternative hypotheses that relies on *mode of inheritance* (MOI): H_r is recessive, H_d is dominant, and H_a is additive MOI. Let us $p_a = P(a)$ is a population frequency of allele a, $p_i = P\{$ number of alleles a is equal $i \mid$ cases $\}$ and $q_i = P\{$ number of alleles a is equal $i \mid$ controls$\}$; $f_i = P\{$ cases $\mid$ number of alleles a is equal i $\}$ are the penetrances, $K = P\{cases\}$ is a population prevalence of disease. The alternative hypothesis is defined by MOI, probability of risk allele p_a, and *the odds ratio* (OR)

$$\psi_i = \frac{f_i(1-f_0)}{f_0(1-f_i)} = \frac{p_i \cdot q_0}{p_0 \cdot q_i},\ i = 1, 2 \tag{1}$$

or *the genetic relative risk* (GRR)

$$\gamma_i = \frac{f_i}{f_0},\ i = 1, 2. \tag{2}$$

The differences between the distributions in the cases and controls groups (P and Q respectively) can be calculated by Kullback-Leibler divergence:

$$D_{KL}(P, Q) = \sum_{i=0}^{2} p_i \ln \frac{p_i}{q_i} \tag{3}$$

or symmetric Kullback-Leibler divergence:

$$\rho_{KL}(P, Q) = D_{KL}(P, Q) + D_{KL}(Q, P) = \sum_{i=0}^{2} (p_i - q_i) \ln \frac{p_i}{q_i} \tag{4}$$

If the H_0 is true then $\rho_{KL} = 0$. Let us Hardy-Weinberg equilibrium (HWE) holds. The Kullback-Leibler divergence is related with p_a, K, ψ_1, and ψ_2 as

$$D_{KL} = \ln \frac{1-K}{\lambda} + \frac{2p_a(1-p_a)\psi_1}{\lambda + K\psi_1} \ln \psi_1 + \frac{p_a^2 \psi_2}{\lambda + K\psi_2} \ln \psi_2, \tag{5}$$

Table 2: The symmetric Kullback-Leibler divergences for different alternative hypotheses

p_a	MOI	ψ_1	ψ_2	ρ_{KL}	MOI	ψ_1	ψ_2	ρ_{KL}	MOI	ψ_1	ψ_2	ρ_{KL}
0.1	R	1.00	1.25	0.00054	A	1.25	1.56	0.00970	D	1.25	1.25	0.00814
0.1	R	1.00	1.50	0.00195	A	1.50	2.25	0.03409	D	1.50	1.50	0.02814
0.1	R	1.00	1.75	0.00398	A	1.75	3.06	0.06831	D	1.75	1.75	0.05558
0.1	R	1.00	2.00	0.00648	A	2.00	4.00	0.10925	D	2.00	2.00	0.08780
0.3	R	1.00	1.25	0.00442	A	1.25	1.56	0.02169	D	1.25	1.25	0.01237
0.3	R	1.00	1.50	0.01560	A	1.50	2.25	0.07339	D	1.50	1.50	0.04046
0.3	R	1.00	1.75	0.03136	A	1.75	3.06	0.14208	D	1.75	1.75	0.07618
0.3	R	1.00	2.00	0.05035	A	2.00	4.00	0.22033	D	2.00	2.00	0.11539
0.5	R	1.00	1.25	0.00979	A	1.25	1.56	0.02479	D	1.25	1.25	0.00886
0.5	R	1.00	1.50	0.03348	A	1.50	2.25	0.08110	D	1.50	1.50	0.02793
0.5	R	1.00	1.75	0.06547	A	1.75	3.06	0.15266	D	1.75	1.75	0.05108
0.5	R	1.00	2.00	0.10250	A	2.00	4.00	0.23116	D	2.00	2.00	0.07557
0.9	R	1.00	1.25	0.00719	A	1.25	1.56	0.00826	D	1.25	1.25	0.00045
0.9	R	1.00	1.50	0.02248	A	1.50	2.25	0.02552	D	1.50	1.50	0.00136
0.9	R	1.00	1.75	0.04084	A	1.75	3.06	0.04594	D	1.75	1.75	0.00244
0.9	R	1.00	2.00	0.06010	A	2.00	4.00	0.06711	D	2.00	2.00	0.00354

where λ can be found from equation

$$\frac{(1-p_a)^2}{\lambda+K}+\frac{2p_a(1-p_a)\psi_1}{\lambda+K\psi_1}+\frac{p_a^2\psi_2}{\lambda+k\psi_2}=1. \tag{6}$$

The Kullback-Leibler divergence is related with p_a, K, γ_1 , and γ_2 as

$$\rho_{KL}=\frac{(1-p_a)^2(1-\tau)\ln\frac{1-K}{\tau-K}+2p_a(1-p_a)(\gamma_1-\tau)\ln\frac{\gamma_1(1-K)}{\tau-\gamma_1K}+p_a^2\cdot(\gamma_2-\tau)\ln\frac{\gamma_2(1-K)}{\tau-\gamma_2K}}{\tau(1-K)}, \tag{7}$$

where

$$\tau=(1-p_a)^2+2p_a(1-p_a)\gamma_1+p_a^2\gamma_2. \tag{8}$$

In the Table 2 the symmetric Kullback-Leibler divergences are calculated for different alternative hypotheses. The symmetric Kullback-Leibler divergence is better mesure of association than OR, because the same OR shows different dependence for different p_a .

2 Association tests

a The Cochran-Armitage trend test

The Cochran-Armitage trend test (CATT) statistic can be defined as [11]

$$T_{CATT} = \frac{\sum_{i=0}^{2} \omega_i \left(sr_i - rs_i\right)}{\left\{ rs \left(\sum_{i=0}^{2} \omega_i^2 n_i - \frac{1}{n} \left(\sum_{i=0}^{2} \omega_i n_i \right)^2 \right) \right\}^{1/2}}.$$

For the recessive MOI the test statistic T_r use score vector $\omega = (0, 0, 1)$, for the additive MOI the test statistic T_a use score vector $\omega = (0, 1, 2)$ and for dominant MOI the test statistic T_d use score vector $\omega = (0, 1, 1)$. The CATT statistic has standard normal distribution under null hypothesis and large sample size. The null hypothesis is rejected for big absolute value of T. The CATT utilizes a set of scores that can be obtained as an efficient score test for a logistic regression [19]. The statistical properties of the optimal test for the additive model were investigated by [17]. It is known that CATT test has substantial loss of power when optimal scores for one model are used, but the data follow a different model. We took the table 2 and table 3 from [4]. If the model is corresponded the optimal scores, then $n\rho_{KL}$ is almost a constant. But if the model is misspecified, then $n\rho_{KL}$ is different for different n. This is demonstrated in the Figure 1. The slope of curve $n(\rho_{KL})$ is higher when model is misspesified.

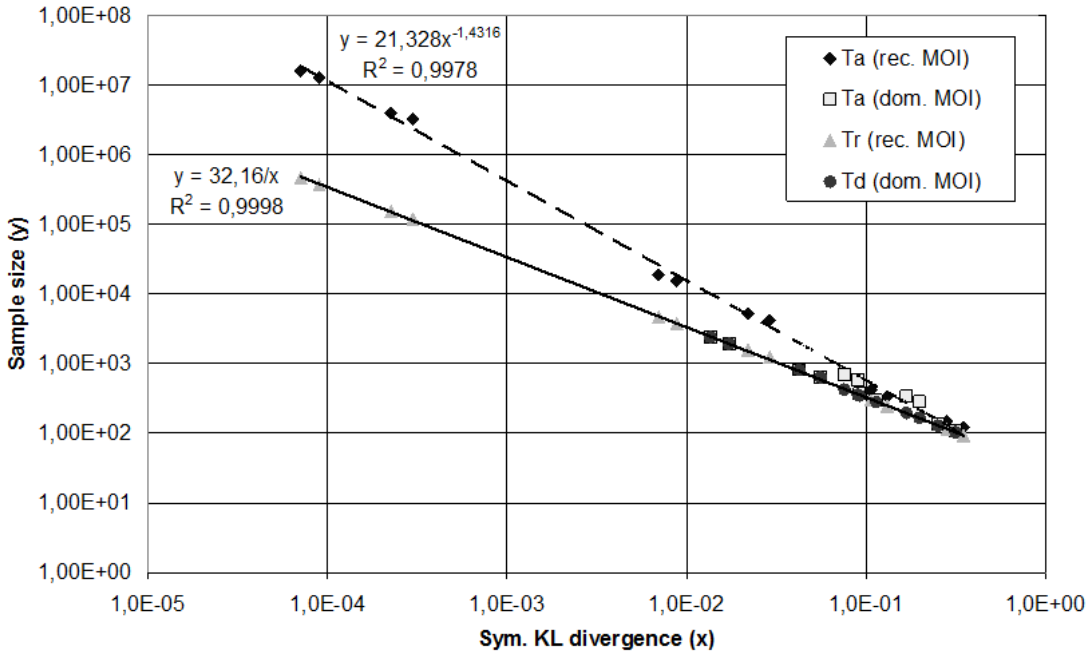

Figure 1: Relationship between required sample size and the symmetric KL-divergence for the Cochran-Armitage trend test with different score vectors ($\alpha = 0.05$ and $\beta = 0.2$)

b The MAX3 test

For robust analysis, when the MOI is misspecified, the Maximum test ($MAX3$) statistic is used [4]

$$T_{\max} = \max\left(|T_r|, |T_a|, |T_d|\right) \tag{9}$$

The optimal sample sizes for $MAX3$ test and different genetic models, when $\alpha = 0.05/m$ ($m = 610000$ is the number of markers), were calculated by Nguyen et al. [10]. We took the data from the tables 1-3 [10] and calculated ρ_{KL}. The relationship between optimal sample size and ρ_{KL} is shown in Figure 2.

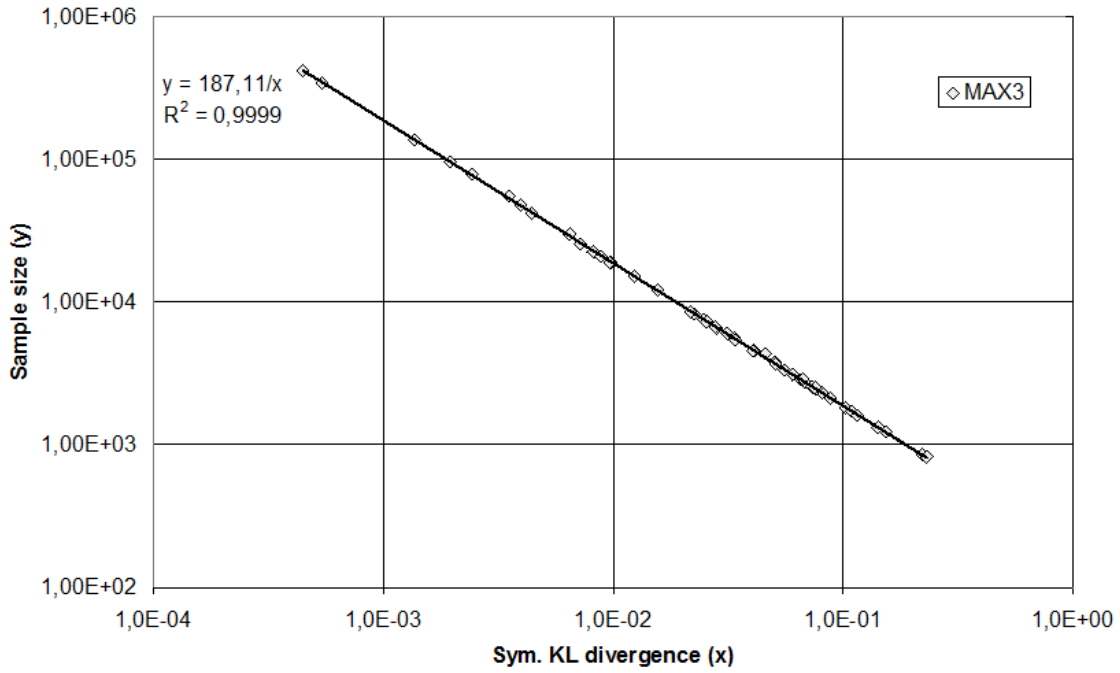

Figure 2: Relationship between required sample size and the symmetric KL-divergence for the $MAX3$ test ($\alpha = 0.05/m$ and $\beta = 0.1$)

3 Relative Efficiency

The association tests can be compared by required sample size. From [22] we can suppose that the required sample size is proportional to

$$\vartheta(\alpha, \beta) = \left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \Phi^{-1}\left(1 - \beta\right)\right)^2, \tag{10}$$

where α and β are the probabilities of first and second type respectively.

Then the relationship between required sample size and the symmetric Kullback-Liebler divergence is

$$n_{CATT} = \frac{4.035 \cdot \vartheta(\alpha, \beta)}{\rho_{KL}}, \tag{11}$$

for the Cochran-Armitage trend test and

$$n_{MAX3} = \frac{4.238 \cdot \vartheta(\alpha, \beta)}{\rho_{KL}}. \tag{12}$$

for $MAX3$ test.

Therefore relative efficiency of $MAX3$ test in compare of he Cochran-Armitage trend test with an optimal score vector is equal 0.95, so $MAX3$ test require in average 5% more observation than the optimal test.

4 Optimal design

The purpose of optimal design is the search optimal experiment parameters for what costs are minimized. For 1-stage design it is equivalent to minimize sample size for specified the errors type I and type II. Optimization of the 2-stage design consists in choosing the proportion of samples between the two stages and critical values in such a manner as to minimize the total cost for specified genome-wide significance level and power. Overall costs of 2-stage design can be calculated as

$$Costs = n\left(C_R + \pi m C_{G_1} + (1-\pi)\left((m-d)\alpha_1 + d(1-\beta_1)\right) C_{G_2}\right), \tag{13}$$

where C_R is costs of phenotyping one person, C_{G_1} is costs of genotyping one marker on the first stage and C_{G_2} on the second stage, m is a number of markers, d is a number of disease associated markers, n_1 is an overall number of person tested on both stage, π is proportion between number of person tested on the first stage and n, α_1 is the probability of the error type I on the first stage, β_1 is the probability of the error type II on the first stage.

The results of an optimal design we obtained by refinement optimal plans from the tables 1-3 [10], where optimal robust two-stage design using $MAX3$ test were considered. The genetic model parameters, optimal n, c_1, c_2, π and overall costs are shown in the Tables 3, 4 and 5. Overall costs can be expired in terms of the symmetric Kullback-Leibler divergence between cases and controls distributions under alternative hypothesis as it is shown in Figure 3.

5 The number of disease associated markers

Since costs are depended on the number of disease associated markers d in 2-stage case-control association study, it is interesting what kind of dependencies that is. We constructed an optimal design for different values of d (Table 6) and revealed that this dependence has a linear shape (Figure 4), but proportional coefficient is so small that it has a slight influence on costs.

Table 3: Optimal design for recessive genetic model

p_a	ψ_1	ψ_2	ρ_{KL}	n	π	c_1	c_2	Costs, $10^9 \times C_{G_1}$	$n\rho_{KL}$
0.1	1.00	1.25	0.00054	383340	0.489	3.6650	5.4876	159.520	209
0.1	1.00	1.50	0.00195	105627	0.490	3.6020	5.5407	44.557	206
0.1	1.00	1.75	0.00398	53239	0.486	3.6797	5.3274	22.019	212
0.1	1.00	2.00	0.00648	33127	0.482	3.6881	5.2479	13.608	215
0.3	1.00	1.25	0.00442	46291	0.489	3.5949	5.3447	19.616	205
0.3	1.00	1.50	0.01560	13457	0.489	3.6902	5.4011	5.595	210
0.3	1.00	1.75	0.03136	6688	0.490	3.6847	5.4268	2.788	210
0.3	1.00	2.00	0.05035	4182	0.488	3.6806	5.4602	1.740	211
0.5	1.00	1.25	0.00979	21219	0.489	3.6851	5.4309	8.839	208
0.5	1.00	1.50	0.03348	6251	0.489	3.6809	5.4672	2.607	209
0.5	1.00	1.75	0.06547	3187	0.489	3.6809	5.4784	1.328	209
0.5	1.00	2.00	0.10250	2039	0.489	3.6841	5.4988	0.850	209
0.9	1.00	1.25	0.00719	28630	0.488	3.6647	5.4056	11.901	206
0.9	1.00	1.50	0.02248	9160	0.489	3.6638	5.4796	3.815	206
0.9	1.00	1.75	0.04084	5079	0.486	3.6626	5.4436	2.106	207
0.9	1.00	2.00	0.06010	3456	0.487	3.6638	5.4005	1.436	208

Table 4: Optimal design for log-additive genetic model

p_a	ψ_1	ψ_2	ρ_{KL}	n	π	c_1	c_2	Costs, $10^9 \times C_{G_1}$	$n\rho_{KL}$
0.1	1.25	1.56	0.00970	21332	0.479	3.6492	5.4345	8.787	207
0.1	1.50	2.25	0.03409	6096	0.480	3.6610	5.4401	2.508	208
0.1	1.75	3.06	0.06831	3072	0.480	3.6623	5.4518	1.263	210
0.1	2.00	4.00	0.10925	1934	0.479	3.6636	5.4746	0.794	211
0.3	1.25	1.56	0.02169	9667	0.480	3.6774	5.4724	3.980	210
0.3	1.50	2.25	0.07339	2860	0.480	3.6808	5.4639	1.176	210
0.3	1.75	3.06	0.14208	1484	0.480	3.6819	5.4668	0.610	211
0.3	2.00	4.00	0.22033	962	0.480	3.6814	5.4735	0.396	212
0.5	1.25	1.56	0.02479	8439	0.480	3.6815	5.4651	3.475	209
0.5	1.50	2.25	0.08110	2589	0.480	3.6791	5.4695	1.066	210
0.5	1.75	3.06	0.15266	1382	0.480	3.6810	5.4700	0.569	211
0.5	2.00	4.00	0.23116	924	0.480	3.6800	5.4700	0.381	214
0.9	1.25	1.56	0.00826	25084	0.481	3.6541	5.4590	10.343	207
0.9	1.50	2.25	0.02552	8198	0.479	3.6581	5.4472	3.370	209
0.9	1.75	3.06	0.04594	4854	0.479	3.6600	5.4447	1.996	223
0.9	2.00	4.00	0.06711	3227	0.478	3.6570	5.4569	1.326	217

Table 5: Optimal design for dominant genetic model

p_a	ψ_1	ψ_2	ρ_{KL}	n	π	c_1	c_2	Costs, $10^9 \times G_1$	$n\rho_{KL}$
0.1	1.25	1.25	0.00814	25324	0.485	3.6608	5.4397	10.488	206
0.1	1.50	1.50	0.02814	7305	0.487	3.6644	5.4602	3.033	206
0.1	1.75	1.75	0.05558	3715	0.488	3.6632	5.4675	1.546	206
0.1	2.00	2.00	0.08780	2356	0.489	3.6698	5.4509	0.980	207
0.3	1.25	1.25	0.01237	16998	0.488	3.6858	5.4322	7.069	210
0.3	1.50	1.50	0.04046	5129	0.489	3.6818	5.4706	2.135	208
0.3	1.75	1.75	0.07618	2725	0.489	3.6839	5.4481	1.135	208
0.3	2.00	2.00	0.11539	1800	0.490	3.6817	5.4656	0.750	208
0.5	1.25	1.25	0.00886	23634	0.485	3.6811	5.4676	9.803	209
0.5	1.50	1.50	0.02793	7495	0.487	3.6805	5.4695	3.116	209
0.5	1.75	1.75	0.05108	4102	0.488	3.6828	5.5008	1.709	210
0.5	2.00	2.00	0.07557	2781	0.489	3.6820	5.4656	1.159	210
0.9	1.25	1.25	0.00045	468107	0.487	3.6696	5.5221	194.310	209
0.9	1.50	1.50	0.00136	154800	0.485	3.6770	5.5774	63.976	211
0.9	1.75	1.75	0.00244	87763	0.483	3.6851	5.2692	36.122	214
0.9	2.00	2.00	0.00354	61325	0.481	3.6935	5.7002	25.149	217

Table 6: 2-stage experiment costs for different number of disease associated markers for recessive genetic model

a	ψ_2	d	n	π	c_1	c_2	Costs. $10^9 \times C_{G_1}$
0.1	2	1	3.312728	0.4820548	3.6881411	5.2479493	13.6084
0.3	2	1	0.418234	0.48787	3.680582	5.460205	1.7398
0.5	2	1	0.2039411	0.4887829	3.684086	5.498811	0.8497
0.9	2	1	0.3456458	0.4874937	3.663774	5.4005	1.4364
0.1	2	3	3.309047	0.4834189	3.69151	5.689438	13.6153
0.3	2	3	0.4187521	0.4884526	3.688388	5.526668	1.7415
0.5	2	3	0.2042659	0.4881194	3.688066	5.530075	0.8499
0.9	2	3	0.3453255	0.4876306	3.660024	5.44371	1.4366
0.1	2	10	3.325978	0.4765964	3.673375	5.544194	13.6053
0.3	2	10	0.4180625	0.4884433	3.681854	5.482262	1.7419
0.5	2	10	0.2041193	0.4883948	3.686623	5.518704	0.8505
0.9	2	10	0.3453255	0.4876306	3.660024	5.44371	1.4377
0.1	2	100	3.327556	0.4758604	3.671292	5.532026	13.7458
0.3	2	100	0.4188474	0.4870465	3.682143	5.465708	1.7594
0.5	2	100	0.204077	0.4884602	3.686299	5.516228	0.8590
0.9	2	100	0.3455149	0.4876147	3.66204	5.443868	1.4526

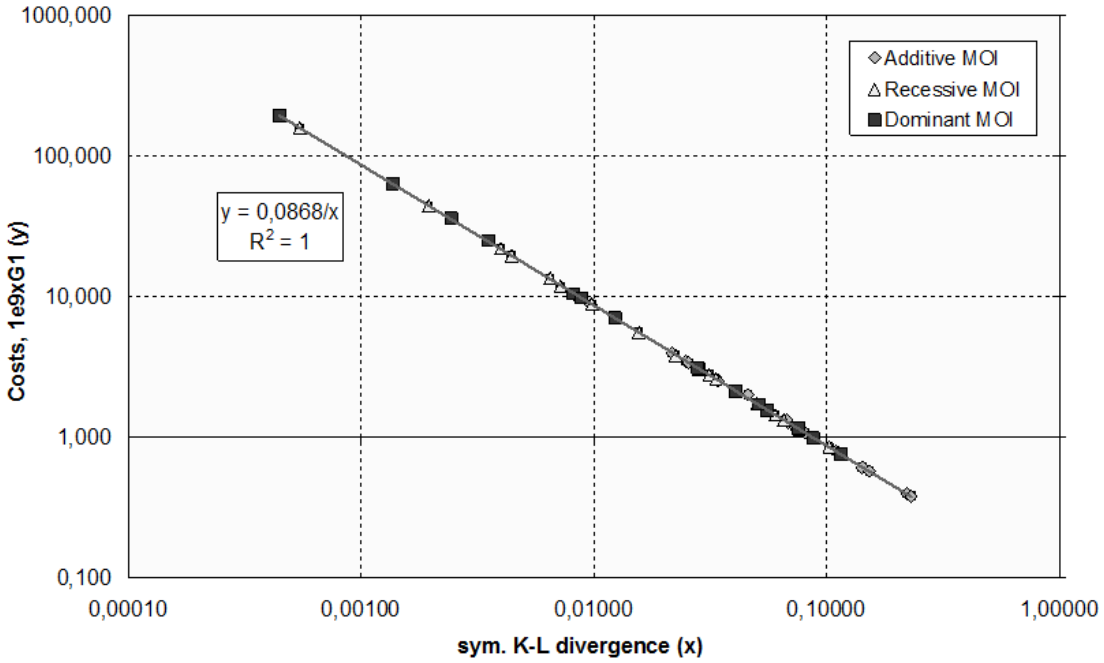

Figure 3: Relationship between overall costs and the symmetric KL-divergence for the $MAX3$ test ($\alpha = 0.05/610000$ $\beta = 0.1$)

6 Discrete design

Modern genotyping technologies allow using 96 and 384 well plates. Thus the number of individuals should be proportional to well plate size. Therefore the optimal study designs can be considered when the first stage sample size is a multiple of the well plate size (On the second stage another genotyping technology is used, because the number of markers is small).

In the table 7 the results for additive genetic model are shown when the number of the well plates is equal 384, $p_a = 0.3$, $\psi_1 = 2$, $\psi_2 = 4$, $K = 0.05$, $m = 610000$, $\alpha = \frac{0.05}{m}$, $\beta = 0.1$, $d = 1$, the proportion of the cases and controls is equal 0.5, $\frac{C_{G_2}}{C_{G_1}} = 100$, $\frac{C_R}{C_{G_1}} = 10^5$. The estimations $\hat{\alpha}$ and $\hat{\beta}$ were found by Monte Carlo method with $N_a = 2.18 \times 10^{10}$ and $N_a = 9.55 \times 10^6$ replicates for α and β respectively.

The optimal plan requires 461 persons on the first stage and 501 person on the second stage. From two available discrete plan (384 and 768 individuals) the first plan is better, because it is more close to restrictions and a little loses the optimal continue plan by costs.

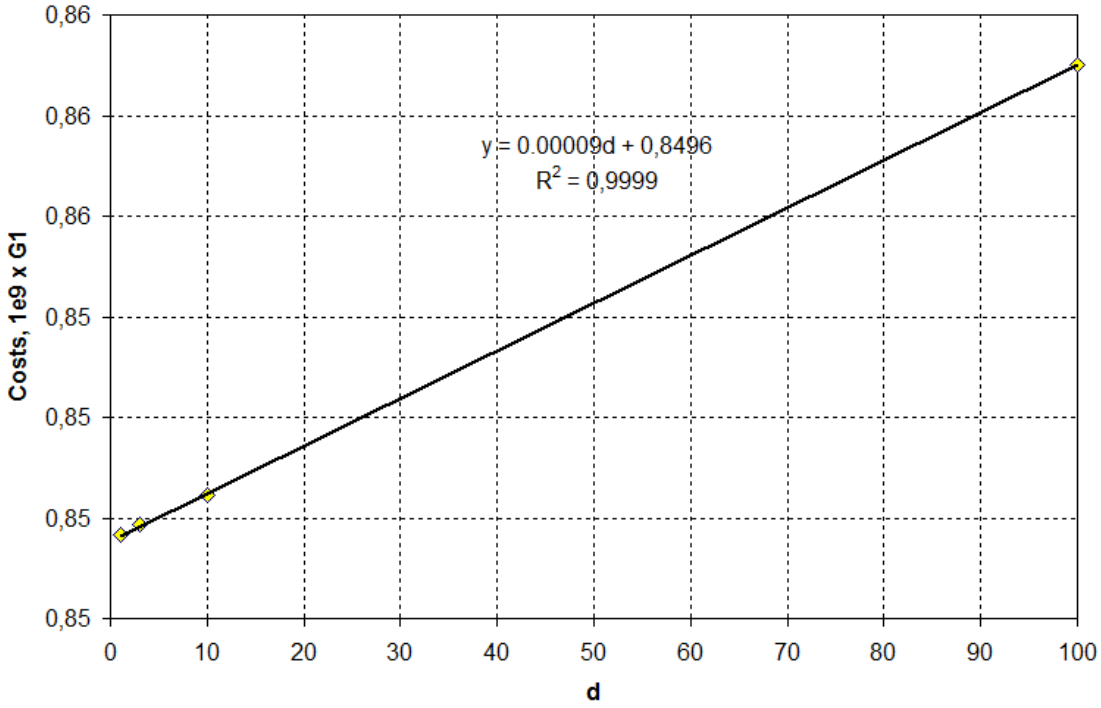

Figure 4: The relationship between costs and number of disease associated markers ($p_a = 0.5, \psi_1 = 2, \psi_2 = 4$)

7 Discussion

The most interesting result of our work is the relationship between optimal sample size (and overall costs) and the symmetric Kullback-Leibler divergence for $MAX3$ test optimal design under different genetic models. So the empirical symmetric Kulback-Leibler divergence

$$\hat{\rho}_{KL} = \frac{1}{rs} \sum_{i=0}^{2} (sr_i - rs_i) \ln \frac{r_i}{s_i}$$

Table 7: Optimal discrete design for additive genetic model (in first row). Well plate size is equal 384

Number of Well plates	n_1	n_2	Costs $10^8 \times C_{G_1}$	c_1	c_2	$\hat{\alpha}$ 10^{-8}	$\hat{\beta}$
1	384	578	4.06	3.3489	5.38015	14.43	0.097
2	768	194	5.65	4.6389	5.58331	7.47	0.046
Optimal	461	501	3.97	3.68142	5.47351	9.69	0.086

can be used as a measure of association. Under null hypothesis $n\hat{\rho}_{KL}/8$ slowly tends to standard exponential random variable.

Acknowledgements

This work was supported by DAAD grant A/11/76161 (first coauthor).

References

[1] White J.E. (1982) A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol* Vol **115**, 119-128.

[2] Breslow N.E., Chatterjee N. (1999) Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *JRSS C* Vol **48**, 457-468.

[3] Elston R.C., Guo X., Williams L.W. (1996). Two-stage global search designs for linkage analysis: use of the mean statistic for affected sib pair. *Genet Epidemiol* Vol **18**, 97–110.

[4] Freidlin B, Zheng G, Li Z, Gastwirth JL. (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity* Vol **53**, 146–152.

[5] Kraft P. (2006) Efficient two-stage genome-wide association designs based on false positive report probabilities. *Pac Symp Biocomputing* Vol **11**,523-534.

[6] Kraft P, Cox DG. (2008) Study designs for genome-wide association studies *Adv Genet* Vol **60**, 465-504.

[7] Kullback, S and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, Vol. **55**, 79–86.

[8] Muller, H.-H., Pahl, R. and Schafer, H. (2007) Including sampling and phenotyping costs into the optimization of two stage designs for genome wide association studies. *Genet Epidemiol* Vol **31**, 844-852.

[9] Saito A, Kamatani N. (2002) Strategies for genome-wide association studies: optimization of study designs by the stepwise focusing method. sl J Hum Genet Vol bf 47, 360-365.

[10] Nguyen TT, Pahl R, Schafer H. (2009) Optimal robust two-stage designs for genome-wide association studies. *Ann Hum Genet.* Vol **73** 638–51.

[11] Sasieni PD (1997) From genotypes to genes: doubling the sample size. *Biometrics* Vol **53**, 1253–1261.

[12] Satagopan J.M., Verbel D.A., Venkatraman E.S., Offit K.E., Begg C.B. (2002) Two-stage designs for gene-disease association studies. *Biometrics* Vol bf58, 163–170.

[13] Satagopan J.M., Elston R.C. (2003) Optimal two-stage genotyping in population-based associaton studys. *Genet Epidemiol* Vol **25**, 149–157.

[14] Satagopan J.M., Venkatraman E.S., Begg C.B. (2004) Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* Vol **60**, 589–597.

[15] Service S.K., Sandkuijl L.A., Freimer N.B. (2003) Cost-effective designs for linkage disequilibrium mapping of complex traits. *Am J Hum Genet* Vol **72**, 1213-1220.

[16] Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2007) Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol* Vol **31**, 776-788.

[17] Slager SL, Schaid DJ (2001) Case-control studies of genetic markers: Power and sample size approximations for Armitages test for trend. *Hum Hered* Vol **52**, 149-153.

[18] Song K, Elston RC (2006) A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine mapping in case-control studies. *Stat Med* Vol **25**, 105-126.

[19] Tarone RE, Gart JJ (1980) On the robustness of combined tests for trends in proportions. *J Am Stat Assoc*. Vol. **75**, pp. 110–116.

[20] Wang K, Sheffield VC (2005) A constrained-likelihood approach to marker-trait association studies. *American Journal of Human Genetics* Vol **77**, 768–780.

[21] Wang, H., Thomas, D. C., Peer, I. and Stram, D. O. (2006) Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol* Vol **30**, 356-368.

[22] Zheng, G and Gastwirth, J.L. (2006) On estimation of the variance in Cochran-Armitage trend tests for genetic association using case-control studies. *Statist Med* Vol **25**, 3150–3159.

Parametric Models in the Analysis of Patients with Multiple Myeloma

Mariya Semenova[1] and Alexander Bitukov[2]
[1] *Novosibirsk State Technical University,*
Novosibirsk, Russia
[2] *The Hematology Center,*
Main Military Clinical Hospital named after N.N.Burdenko,
Moscow, Russia
e-mail: vedernikova.m.a@gmail.com

Abstract

The research of various schemes of chemotherapy for the patients with multiple myeloma has been carried out. The purpose of the investigation is to compare the response time to the treatment in two groups of patients who received different treatment. It has been proposed to use the accelerated life models with cross-effect for relating the distribution of response time to the scheme of chemotherapy, type of the response, etc. We have ascertained the fact that such as a complete response, partial response, minimal response, stabilization and progression of the disease in the group of patients treated by Bortezomibe were achieved faster than in the control group. We are continuing these studies.

Keywords: lifetime analysis, censored data, regression models, cross-effect of survival functions.

Introduction

Accelerated life models are used more and more often in oncology and hematology studies for estimation of the effect of explanatory variables on lifetime distribution and for estimation of the survival function under given covariate values see [6], [8] and [9].

The most popular and most widely applied survival regression model is the proportional hazards model (called also the Cox model) introduced by Sir David Cox. The popularity of this model is based on the fact that there are simple semiparametric estimation procedures which can be used when the form of the survival distribution function is not specified, see [4]. The survival functions for different values of the covariates according to the Cox proportional hazard (PH) model do not intersect. However, in practice this condition often does not hold. Then we need to apply some more complicated models which allow decreasing, increasing or nonmonotonic behavior of the ratio of hazard rate functions.

Following [1],[2] and [7] we illustrate possible applications of the Hsieh model (see [5]), which is particularly useful for the analysis of survival data with one crossing point.

1 Parametric models

Suppose that each individual in a population has a lifetime T_x under a vector of covariates $x = (x_1, x_2, ..., x_m)^T$. Let us denote by $S_x(t) = P(T_x \geq t) = 1 - F_x(t)$ the survival function and by $\lambda_x(t)$ and $\Lambda_x(t)$ the hazard rate function and the cumulative hazard rate function of T_x, respectively.

In survival analysis, lifetimes are usually right censored. The observed data usually are of the form $(t_1, \delta_1), ..., (t_n, \delta_n)$, where $\delta_i = 1$ if t_i is an observed complete lifetime, while $\delta_i = 0$ if t_i is a censoring time, which simply means that the lifetime of the i-th individual is greater than t_i.

a Proportional hazards model

The cumulative hazard rate for the Cox proportional hazards model is given by

$$\Lambda_x(t;\beta) = \exp\left(\beta^T \cdot x\right) \Lambda_0(t), \tag{1}$$

where β is the vector of unknown regression parameters, $\Lambda_0(t;\theta)$ is the baseline cumulative hazard rate function, which is presented in Table 1 for some commonly used baseline distributions.

This model implies that the ratio of hazard rates under different values of covariate x_2 and x_1 is constant over time:

$$\frac{\lambda_{x_2}(t)}{\lambda_{x_1}(t)} = \frac{\exp\left(\beta^T \cdot x_2\right)}{\exp\left(\beta^T \cdot x_1\right)} = \text{const} \tag{2}$$

However, this model is rather restrictive and is not applicable when the ratios of hazard rates are not constant in time. There may be an interaction between covariates and time, in which case hazards are not proportional.

Table 1: The cumulative hazard rate functions for baseline distributions

Distribution	$\Lambda_0(t;\theta)$
Exponential	t/θ_1
Weibull	$(t/\theta_1)^{\theta_2}$
Gamma	$-\log\left(1 - \frac{\Gamma(t/\theta_1, \theta_2)}{\Gamma(\theta_2)}\right)$
Lognormal	$-\log\left(\frac{1}{2} - \frac{1}{2\sqrt{\pi}}\Gamma\left(\frac{1}{2\theta_2}\log^2(t/\theta_1), \frac{1}{2}\right)\right)$

b Hsieh model

According to the idea of Hsieh, one possible way to obtain a nonmonotonic behavior of ratios of hazard rates is to take a power function of the baseline cumulative hazard function. Namely, Hsieh proposed the model given by

$$\Lambda_x(t;\beta,\gamma) = \exp\left(\beta^T \cdot x\right) \{\Lambda_0(t)\}^{\exp\left(\gamma^T \cdot x\right)}. \quad (3)$$

The parameters β and γ are m-dimensional. It is a generalization of the proportional hazards model taking the power $exp(\gamma^T x)$ of $\Lambda_0(t)$ instead of the power 1. It is easy to show that the Hsieh model implies that the hazard ratio between different fixed covariates is increasing from 0 to ∞ or decreasing from ∞ to 0. So, we have a cross-effect of hazard rates functions and survival functions [5].

To test the goodness-of-fit of the proportional hazards model to an observed data we will use the approach based on the residuals, which should fit closely to the standard exponential distribution if the model is indeed "correct". Testing the hypothesis H_0 whether the samples of observed residuals belong to a particular distribution can be carried out by means of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests and using the maximum likelihood estimates of unknown parameters.

In this paper, we also use these test statistics to choose the most suitable baseline distribution for observed data. The distributions of these test statistics are different for different baseline distributions, i.e. test statistics are measured in different scales. Therefore, it is logical to use the value $p = 1-G(S_n|H_0)$ as a numerical measure, where $G(S|H_0)$ is the distribution function of the statistic corresponding to the presumed parametric model and S_n is the value of the statistic calculated from the residuals. The greater the value of p, the better the fit is for the baseline lifetime distribution in the considered model to the data at hand [3].

2 Analysis of patients with multiple myeloma

This investigation of patients with multiply myeloma was carried out in The Hematology Center, in the Main Military Clinical Hospital named after N.N.Burdenko. The purpose of the investigation is to compare the response time to the treatment in two groups of patients. The difference in these groups is in the fact that the first group received chemotherapy with Bortezomibe, which is marketed as Velcade by Millennium Pharmaceuticals.

a Data description

The data include observations of 60 patients, 4 of which were randomly censored. Patients in the study were randomly assigned to one of two treatment groups: chemotherapy without Bortezomibe ($x_1 = 0$) or chemotherapy together with Bortezomibe ($x_1 = 1$).

In addition to treatment, several factors were also observed: type of response (the value $x_2 = 1$ corresponds to the general response, $x_2 = 0$ - the progression of the disease), sex ($x_3 = 1$ means that the patient is male, $x_3 = 0$ means that the patient is female), and age in years (x_4).

Table 2 also gives the response times in months (t) and the censoring indicator δ.

Table 2: The data of patients with multiply myeloma

t	δ	x_1	x_2	x_3	x_4	t	δ	x_1	x_2	x_3	x_4	t	δ	x_1	x_2	x_3	x_4
61	1	1	1	1	64	62	1	1	0	1	75	7	1	0	1	0	66
50	1	1	0	1	81	3	1	1	1	1	64	2	1	0	1	0	60
2	1	1	1	0	71	26	1	1	1	1	61	262	0	0	1	1	68
36	1	1	0	1	69	22	1	1	0	1	72	81	1	0	1	0	81
14	1	1	0	0	74	46	1	1	0	1	59	33	1	0	0	0	79
27	1	1	0	1	83	3	1	1	1	1	77	215	1	0	0	1	65
1	1	1	1	0	46	16	1	1	1	1	66	57	1	0	0	0	85
4	1	1	1	1	80	10	1	1	0	0	46	17	1	0	0	1	89
27	1	1	0	0	58	25	1	1	1	0	55	26	1	0	1	0	75
115	0	1	0	1	50	6	1	1	1	1	48	7	1	0	0	1	47
13	1	1	0	1	85	5	1	1	1	0	51	30	1	0	0	1	75
2	1	1	1	1	56	30	1	1	0	1	81	2	1	0	0	1	66
3	1	1	1	1	57	25	1	1	0	1	58	26	1	0	1	1	76
25	1	1	1	1	71	39	1	1	0	1	77	20	0	0	1	1	37
4	1	1	1	1	64	6	1	1	1	1	65	5	0	0	0	0	57
62	1	1	0	0	57	83	1	1	0	0	69	8	1	0	1	0	73
9	1	1	1	0	71	24	1	1	1	1	52	127	1	0	0	0	79
10	1	1	0	0	56	3	1	1	1	1	49	149	1	0	0	1	87
7	1	1	1	1	55	7	1	0	1	1	61	10	1	0	1	1	65
54	1	1	0	1	75	2	1	0	1	1	45	8	1	0	1	1	61

So, there are 38 observations in the first group and 22 observations in the second one. It should be noted that 4 observations are independent randomly censored observations.

b Simulation results

First of all, we estimated survival functions for patients in two groups of treatment using nonparametric Kaplan-Meier estimates since the sample is censored. On Figure 1, these estimates are presented.

As can been seen from Figure 1, the estimates of survival functions intersect once. By this reason the proportional hazards model can be inappropriate for these data (proportional hazard assumption was not hold (2)); and we propose using the Hsieh model with a cross-effect of survival functions for relating the distribution of response time to the scheme of chemotherapy and other factors.

We analyzed different parameterizations of Hsieh models for these data and considered the problem of choosing a distribution law for the baseline survival function that best fits these data. So, we obtained the values p for exponential (Exp), Weibull (Wei), gamma (Gam) and lognormal (LnN) Hsieh models. In Table 3, the maximum likelihood estimates of the model parameters θ, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ and $\gamma =$

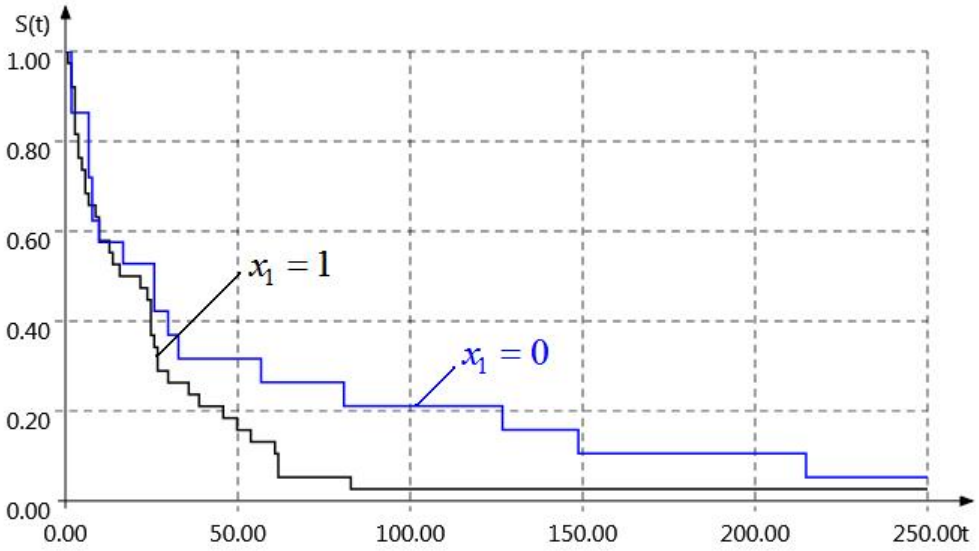

Figure 1: The Kaplan-Meier estimates of survival functions

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4)^T$, the statistics of Kolmogorov (S_k), Cramer-von Mises-Smirnov(S_{ω^2}) and Anderson-Darling (S_{Ω^2}) tests and the corresponding values of p are all presented.

As can been seen from Table 3, the lognormal Hsieh model fits the data much better than all other considered models. Now, let us consider the significance of the parameters β_1 and θ_1 by the Wald test. The null hypothesis to be tested is $H_0 : \beta_1 = \theta_1 = 0$, the statistic

$$W(\eta_0) = (\hat{\eta} - \eta_0)^T I(\hat{\eta}) (\hat{\eta} - \eta_0), \tag{4}$$

where $\hat{\eta}$ is vector $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4)$, $\eta_0 = (0.0, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, 0.0, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4)$ and $I(\hat{\eta})$ is the following estimation of Fisher information matrix

$$I(\eta) = -\frac{\partial^2 \log L(T_n; \eta)}{\partial \eta^2}.$$

For obtained value $W(\eta_0) = 13.09$ and corresponding χ^2 distribution the p-value is less than 0.01, therefore the hypothesis H_0 is rejected, parameters β_1 and θ_1 are not equal to 0, i.e. parameters for the first covariate in the model, namely type of chemotherapy, are significant.

Acknowledgements

This research has been supported by the Russian Ministry of Education and Science as part of the state task (project 8.1274.2011) and the Federal Target Program "Research and scientific-pedagogical personnel of innovative Russia" (contract number 14.B37.21.0860).

Table 3: Comparison of different baseline distributions for the Hsieh model

	$\hat{\theta}$	$\hat{\beta}$	$\hat{\gamma}$	S_k	p_{S_k}	S_{ω^2}	$p_{S_{\omega^2}}$	S_{Ω^2}	$p_{S_{\Omega^2}}$
Exp	106.14	1.26, 0.71, -0.56, 0.007	0.31, -0.57, -0.10, 0.004	1.05	0.01	0.19	0.01	1.10	0.01
Wei	4.21, 0.24	0.07, 2.37, -0.67, -0.05	0.41, -0.39, 0.06, 0.02	0.79	0.16	0.11	0.11	0.67	0.10
Gam	5.87, 3.93	0.54, 1.50, -0.49, -0.02	0.01, -0.89, -0.06, 0.003	0.71	0.12	0.11	0.09	0.75	0.08
LnN	15.68, 0.97	0.53, 1.65, -0.49, -0.02	0.11, -0.94, 0.03, 0.006	0.61	0.31	0.08	0.18	0.60	0.16

References

[1] Bagdonavicius, V. and Nikulin, M. *A Kolmogorov-Smirnov test for censored samples*. Boca Raton: Chapman and Hall/CRC, 2002.

[2] Bagdonavicus, V., Levuliene, R. and Nikulin, M. Testing absence of hazard rates crossings. *Comptes Rendus de lAcademie des Sciences de Paris, Ser. I*, 346, 7-8, 445-450, 2009.

[3] Balakrishnan, N., Chimitova, E., Galnova, N. and Vedernikova, M. Testing goodness-of-fit of parametric AFT and PH models with residuals.*Communications in Statistics - Simulation and Computation*, 42, 1352-1367, 2013.

[4] Cox, D.R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187-220, 1972.

[5] Hsieh, F. On heteroscedastic hazards regression models: theory and application. *Journal of the Royal Statistical Society, Series B*, 63, 63-79, 2001.

[6] Klein, J.P. and Moeschberger, M.L. *Survival analysis*. New-York:Springer, 1997.

[7] Nikulin, M. and Wu, H.-D. Flexible regression models for carcinogenesis data. *Probability and Statistics, Steklov Mathematical Institute in St.Petersburg, RAS*, 78-101.

[8] Piantadosi, S. Clinical Trials, *J.Wiley : New York*, 1997.

[9] Zeng, D. and Lin, D.Y. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of Royal Statistical Society, Series B* 69, 1-30, 2007.

On Systems with "Tube" Structure

A.V. Strelnikov and A.V. Medvedev
Siberian state aerospace university,
Krasnoyarsk, Russia
e-mail: ArkStrelVlad@mail.ru, saor_medvedev@sibsau.ru

Abstract

The problem of inertialess processes with "tube" structure identification is considered. The parametric and nonparametric algorithms are used for solution of this problem. The results of numeracy experiments are given.

Keywords: inertialess process, "tube" structure, nonparametric model.

Introduction

The problem of discrete continuous processes identification will be a long time one of the most of actual in control theory. On the initial stage of the process research a priori information availability is one of the most factor, which causes a variety of identification problems. Depending on this, a large variety of mathematical tasks are formulated. Below the discrete continuous processes identification in nonparametric uncertainty conditions, in other words in the conditions, when there is no a priori not only the laws of distribution of random factors, and information on the parametric model of the process, is told about. After that the special case, when inputs variables, influencing on the process, are stochastic depended form each other, is researched. Apparently this case was researched in [1] at the first time. The analysis of this case leads to the necessity to revise the identification problem of inertialess systems and the introduction of a new class of models describing the processes that have a "tube" structure in the space of input and output variables.

1 The general problem of identification

Recently the inertialess processes get more actual. It's connection with the fact that the measurements of the variables are made through significant periods of time. In particular it's happened when step of discretisation is larger than stability time. Let's consider the typical task of identification of [2], in which the process belongs to the class of static and corresponds to the scheme.

On the fig. 1 A is operator, describing the process, it's known with an accuracy of up to parameters (but the class of operator is known); $x(t) \in R^1$ is scalar output variable of the process; $u(t) \in R^n$ is vector input variable of the process; $\xi(t)$ is random action, $h^u(t)$ and $h^x(t)$ are noises, executing in the measuring lines; u_t and x_t are measurements of input and output variables at the moment t. The random action has dual character: it describes influence to the process by environment and random actions inside of the process. Type and distribution of the random action $\xi(t)$ are unknown. The random action $\xi(t)$ is centered random variable with limited variance.

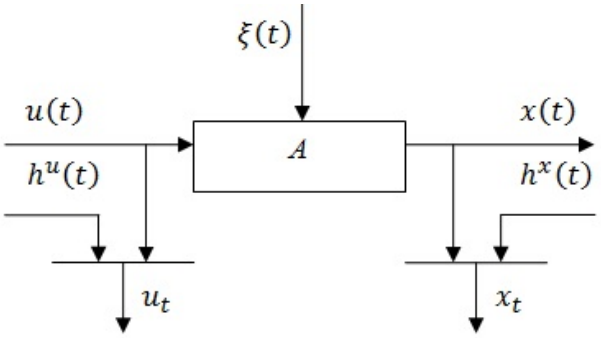

Figure 1: Scheme of the process

The distribution laws of the noises $h^u(t)$ and $h^x(t)$ are also unknown, centered and have limited variances.

Mathematically the process can be shown as the equation:

$$x(t) = A\left(u(t), \xi(t)\right). \tag{1}$$

The task is concluded to construct the mathematical model of dependence between output and input variables. Depending on the type of operator A task will have a different character. If a priori model structure is known, then we have identification problem in "narrow" sense. The parametric model will be used to solve this problem:

$$\hat{x}(t) = B_\alpha\left(u(t), \alpha\right), \tag{2}$$

when B_α is parametric operator, α is a set of parameters.

If the parametric model structure is unknown, then the identification problem has "wide" sense. The nonparametric model [3] will be used to solve this problem:

$$x_s(t) = B_s\left(u(t), \vec{x}_s, \vec{u}_s, \beta\right), \tag{3}$$

where B_s is nonparametric operator, β is parameters set of nonparametric model, $\vec{x}_s = (x_1, \ldots, x_s)$, $\vec{u}_s = (u_1, \ldots, u_s)$ are temporal vectors. Asymptotical properties of inertialess processes nonparametric estimations are considered in [4]. Unlike "wide" identification of "narrow" lies in the fact that there is no a selection stage of the variational structure. However the parameters are also configured with help samples $x_i, u_i, i = 1, \ldots, s$ and quality criterion.:

$$W(\beta) = M\left\{\left(x_i - x_s(u_i, \vec{x}_s, \vec{u}_s, \beta)\right)^2\right\} \to \min_\beta. \tag{4}$$

The similar criterion is used for configure parameters α in (2).

2 The processes with "tube" structure

This paper focuses on the modeling of processes having a "tube" structure. In the such processes there may be dependence between components of input vector $u(t)$. It's necessary more clear idea about "tube" structure to understand the problem.

In the input-output space $\Omega(x, u)$ the process have limited domain $\Omega^H(x, u)$. For clarity let's consider case, when $n = 2$. Type of the dependence between inputs variables has functional character:

$$\begin{cases} x(t) = f(u(t)) + \xi(t); \\ F(u(t)) + \varphi(t) = 0. \end{cases} \tag{5}$$

Here $f(u(t))$ is some unknown function and $\varphi(t)$ is centered random variable with limited variance. Schematically it can look like on the fig. 2.

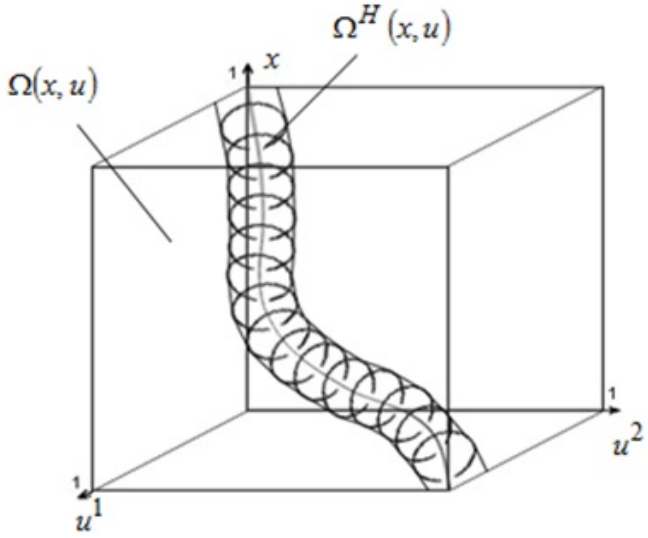

Figure 2: Scheme of the "tube" structure

If to look on the scheme (fig. 2), then you can see, that the domain area is limited $\Omega^H(x, u) \subset \Omega(x, u)$, as we told previously. Shown structure is like curved tube, therefore it was called "tube" structure.

For researched process the input-output variable space $\Omega(x, u)$ can be submitted as hypercube without loss of generality. This space is always known. On the scheme, shown on fig. 2, the hypercube is three-dimensional unit cube. If it's possible, then it' always advisable to transform space $\Omega(x, u)$ to unit hypercube to simplify calculations. The limited domain $\Omega^H(x, u)$ will be called "tube". Only in this field contain the values of the input variables, in which the process operates. It should be noted that this space is never known. This is the whole complexity of the studied problem.

If we use usual models (constructed without considering "tube" structure) for identification, then these models may contain value of input variables, in which the researched process can not operate. In other words, such models contain dots in hypercube $(x, u) \in \Omega(x, u)$, but they don't belong "tube" area $(x, u) \notin \Omega^H(x, u)$.

Work with such processes can be with the help of three instruments:

1. variance analysis;

2. indicator function;

3. cluster analysis.

The first method is concluded in search functional dependencies between variables and reduction the space $\Omega(u)$, until the "tube" structure will not disappear.

The second method enter the new special function, which estimate ownership researched dot to "tube" area: $u \in \Omega^H(u)$.

The third method separate the space $\Omega(u)$ on some subspaces with same dimension. However every space hasn't dependence between inputs variables, i.e. the process hasn't "tube" structure.

In this paper only the second modeling method is researched. Introduce indicator function:

$$\begin{cases} \Theta^H(u) = 1 & if \quad \sum_{s=1}^{s} \prod_{j=1}^{n} \Phi\left(c_{j,s}^{-1}(u_j - u_{j,i})\right) > 0; \\ \Theta^H(u) = 0 & otherwise, \end{cases} \tag{6}$$

where $\Phi(\cdot)$ is finite kernel function, $c_{j,s}$ are smooth parameters. When value of the indicator function equals one, then $u \in \Omega^H(u)$, when it equals zero, then $u \notin \Omega^H(u)$.

Now the parametric model (2) is transformed into:

$$\tilde{x}(t) = B_\alpha(u(t), \alpha) \cdot \Theta^H(u). \tag{7}$$

As the result we get absolutely new class of models with "tube" structure (7). Their speciality consists of the fact that, in addition to the estimates values of the output variables, they return again and ownership of the $\Omega^H(u)$.

3 The numeracy experiments

Let's demonstrate the advantage of the new class of models with the "tube" structure in the following example:

$$\begin{cases} x(t) = 0.5u_1(t) + 0.5u_2(t) + \xi(t); \\ u_2(t) = u_1(t) + \varphi(t), \end{cases} \tag{8}$$

where $\xi(t)$ and $\varphi(t)$ are random variables, distributed on the uniform law in range $[-0.05; 0.05]$.

Create five samples of statistically independent measurements of the volume $s = 200$. The first component of vector input variable $u_1(t)$ is random variable, distributed on uniform law in range $[0; 1]$. The random variables generator [5]. Values of the second component $u_2(t)$ are calculated according to the second equation in system (8), and according the first equation values of the output variable $x(t)$. The parametric model (2) is created for every case with help method of the least squares. The results

of modeling are shown on fig. 3. We have five models with different estimations of coefficients:

$$\begin{cases} \hat{x}_1(t) = 0.2u_1(t) + 0.8u_2(t); \\ \hat{x}_2(t) = 0.4u_1(t) + 0.6u_2(t); \\ \hat{x}_3(t) = -5u_1(t) + 6u_2(t); \\ \hat{x}_4(t) = 0.6u_1(t) + 0.4u_2(t); \\ \hat{x}_5(t) = 1.3u_1(t) - 0.3u_2(t). \end{cases}$$

As the result we have the researched process is described same structure models, but the have absolutely different coefficients. In the real situation it means that researcher need estimate parameters of the model in every new experiment. Two of the obtained models are out from limit of the unit cube $\hat{x} \notin \Omega(x)$. In other words got models aren't adequate. However values of the least-mean square criteria for every models are not exceed the one thousandth.

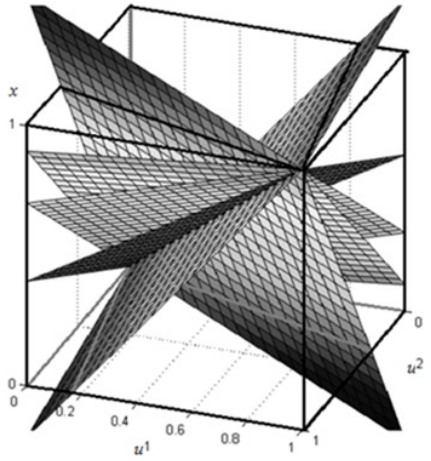

Figure 3: Results of parametric modeling

As we told, the parametric model estimates process great into area $\Omega^H(u)$, but out this area $u \notin \Omega^H(u)$ and $u \in \Omega(u)$ (i.e. the process doesn't operate with this values of the input variables) the model estimates very strangely. And fact, that "tube" area is never known, makes more complex the problem. In the control problem the model returns unrealizable control.

The results of modeling with using indicator function are shown on fig. 4.

Dark grey color denotes estimation of "tube" area $\Omega^H(u)$, light grey color denotes parametric model and black points denote samples.

We will spend one more experiment. This time we will solve the problem of "wide" identification. Let the object is described by the equation:

$$\begin{cases} x(t) = 0.5u_1^2(t) + 0.5u_2^2(t) + u_1(t)u_2(t) + \xi(t); \\ (u_1^2(t) - 0.5)^2 + (u_2^2(t))^2 = (0.4 + \varphi(t))^2. \end{cases} \tag{9}$$

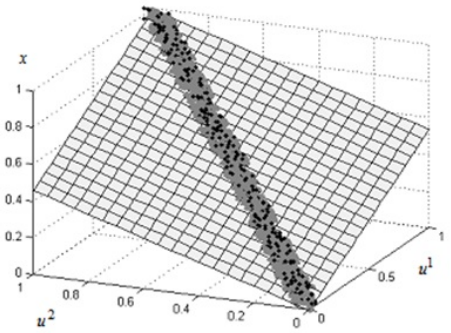

Figure 4: Results of "tube" modeling

As a nonparametric model we use Nadaraja-Watson's estimation [6]:

$$x(t) = \sum_{i=1}^{s} x_i \prod_{j=1}^{n} \Phi\left(c_{j,s}^{-1}(u_j(t) - u_{j,i})\right) \Big/ \sum_{i=1}^{s} \prod_{j=1}^{n} \Phi\left(c_{j,s}^{-1}(u_j(t) - u_{j,i})\right). \qquad (10)$$

The result is shown on fig. 5.

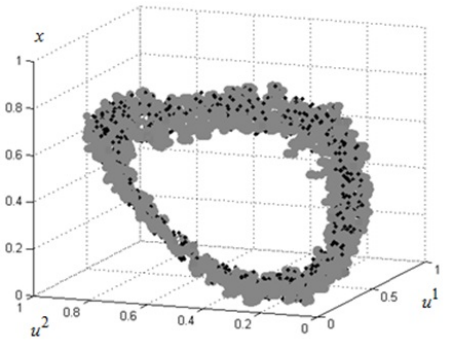

Figure 5: Results of nonparametric modeling

On the fig. 5 the result of nonparametric modeling without indicator function is shown. The fact is, that such model in a natural form has a "tube" structure, because in the dots of the space $\Omega(u)$ (when the process doesn't operate and there are no measurements) the model returns uncertainty, which is not operable by computer. It's just need exception.

4 The volume of "tube"

The researched processes obtain one property, it's "tube" area. If this area is exist, then its volume can be estimated.

In the considered examples we can see, that the volume of the hypercube is rather larger than volume of the "tube". The true volume of the space $\Omega^H(u)$ is also unknown. But it can be estimated some other way. Let's consider example with variate dimension:

$$\begin{cases} x(t) = n^{-1} \sum_{j=1}^{n} u_j(t) + \xi(t); \\ u_j(t) = u_1(t) + \varphi(t) \qquad j = 2, \ldots, n. \end{cases} \tag{11}$$

Estimation the volume of "tube" will be in two methods. The first method is geometrical. Let's imagine, that we can evaluate length of the "tube", i.e. depended inputs can be expressed:

$$u_j(t) = f_j(u_1(t)) + \varphi(t) \qquad j = 2, \ldots, n, \tag{12}$$

where $f_j(\cdot)$ are unknown functions. Then the volume could be estimated:

$$\hat{V}_1 = l \cdot \pi^{0.5n} \cdot r^n \cdot \Gamma^{-1}(0.5(n+2)), \tag{13}$$

where l is length of the "tube", r is its radius, Γ - gamma-function. Length and radius are evaluated by help estimations of functions $f_j(\cdot)$ between input variables.

The second method is concluded in the follow. In the space $\Omega(u)$ set of random dots, distributed on uniform law, is generated. The volume of "tube" equals fraction of amount of dots, owned by space $u \in \Omega^H(u)$, and amount of all generated dots. Ownership to "tube" area is estimated by indicator function (6). Then the volume could be estimated:

$$\hat{V}_2 = k_\Omega / k, \tag{14}$$

where k_Ω is amount of dots, owned by "tube" area, k is amount of all generated dots.

Let's experiment with different dimensions n. Value of noises is fixed 5%. The results of computing modeling are shown in the table 1.

Table 1: Results of volume evaluating

Method	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$
$\hat{V}_1$	$4.8 \cdot 10^{-3}$	$4.1 \cdot 10^{-4}$	$5.2 \cdot 10^{-5}$	$5.2 \cdot 10^{-6}$	$6.0 \cdot 10^{-7}$	$8.2 \cdot 10^{-8}$
$\hat{V}_2$	$7.7 \cdot 10^{-3}$	$4.5 \cdot 10^{-4}$	$5.2 \cdot 10^{-4}$	$2.0 \cdot 10^{-6}$	$8.2 \cdot 10^{-7}$	$2.3 \cdot 10^{-7}$
Truth	$5.6 \cdot 10^{-3}$	$5.9 \cdot 10^{-4}$	$6.2 \cdot 10^{-5}$	$6.4 \cdot 10^{-6}$	$6.8 \cdot 10^{-7}$	$6.9 \cdot 10^{-8}$

The first method is rather closer to truth and needs less time on evaluations. However the first method can be used in one case, when the dependence (12) between inputs is exist. When we get volume of the $\Omega^H(u)$, we can say about available "tube" structure of researched process.

Conclusions

The modeling of inertialess processes, having "tube" structure, was considered. The usual parametric models can't be used for estimation of these processes, but introduce the indicator function make models adequate. The new class of models with indicator function was got. The nonparametric models can be used for estimation of processes with "tube" structure in nature type. Also volume of "tube" structure was estimated accurate within order. In the future the control system of inertialess processes with "tube" structure will be constructed.

References

[1] Medvedev A.V. (1995). Data analysis in the identification problems. *Computing analysis and modeling*. Vol **2**, pp. 201-206.

[2] Eykhoff P. (1975). *Foundations of control systems identification*. Publ. of BSU, Minsk.

[3] Medvedev A.V. (1983). *Nonparametric systems of adaptation*. Nauka, Novosibirsk.

[4] Koshkin G., Piven I. (2009). *Nonparametric identification of stochastic objects*. Russian Academy of Science, Krasnoyarsk.

[5] Pervushin V.F., Sergeeva N.A., Strelnikov A.V. (2011). On random-number generator of given distribution. *Proceedings of the International Workshop "Applied Method of Statistical Analysis. Simulations and Statistical Inference" - AMSA'2011*, pp. 37-43.

[6] Nadaraja E.A. (1965). Nonparametric estimations of the regression curve. *Works VU AN GSSR*. Vol. **5**, pp. 56-68.

The Research of Distribution of the Ramsey RESET-Test Statistic

V. M. Volkova and V. L. Pankina
Novosibirsk State Technical University,
Novosibirsk, Russia
e-mail: vik-ky@yandex.ru, gfymrbyf@yandex.ru

Abstract

It is demonstrated that Ramsey Regression Equation Specification Error Test (RESET-test) is robust to disturbance in the assumption of a Gaussian distribution model errors. The power of Ramsey RESET-test is investigated.

Keywords: Model specification, a linear regression model, Ramsey RESET-test.

Introduction

RESET tests the significance of a regression of the residuals on a linear function vectors, which are obtained from the least-squares estimates of the depended variable. The specification errors considered are omitted variables, incorrect functional form. The effects on the Ramsey RESET statistics distribution of residuals of model specifications are considered. However, this test is designed with the assumption of Gaussian distribution of the random component of the model, which often disturbs in practice. Computer modelling methods provide an opportunity to study the influence of disturbance of this assumption on the properties of the distribution statistics.

1 Research of distribution of the Ramsey RESET-test statistic

Ramsey RESET-test tests whether non-linear combinations of the fitted values help explain the response variable. More specifically, if the null-hypothesis that all regression coefficients of the non-linear terms are zero is rejected, then the model suffers from mis-specification. F test used for proof of hypothesis. To research of distribution of the Ramsey RESET-test statistic under failure the assumption of a Gaussian distribution model errors we used bilateral exponential distribution (Be), distribution of maximum value (Max) and distribution of minimum value (Min). The significance level is equal 0.05 in all researches. The samples of distributions of the statistics are modeled with sample size is equal 20000, which made it possible to estimate the attainable significance point and the power with error to within 0.01. To implement RESET, we must decide how many functions of the fitted values to include in an expanded regression. To answer this question, we researched of distribution of the test statistic for different number of added regressors. The results of this researching are presented in Table 1. According to Table 1, the squared and cubed terms have

Table 1: Goodness-of-fit test's attainable significance point for different number of added regressors

Errors distribution — Number of added regressors	1	3	5	7
Gaussian	0.71	0.68	0.11	0.65
Be (0.5)	0.01	0.00	0.00	0.00
Be (10)	0.70	0.78	0.00	0.00
Max	0.90	0.06	0.08	0.00
Min	0.27	0.08	0.04	0.00

proven to be useful in most situations. Also, we researched the dependence of the test statistics distribution from the noise level. The results of this researches are presented in Table 2. According to Table 2, Ramsey RESET-test is robust as to a

Table 2: Goodness-of-fit test's attainable significance point for different noise level

Errors distribution — Noise level (%)	10	30	50	70
Gaussian	0.70	0.28	0.65	0.39
Be (0.5)	0.05	0.00	0.00	0.02
Be (10)	0.23	0.61	0.19	0.60
Max	0.11	0.49	0.06	0.25
Min	0.68	0.50	0.21	0.14

high noise level as to disturbance the assumption of model errors distribution.

2 Power of the Ramsey RESET-test

The power was analysed of the relative to the following alternative: regression coefficient of the squared term is not zero. We researched the dependence of the test power from the sample's size and from the noise level. The results of this researches are presented in Table 3 and Table 4, respectively.

According to Tables 3, Ramsey RESET-test's power is hight for different distributions of the observational errors for different sample size. According to Tables 4, Ramsey RESET-test's power is the highest for Gaussian distribution for different noise levels.

Conclusions

Ramsey RESET-test is robust as to a high noise level as to disturbance the assumption of model errors distribution. To implement Ramsey RESET-test, we must decide how

Table 3: Test's power for different sample size

Errors distribution — Sample size	20	40	60	80
Gaussian	1.00	1.00	1.00	1.00
Be (0.5)	0.96	1.00	1.00	1.00
Be (10)	0.99	1.00	1.00	1.00
Max	0.97	1.00	1.00	1.00
Min	0.98	1.00	1.00	1.00

Table 4: Test's power for different noise level

Errors distribution — Noise level (%)	10	30	50	70
Gaussian	1.00	1.00	1.00	1.00
Be (0.5)	1.00	0.95	0.86	0.77
Be (10)	1.00	0.98	0.86	0.72
Max	1.00	0.96	0.85	0.73
Min	1.00	0.97	0.88	0.75

many functions of the fitted values to include in an expanded regression. There is no right answer to this question, but the squared and cubed terms have proven to be useful in most situations. Ramsey test's power is approximately the same for different distributions of the observational errors.

References

[1] J.B. Ramsey. (1969). Test for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*. Vol. **31**, pp. 350-371.

[2] J. M. Wooldridge. (2009). *Introductory Econometrics. A Modern Approach, Fourth Edition*. South-Western Cengage Learning, Mason.

Joint Use of Monte Carlo and Geostatistical Techniques for Global Estimating of Functions Presented in an Integral Form

Anton V. Voytishek
Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia; e-mail: vav@osmf.sscc.ru

Abstract

In this paper the methodological and theoretical grounds for usage of kriging in the discrete-stochastic numerical algorithms are considered.

Keywords: functional Monte Carlo estimators, discrete-stochastic numerical algorithm (D-SNA), L_2-approach to conditional optimization of the D-SNA, "smoothness" of data at grid nodes, kriging

Introduction

Last thirty years the theory and applications of *the functional Monte Carlo estimators* are developed rapidly [6, 7, 8]. As a rule, corresponding *Monte Carlo discrete-stochastic schemes* are used for global solution of *Fredholm integral equations of the second kind*. They look as follows. A grid is introduced and some Monte Carlo estimators (*dependent, independent, "weakly" dependent*) are used for approximation of the solution in grid nodes. Then the "fulfilment" of solution is realized using approximate values at grid nodes and some basis of known functions (in this paper, *the local finite elements* [4] are used). New approaches to investigating errors of such schemes and to choosing the conditionally optimal parameters are elaborated [6, 7, 8].

In significant applications the independent estimators at grid nodes are used because of existence of singularities in required functionals and in kernels and free terms of integral equations. Independence of estimators leads to loss of smoothness for approximate values at grid nodes. Thus, some smoothing procedures are needed and some *geostatistical techniques* (in particular, *kriging*; see, for example, [3, 5, 9]) can be useful in framework of *L_2-approach* to conditional optimization of the corresponding discrete-stochastic schemes (see [6, 7, 8] and Section 2 of this paper). In this paper the theoretical foundations of such usage are considered.

1 Classical Monte Carlo estimators

1.1. The collision estimator. The main practical example of Monte Carlo calculations is a *linear functional*

$$I = (\varphi, h) = \int \varphi(x) h(x)\, dx \tag{1.1}$$

on the solution $\varphi(x)$ of Fredholm integral equations of the second kind

$$\varphi = K\varphi + l \;\text{ or }\; \varphi(x) = \int k(x', x)\varphi(x')\, dx' + l(x), \tag{1.2}$$

where K is an integral operator with a kernel $k(x', x)$ and $l(x)$ is a free term of the equation (1.2). The following Monte Carlo estimator – unbiased *collision estimator*

$$\zeta = \sum_{m=0}^{N} Q^{(m)} h(\xi^{(m)}) \tag{1.3}$$

is usually used for the value (1.1) (see, for example, [7]). Here

$$\xi^{(0)}, \xi^{(1)}, \ldots, \xi^{(N)} \tag{1.4}$$

are the states of a uniform Markov chain, which breaks with probability one (N equals to number of the break-state), defined by the initial density $\pi(x)$ and transition function $p(x', x) = (1 - q_a(x')) \times q(x|x')$; here $q_a(x)$ is the probability of breaking ("absorption") and $q(x|x')$ is the transition conditional density. Stochastic weights $Q^{(m)}$ can be defined reccurently:

$$Q^{(0)} = f(\xi^{(0)})/\pi(\xi^{(0)});\; Q^{(m)} = Q^{(m-1)} k(\xi^{(m-1)}, \xi^{(m)})/p(\xi^{(m-1)}, \xi^{(m)}). \tag{1.5}$$

The Monte Carlo method is an application of *the big numbers law* (see, for example, [7]) which is used for the estimator (1.3) as follows:

$$I = \mathbf{E}[\zeta] \approx S_n = \frac{1}{n} \sum_{j=1}^{n} \zeta_j. \tag{1.6}$$

Using *the central limit theorem* (see, for example, [2]) it is easy to get the ratio for the error of the method (1.6) (see, for example, [7]): with probability close to unit

$$|I - S_n| \;\sim\; H \times \frac{\sqrt{\mathbf{V}\mathrm{ar}[\zeta]}}{\sqrt{n}}; \quad H = \mathrm{const}, \tag{1.7}$$

where $\mathbf{V}\mathrm{ar}[\zeta]$ is the variance of the estimator ζ and $H < 3$.

For *optimization* of choice of the estimator ζ (it means the choice of the functions $\pi(x)$ and $p(x', x)$ for the chain (1.4)) we have to minimize *the cost* of the algorithm (1.6):

$$\mathrm{Cost} = t \times \mathbf{V}\mathrm{ar}[\zeta], \tag{1.8}$$

where t is an average time for computer calculating of an individual sample value ζ_j of the random value ζ (see, for example, [7]).

1.2. The local Monte Carlo estimator. For further reasoning it is necessary to note that the estimators of the type (1.3) can be derived for a value $\varphi(x_0)$ of the solution of the equation (1.2) in some individual point x_0.

There are exist two approaches for constructing an estimator for the value $\varphi(x_0)$ [7, 8]. The first one has a little bit "strange" name – *local estimator* (in fact, it is a kind of *"global"* estimator). The idea is to treat the first term in the right side of the equation (1.2) as a parametric functional $\int k(x',x)\varphi(x')\,dx' = (\varphi, h_x) = I_x$ and construct an unbiased collision estimator (1.3) for this functional, and finally

$$\varphi(x_0) = \mathbf{E}[\zeta(x_0)], \quad \zeta(x_0) = \sum_{m=0}^{N} Q^{(m)} k(\xi^{(m)}, x_0) + l(x_0) \tag{1.9}$$

(see [7, 8]). Note that in this case the same trajectories of the Markov chain (1.4) can be used for various points x (in this sense (1.9) is a *"global"* estimator as it was mentioned above). Moreover, it is possible to prove [8] that the error of the "continuous" Monte Carlo approximation $Z_n(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{m=0}^{N_i} Q_i^{(m)} k(\xi_i^{(m)}, x)\right) + l(x)$ has the same order with respect to n of the error in the C-norm as for "ordinary" Monte Carlo scheme – see formula (1.7):

$$\rho^{(C)}(\varphi, Z_n) = \sup_{x\in X} |\varphi(x) - Z_n(x)| \;\sim\; \frac{H_C}{\sqrt{n}}; \tag{1.10}$$

here X is a bounded compact domain in R^d. Nevertheless, for this result the function $\varphi(x)$ must be smooth enough with respect to the x (the second derivative on x must be bounded), thus, the kernel $k(x',x)$ and the free term $l(x)$ of the equation (1.2) must have the same smoothness. But for the most part of actual applied problems this smoothness does not exist.

EXAMPLE 1. One of the main applications of the estimator (1.3) is the modelling of heat transfer (see, for example, [6, 7]). The radiation flow is treated as as a steam of "small" particles which interact with "big" particles of medium. Under interaction the "small" particle can be absorbed by the "big" one or disperse stochastically. In this case $\varphi(x)$ is a total density of collisions and

$$k(x',x) = p_s(r')\chi_s(\omega|x')\left(\lambda e^{-\lambda|r-r'|}\right) \times \delta\left(\omega - \frac{r-r'}{|r-r'|}\right), \tag{1.11}$$

where $x = (r,\omega)$; $x' = (r',\omega')$ are the points of "small" particle's collision with "big" ones: r is a three-dimensional coordinate of a collision point, ω is a unit vector of the moving direction of the "small" particle before collision; $p_s(r')$ is the probability of survival of the "small" particle in the collision point r' (and $p_a(r') = 1 - p_s(r')$ is the probability of absorption); $\chi_s(\omega|x')$ is conditional density for new direction of moving (which includes the dispersion index); $\lambda e^{-\lambda|r-r'|}$ – is the density of length of free run (without collisions); for homogeneous environment it is exponential (this case is considered); $\delta(\omega-(r-r')/|r-r'|)$ is *a generalized function* (or *delta-function*); it is reflects that fact that free run occurs on a straight line.

Availability of the delta-function excludes the smoothness of the function (1.11) with respect to x. In practice, the delta-function is approximated by the smooth function in small neighborhood of the corresponding value of independent variable x,

and the estimator (1.9) is constructed for this comparatively small domain (that is why the stochastic function (1.9) is called *"local"* estimator).

1.3. The conjugate wandering method. In many practically considerable cases, the presence of singularities (in particular, delta-functions – like in the formula (1.11)) forces to use the following approach for calculating for the value $\varphi(x_0)$ (see also [6, 7, 8]). Note that it is possible to rewrite this value in the form (1.1):

$$\varphi(x_0) = \int \varphi(x) h_{x_0}(x)\, dx; \quad h^{(x_0)}(x) = \delta(x - x_0). \tag{1.12}$$

But the estimator (1.3) can not be used in this case because it is impossible to get the value of the delta-function in an individual point ξ. The following *dual representation* of the functional (1.1) can be easily derived (see, for example, [7]):

$$I = (\varphi, h) = (\varphi^*, l), \tag{1.13}$$

where $\varphi^*(y)$ is the solution of the *conjugate* (with respect to functional (1.1)) *equation*

$$\varphi^*(y) = \int k^*(y', y)\varphi^*(y')\, dy' + h(y); \quad k^*(y', y) = k(y, y'). \tag{1.14}$$

Thus, instead of the estimator (1.3) we can use the collision estimator of the functional (φ^*, l) from (1.13) on the solution of the integral equation (1.14)

$$\zeta^* = \sum_{m=0}^{N^*} Q^{*(m)} l(\xi^{*(m)}) \tag{1.15}$$

for the corresponding uniform Markov chain

$$\xi^{*(0)}, \xi^{*(1)}, \ldots, \xi^{*(N)} \tag{1.16}$$

(which reflects *the conjugate wandering*) with initial density $\pi^*(y)$ and transition function $p^*(y', y)$ (here N^* is the stochastic number of break-state of the chain);

$$Q^{*(0)} = h(\xi^{*(0)})/\pi^*(\xi^{*(0)});\ Q^{*(m)} = Q^{*(m-1)} k(\xi^{*(m)}, \xi^{*(m-1)})/p^*(\xi^{*(m-1)}, \xi^{*(m)}). \tag{1.17}$$

In the case of the indicating function (1.12) for estimator (1.16) of the value $\varphi(x_0)$ it is impossible to calculate the stochastic weight $Q^{*(0)}$ in ratios (1.17). Here the method of *including singularity into a density* (see, for example, [7]) can be used. Namely, we can choose *the delta-density* $\pi^{*(0)}(y) = \delta(y - x_0)$. It means that

$$\xi^{*(0)} = x_0 \quad \text{and} \quad Q^{*(0)} = 1 \tag{1.18}$$

with probability one (thus, the randomness "¡dissapear"¿ for the delta-density) and

$$\varphi(x_0) = \mathbf{E}[\zeta^{*(x_0)}]; \quad \zeta^{*(x_0)} = l(x_0) + \sum_{m=1}^{N^*} Q^{*(m)} l(\xi^{*(m)}). \tag{1.19}$$

Note that because of (1.18) the chain (1.16) and estimator (1.19) are realized exclusively for one point x_0 (thus, the estimator (1.19), compared with (1.9), is really *"local"*).

2 Functional Monte Carlo estimators

2.1. The discrete-stochastic algorithm. For many useful applications there is needed the "global" approximation of the solution $\varphi(x)$ of the equation (1.2) in a bounded compact domain $X \subset R^d$ [6, 7, 8]. We can use some "classical" method of numerical function approximation (see, for example, [1]) and introduce a grid $X^{(M)} = \{x_1, \ldots, x_M\}$ and then consider the approach

$$\varphi(x) \approx L_M\varphi(x) = \sum_{i=1}^{M} w_i(\boldsymbol{\varphi}^{(M)})\chi_i(x), \tag{2.1}$$

where $\Xi^{(M)} = \{\chi_1(x), \ldots, \chi_M(x)\}$ are known "basis" functions, and
$W^{(M)} = \{w_1, \ldots, w_M\}$ are some coefficients which depend on the values
$\boldsymbol{\varphi}^{(M)} = \{\varphi(x_1), \ldots, \varphi(x_M)\}$ of the function $\varphi(x)$ at grid nodes $X^{(M)}$; as a rule

$$w_i(\varphi(x_1), \ldots, \varphi(x_M)) = \varphi(x_i). \tag{2.2}$$

In our case the values $\{\varphi(x_1), \ldots, \varphi(x_M)\}$ are not given: we have to use Monte Carlo method (with estimators (1.9) or (1.19)) to get them:

$$\varphi(x_i) \approx S^{(i)}_{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} \zeta^{(i)}_j. \tag{2.3}$$

As a result we have the following approximation instead of (2.1):

$$\varphi(x) \approx L_M\tilde{\varphi}(x) = \sum_{i=1}^{M} w_i(S^{(1)}_{n_1}, \ldots, S^{(M)}_{n_M})\chi_i(x). \tag{2.4}$$

This formula presents *the discrete-stochastic algorithm* for approximation of the function $\varphi(x)$ (see [7, 8]).

2.2. Conditional optimization of the discrete-stochastic algorithm. One of the main problems in realization of the approximation (2.4) is: *how to choose the parameters M (number of nodes in the grid $X^{(M)}$) and number of sample values $n = \min(n_1, \ldots, n_M)$ of the estimators $\{\zeta^{(i)}\}$ of the values $\{\varphi(x_i)\}$?* The approach is highly natural [6, 7, 8]: we construct the upper boundary for the error of the approximation (2.4)

$$\delta^{(B)} = \rho_{B(X)}(\varphi, L_M\tilde{\varphi}) < T(M, n); \tag{2.5}$$

here $B(X)$ is a corresponding Banach functional space. The we consider the equation

$$T(M, n) = \Delta, \tag{2.6}$$

where Δ is an admissible level of error. From the equation (2.6) we express one parameter by means of another, for example, $n = v(M, \Delta)$, and then use this ratio for expression of the cost of the algorithm (2.4). As a rule, this expression has the form $\tilde{s} = H \times M \times n$ (here $H = \text{const}$), thus, we get the function

$$s(M) = H \times M \times v(M, \Delta)$$

on one independent variable M. Then we can find the minimum point M_{min} of this function using the approaches of mathematical or numerical functional analysis. Finally we get *the conditionally optimal parameters*

$$M = M_{min}(\Delta), \quad n = v(M_{min}, \Delta), \tag{2.7}$$

which can be used in calculations according the formula (2.4). Here we exploit the term *"conditional"* (see also [7, 8]) because in the equation (2.6) we use *the upper boundary* $T(M, n)$ (see the inequality (2.5)), but *not the exact dependence* of the error $\delta^{(B)}$ on parameters M and n.

2.3. Construction of upper boundaries for the error of the discrete-stochastic algorithm. Note that the error $\delta^{(B)}$ is a stochastic value. Thus, the inequality (2.5) must be treated in some probabilistic sense. Another problem is: how to choose the Banach space $B(X)$.

For using the geostatistical technique it is appropriate to use the so called L_2-approach (see [6, 7, 8]), where the upper boundary is constructing for the averaged second order deviation:

$$\left(\mathbf{E}\left[\delta^{(L_2)}\right]\right)^2 = \left(\mathbf{E}\left[\left(\int_X \left(\varphi(x) - L_M\tilde{\varphi}(x)\right)^2 dx\right)^{1/2}\right]\right)^2 < T^{(L_2)}(M, n), \tag{2.8}$$

where $T^{(L_2)}(M, n) \to 0$ while $M, n \to \infty$.

As basic functions $\Xi^{(M)}$ it is expedient to use local finite elements of the so called *Strang–Fix approximation* (because of their good properties concerning *stability*) [4].

3 Usage of kriging

3.1. Expediency for usage of some smoothing procedure for the discrete-stochastic algorithm. Note that for many important applications the independent stochastic estimators (namely, the estimators (1.19) of the conjugate wandering method) for getting the approximations $\{S_{n_i}^{(i)}\}$ of values $\{\varphi(x_i)\}$ are used. The main reason of such usage is the existence of singularities in required functionals and in kernels and free terms of corresponding integral equations (see, for example, formula (1.11)).

It is well-known (see [7, 8]) that the usage of independent estimators in grid nodes can disturb the smoothness of the solution $\varphi(x)$ even in the cases when this solution is continuous or smooth. Sometimes some tricks for involving of an artificial dependence are used [6, 7, 8], but they are not so effective. Thus, smoothing procedures for the values $\{S_{n_i}^{(i)}\}$ (together with or instead of "fulfilment" (2.1))) can give essential positive effect in global approximation of the function $\varphi(x)$.

3.2. Accuracy of data. Note that in practice for every i calculation of the value $S_{n_i}^{(i)}$ is laborious. Thus, the number of grid nodes M can not be rather big. In this case, according to formulas like (2.7), it is not necessary to use too many computer realizations of Markov chain trajectories (thus, their number n is not so

big). In this case the accuracy of data $\{S_{n_i}^{(i)}\}$ in grid nodes can be not so high. Thus, some *geostatistical smoothing technique* (in particular, *the kriging*; see, for example, [3, 5, 9]) seems to be useful as a modification of the corresponding discrete-stochastic algorithm (2.4).

Note also, that in opposite to standard setting of geostatistical problems we can use not only the final data $\{S_{n_i}^{(i)}\}$, but also some intermediate information $\{S_{\hat{n}_i}^{(i)}\}$; here $\hat{n}_i < n_i$.

3.3. Numerical approximation of the variogram. Note that the criteria (1.8) for optimization of the classical Monte Carlo method can not be used directly because the problem of calculation of variance $\mathbf{V}\mathrm{ar}[\zeta]$ has the same difficulty as the initial problem of calculating of the value $I = \mathbf{E}[\zeta]$. In practice the value $\mathbf{V}\mathrm{ar}[\zeta]$ is approximated by the preliminary calculations which use the sample values $\zeta_1, \ldots, \zeta_{\hat{n}}$ (the number $\hat{n}$ is rather less than the value n from the approximate ratio (1.6)). The simplest approximation has the form stochastic approximation of the variance has the form

$$\mathbf{V}\mathrm{ar}[\zeta] = \mathbf{E}[\zeta^2] - (\mathbf{E}[\zeta])^2 \approx V_{\hat{n}} = \frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \zeta_i^2 - (S_{\hat{n}})^2. \tag{3.1}$$

It is well known (see, for example, [7]) that if we treat ζ_i as independent identically distributed random variables, then $\mathbf{E}\,[V_{\hat{n}}] = (1 - 1/\hat{n})\mathbf{V}\mathrm{ar}[\zeta]$, i. e. the estimator $V_{\bar{n}}$ of the variance $\mathbf{V}\mathrm{ar}[\zeta]$ is *biased* (but *consistent*). It is easy to construct an *unbiased estimator* of variance:

$$\mathbf{V}\mathrm{ar}[\zeta] \approx \hat{V}_{\hat{N}} = \frac{V_{\hat{n}}}{1 - 1/\hat{n}} = \frac{1}{\hat{n} - 1} \sum_{i=1}^{\hat{n}} \zeta_i^2 - \frac{1}{\hat{n}(\hat{n} - 1)} \left(\sum_{i=1}^{\hat{n}} \zeta_i \right)^2. \tag{3.2}$$

The value t from criteria (1.8) can be also easily approximated using the preliminary calculations with the sample values $\zeta_1, \ldots, \zeta_{\hat{n}}$:

$$t \approx \frac{t_1 + \ldots + t_{\hat{n}}}{\hat{n}},$$

where t_i is the cost of calculation of the value ζ_i.

The approximations of the type (3.1), (3.2) can be also used for approximation of *the variogram* as follows (see, for example, [3, 5, 9]). Note that in this approximation the corresponding average is realized with respect the parameter M (the number of grid nodes in the set $X^{(M)}$) instead of the parameter n (the number of realizations of Monte Carlo estimators at grid nodes).

Let us consider (for simplicity) the one-dimensional case:

$$X = [0, 1], \quad x_i = i \times s; \quad s = 1/(M - 1); \quad i = 1, \ldots M,$$

and suppose $n_1 = \ldots = n_M = n$. Introduce the stationary random process

$$Z_n(x) = \frac{1}{n} \sum_{j=1}^{n} \zeta_j(x) - \varphi(x).$$

Consider also *the variogram* of this process (see, for example, [3, 5, 9]):

$$\gamma_n(h) = \frac{1}{2}\mathbf{V}\mathrm{ar}[Z_n(x+h) - Z_n(x)].$$

For the discrete-stochastic algorithm (2.4) we have a kind of *"geostatistical situation"*: we can use only and the values in corresponding grid nodes

$$Z_n^{(i)} = S_n^{(i)} - \varphi(x_i); \quad i = 1, \dots M;$$

moreover, we have to approximate somehow the unknown value $\varphi(x_i)$ (for example, using some intermediate information $\{S_{\hat{n}_i}^{(i)}\}$). Thus we can use only the values $h = k \times s$; $k = |i_1 - i_2|$ for corresponding indexes of the grid nodes x_{i_1} and x_{i_2} (as in the cases of practical use of kriging, see [3, 5, 9]).

Conclusions

In this paper the reasoning for usage of geostatistical smoothing technique in discrete-stochastic numerical algorithms together with corresponding numerical schemes are formulated.

References

[1] Bahvalov N. S., Zidkov N. P., Kobelkov G. M. *Numerical Methods.* Moscow: Binom, 2003 (in Russian).

[2] Borovkov A. A. *Probability Theory.* Moscow: "Nauka", 1986 (in Russian).

[3] Chilés J. P., Delfiner P. *Geostatistics – Modeling Spartial Uncertainty.* New York: John Wiley & Sons, 1998.

[4] Marchuk G. I., Agashkov V. I. *Introduction in Projection Grid Methods.* Moscow: "Nauka", 1988 (in Russian).

[5] Matheron G. *Random Sets and Integral Geometry.* New York: John Wiley & Sons, 1975.

[6] Mikhailov G. A. *Weighted Monte Carlo Methods.* Novosibirsk: SD RAS, 2000 (in Russian).

[7] Mikhailov G. A., Voytishek A. V. *Numerical Stochastic Modeling. Monte Carlo Methods.* Moscow: Publishing center "Akademiya", 2006 (in Russian).

[8] Voytishek A. V. *Functional Monte Carlo Estimators.* Novosibirsk: Novosibirsk State University, 2007 (in Russian).

[9] Wachernagel H. *Multivariate Geostatistics – An Intriduction and Applications.* Berlin: Springer, 1995.

Asymptotics of the Embryos Number for the Double Stochastic Model of First Order Phase Transitions

Svyatoslav I. Zolotukhin and Victor B. Kurasov
St. Petersburg State University, Physics Faculty, Russia
e-mail: tothejudge@gmail.com

Abstract

Kinetics of the first order phase transition is considered. Double stochastic processes controlled by inhomogeneous Poisson process were investigated. It has been shown that the consideration of double stochastic processes leads to a heavy-tailed probability distributions (as compared with a normal distribution).

Keywords: phase transition, Laplace transform, Cox process.

Introduction

The theoretical derivation of the rate for the first order phase transition is one of the actual tasks of the first order phase transition kinetics. To construct the global picture of the phase transition one has to take into account a depletion of a mother phase around growing objects of a new phase. Such problems have been studied, for example, in [8, 9, 10, 11, 12]

One of the first attempts to take into account the effect of the interaction of volumes free from the formation of the new phase was made in [5]. The process of the phase transition can be described by the following way: At some moment (let it be an initial moment) the embryos of a new phase begin to appear in the mother phase. Later these embryos begin to consume the mother phase and the regions of essential exhaustion will inevitably grow in the mother phase. Certainly, in the unexhausted regions the embryos of a new phase can appear later, but the total volume of the unexhausted region will decrease in time.

The process of the embryos formation satisfies the following natural assumptions:

1. The potential (without exhaustion) intensity of the embryos formation does to depend on the spatial point which implies the homogeneity of the substance. This value is assumed to be ***constant***.

2. The formation of embryos is the which implies that the probability of formation of n embryos in the time interval $[0, t]$ is given by:

$$p_n(t) = \frac{(\lambda t)^n}{n!} \exp(-\lambda t),$$

where λ is the constant intensity of the embryos formation.

This condition can be rewritten as:

$$\begin{cases} p_1(dt) = \lambda dt \\ p_{\geq 2}(dt) = o(dt) \end{cases},$$

where $p_n(dt)$ is the probability of embryos formation of n in the time interval dt.

3. The part of the exhausted volume at time t is proportional to the probability for some point to be out of the process of the embryos formation at t (***ergodic*** property).

Let us introduce the following notation.
$\alpha(t)$ is a intensity of the embryos formation in unit volume per unit time;
$W(t)$ is the relative volume of exhausted phase;
$V(t)$ is the relative volume of unexhausted phase;
$p(t)$ is the probability that an arbitrary point will be in the exhausted volume;
$q(t)$ is the probability that an arbitrary point will be in the unexhausted volume.
$N(t)$ is the number of embryo formed in the time interval $[0, t]$;
the volume of embryo grows in time as some function $\omega(t)$.

It is obvious: $p(t) + q(t) = 1$.
Due to ergodicity: $V(t) = q(t);\ W(t) = p(t)$.

Kolmogorov has received the following formulas:

$$q(t) = \exp\left(-\int_0^t \alpha(\tau)\omega(t-\tau)d\tau\right);$$

$$N(t) = \int_0^t \alpha(\tau)q(\tau)d\tau.$$

Similar results were obtained by the authors in a manner which allows some further generalization in the paper [14].

Note that these expressions are a ***deterministic*** functions of time. Now we going to study ***stochastic processes*** $N(t)$ and $W(t)$.

1 Main results

- probability generating function for stochastic process $N(t)$ and the Laplace transform for stochastic process $W(t)$ are found;

- the basic numerical characteristics of these processes are calculated;

- it is shown that the expectations of these processes are the same as corresponding deterministic functions, found earlier by Kolmogorov;

- asymptotic distributions of these processes as $t \to +\infty$ processes are found;

- ***double stochastic processes*** $N_2(t) \equiv N(\alpha(t))$ and $W_2(t) \equiv W(\alpha(t))$, controlled by ***stochastic process*** $\alpha(t)$, are investigated;

- asymptotic distributions of $N_2(t)$ and $W_2(t)$ are also found;

- it has been shown that the consideration of double stochastic processes leads to a heavy-tailed probability distributions (as compared with a normal distribution).

2 Process $N(t)$ as an inhomogeneous Poisson process

It is easy to show that by the assumptions 1-3 process $N(t)$ is an ***inhomogeneous Poisson process*** with local intensity

$$\lambda(\tau) = \alpha(\tau)q(\tau);$$

(see [4, 3]).

Khinchin showed that in this case the probability of formation of k events in the time interval $[\tau, t+\tau]$ is given by:

$$P_k(\tau, t) = \frac{(\Lambda(\tau, t))^k}{k!} \exp(-\Lambda(\tau, t)),$$

where

$$\Lambda(\tau, t) = \int_\tau^t \lambda(u)du.$$

The function $\Lambda(\tau, t)$, according to Khinchin, is called the leading function of inhomogeneous Poisson process. According to [7], we can derive the following formula:

$$P(N(t) = k) = \frac{(\Lambda(0, t))^k}{k!} \exp(-\Lambda(0, t)),$$

where

$$\Lambda(0, t) = \int_0^t \alpha(u)q(u)du.$$

Now we can calculate the generating function for process $N(t)$:

$$\Phi(z, t) \equiv \sum_{k=0}^{\infty} z^k P(N(t) = k).$$

Calculations yield

$$\Phi(z, t) = \exp\left(\Lambda(0, t)\,(z-1)\right).$$

Now, with the calculation of derivatives of the generating function, we can obtain the moments of the process $N(t)$. In particular:

$$EN(t) \equiv \left.\frac{\partial\Phi(z, t)}{\partial z}\right|_{z=1} = \Lambda(0, t) = \int_0^t \alpha(\tau)q(\tau)d\tau.$$

This result is analogous to that of Kolmogorov. Thus, the formula derived by Kolmogorov actually is a formula for the expectation of the (random) number of embryos. We can now calculate the variance of the number of embryos (and get the same value):

$$DN(t) \equiv \left.\frac{\partial \Phi(z,t)}{\partial z}\right|_{z=1} + \left.\frac{\partial^2 \Phi(z,t)}{\partial z^2}\right|_{z=1} - \left(\left.\frac{\partial \Phi(z,t)}{\partial z}\right|_{z=1}\right)^2 = \int_0^t \alpha(\tau) q(\tau) d\tau.$$

Now let's look for the asymptotic distribution of the process $N(t)$ as $t \to +\infty$.
Let

$$a = \int_0^\infty \alpha(\tau) d\tau \exp\left(-\int_0^\tau \alpha(u)\omega(\tau - u) du\right).$$

This integral converges, hence as $t \to +\infty$

$$\Phi(z,t) \to \exp(a(z-1)) \equiv \Phi_1(z).$$

This means that at infinity the process $N(t)$ converges to a Poisson random variable with parameter a.

Consider an important special case $\alpha(t) = \alpha$.

Then

$$\Phi_1(z) = \exp(\alpha(z-1) \int_0^\infty -\alpha \exp\left(-\int_0^\tau \omega(\tau - u) du\right) d\tau).$$

If more $\omega(t - \tau) = t - \tau$, then we have

$$\Phi_1(z) = \exp\left((z-1)\sqrt{\frac{\pi\alpha}{2}}\right).$$

3 Double stochastic Poisson process $N_2(t)$

Now, we assume that $\alpha(t)$ is itself a stochastic process. Then define the (double stochastic) process $N_2(t) \equiv N(\alpha(t))$. We have the so called Cox process controlled by $\alpha(t)$ (see [6]).

At first, we consider a special case, $\alpha(t) = \alpha$. Here α is a positive random variable with the density $f_\alpha(\alpha)$.

Let's calculate the generating function for double stochastic process $N_2(t)$:

$$\Phi_2(z,t) = \int_0^\infty \exp\left((\alpha(z-1) \int_0^t -\alpha \exp\left(-\int_0^\tau \omega(\tau - u) du\right) d\tau\right) f_\alpha(\alpha) d\alpha.$$

As $t \to +\infty$

$$\Phi_2(z,\infty) = \int_0^\infty \exp(a(z-1)) f_a(a) da,$$

where

$$a = \alpha \int_0^\infty \exp\left(-\alpha \int_0^\tau \omega(\tau - u)du\right) d\tau.$$

We can see that random variable a is a deterministic function of random variable α. That is, the distribution of a uniquely determines the distribution of α, and vice versa. We can say that we have a mixed Poisson distribution, and mixing distribution will have the density $f_a(a)$.

Here are some specific examples.

- if a has a gamma distribution, then $N_2(\infty)$ has a negative binomial distribution.
- if a has a exponential distribution, then $N_2(\infty)$ has a negative geometric distribution.
- if a has a shifted gamma distribution, then $N_2(\infty)$ has a **Delaporte** distribution.

The main difference of these distributions is that now the tails of the distributions decreases much more slowly than for the determined value of a (or, equivalently, α). Namely, Willmot theorem holds (see [13]).

Theorem (Willmot).

If

$$P(a > \lambda) \sim L(\lambda)\lambda^\alpha \exp(-\beta\lambda),$$

as $\lambda \to \infty$ and $L(\lambda)$ is a **slowly varying function** $\left(\lim\limits_{\lambda\to\infty} \frac{L(a\lambda)}{L(\lambda)} = 1\right)$ then as $\lambda \to \infty$

$$P(N_2(\infty) > n) \sim L(n)n^\gamma(\beta + 1)^{-1-n-\gamma},$$

with $\gamma < 0$ for $\beta \geq 0$.

This theorem implies that the rate of decrease of the tails of distribution $N_2(t)$ is a **power** function. It is not exponential function, as for the deterministic case.

In general, we can use the **theorem of continuity** (see [1]).

Theorem.

Suppose that there is a weak convergence of the process $\alpha(t)$ to the (positive) random variable αas $t \to \infty$.

Then the generating function of the process $N_2(t) \equiv N(\alpha(t))$ converges to

$$\Phi_2(z, \infty) = \int_0^\infty \exp(a(z - 1))f_a(a)da,$$

where

$$a = \alpha \int_0^\infty \exp\left(-\alpha \int_0^\tau \omega(\tau - u)du\right) d\tau.$$

This means that the process $N(\alpha(t))$ has the same asymptotic distribution as for special case $\alpha(t) = \alpha$ (where α is a random variable).

4 Process $W(t)$ with the deterministic control function $\alpha(t)$

We can calculate the Laplace transform of process $W(t)$:

$$\psi(s,t) \equiv E\exp(-sW(t)).$$

We'll obtain

$$\psi(s,t) = \exp\left(\Lambda(0,t)(E_\tau \exp(-s\omega(t-\tau)) - 1)\right).$$

Hence we have

$$EW(t) \equiv \left.\frac{\partial\psi}{\partial s}\right|_{s=0} = \Lambda(0,t)E_\tau\omega(t-\tau).$$

Let's define

$$\Lambda_1(0,t) \equiv \int_0^t \alpha(\tau)q_1(\tau,t)d\tau,$$

where

$$q_1(\tau,t) \equiv \exp\left(-\int_\tau^t \alpha(u)\omega(t-u)du\right);$$

Next, we consider the case when $\Lambda(0,t) = \Lambda_1(0,t)$(for example, this holds if $\alpha(\tau) = \alpha$, this is verified by direct calculation). Then we can find the expectation

$$EW(t) = \Lambda(0,t)E_\tau\omega(t-\tau) = 1 - \exp\left(-\int_0^t \alpha(u)\omega(t-u)du\right) = p(t).$$

This answer coincides again with the result obtained by Kolmogorov!

And we get the formula for the variance:

$$DW(t) = \int_0^t \omega^2(t-\tau)\alpha(\tau)\exp\left(-\int_0^\tau \alpha(u)\omega(t-u)du\right)d\tau.$$

These formulas follow from the formula for the Laplace transform:

$$\psi(s,t) = \exp\left(\int_0^t \left(\exp(-s\omega(t-\tau)) - 1\right)\alpha(\tau)\exp\left(-\int_\tau^t \alpha(u)\omega(t-u)du\right)d\tau\right).$$

Asymptotic distribution of process $W(t)$ when $\alpha(t) = \alpha;\ \omega(t-\tau) = t-\tau$.

Try to calculate explicitly the Laplace transform as $t \to +\infty$ in this important case. The result is as follows:

$\psi(s,\infty) = \exp\left(\sqrt{\frac{\pi\alpha}{2}}\left(2\exp\left(-\frac{s^2}{2\alpha}\left(1-\Phi\left(-\frac{s}{\sqrt{\alpha}}\right)\right)\right)-1\right)\right).$

Here $\Phi(\bullet)$ is the Laplace function.

Bivariate Laplace transform of the process $W(t)$.

$$\psi(s_1, s_2, t_1, t_2) =$$

$$= \exp(-\Lambda(0, t_2)) \exp\left(\int_{t_1}^{t_2} \exp(-s_2\omega(t_2 - \tau)))\alpha(\tau) \exp\left(-\int_{\tau}^{t_2} \alpha(u)\omega(t_2 - u)du\right) d\tau\right)$$

$$\exp\left(\int_{0}^{t_1} \exp(-i(s_1\omega(t_1 - \tau) + s_2\omega(t_2 - \tau)))\alpha(\tau) \exp\left(-\int_{\tau}^{t_1} \alpha(u)\omega(t_1 - u)du\right) d\tau\right).$$

Correlation function of the process $W(t)$.

If we know the Laplace transform, we can easily calculate the correlation function of the process.

$$K_W(t_1, t_2) =$$

$$= \int_0^{\min(t_1,t_2)} \omega(t_1 - \tau)\omega(t_2 - \tau)\alpha(\tau) \exp\left(-\int_{\tau}^{\min(t_1,t_2)} \alpha(u)\omega(\min(t_1, t_2) - u)du\right) d\tau.$$

5 Asymptotic distribution of the double stochastic process $W_2(t)$

Suppose that there is a weak convergence of the process $\alpha(t)$ to the (positive) random variable α as $t \to \infty$.

Then the Laplace transform of the process $W_2(t) \equiv W(\alpha(t))$ converges to

$$\psi_2(s, \infty) = \int_0^{\infty} \psi(s, \infty) f_\alpha(\alpha) d\alpha.$$

In particular for $\omega(t - \tau) = t - \tau$ we have

$$\psi_2(s, \infty) = \int_0^{\infty} \exp\left(\sqrt{\frac{\pi\alpha}{2}}\left(2\exp\left(-\frac{s^2}{2\alpha}\left(1 - \Phi\left(-\frac{s}{\sqrt{\alpha}}\right)\right)\right) - 1\right)\right) f_\alpha(\alpha) d\alpha.$$

Asymptotics for the tail distributions can be found using the Tauberian theorem (see [2]).

Let's show an example of the application of this theorem. Let the density of the random variable $\alpha^{-\frac{1}{2}}$ has a gamma distribution:

$$f_{\alpha^{-\frac{1}{2}}}(x) = \frac{\lambda^\rho}{\Gamma(\rho)} x^{\rho-1} e^{-\lambda x},\ x > 0.$$

Then as $s \to 0$

$$\psi_2(s) \sim \lambda^\rho \left(\lambda + \sqrt{\frac{\pi}{2}} \frac{s^2}{2} \right)^{-\rho},$$

and according to the Tauberian theorem density of $W_2(\infty)$ decreases as a power function as $x \to \infty$.

References

[1] *A.A. Borovkov* Probability theory, //M., Nauka, 1986 (in Russian).

[2] *William Feller* An Introduction to Probability Theory and Its Applications, Vol. 2.

[3] *B.V. Gnedenko, I.N. Kovalenko* vvedenie v teoriyu massovogo obsluzhivaniya //M., Nauka, 1987 (in Russian), 336p.

[4] *A.Ya. Khinchin* Works on the mathematical theory of queuing //M., Fizmatlit, 1963 (in Russian).

[5] *A.N. Kolmogorov* Isv. AN SSSR. Ser. Mat. 3, 1937 (in Russian), pp.355-360.

[6] *V.Yu. Kovalev, V.E. Bening, S.Ya. Shorgin* Matematicheskie osnovy teorii riska //M., Nauka, 2007 (in Russian), 542p.

[7] *I.N. Kovalenko, N.Yu. Kuznetsov, V.M. Shurenkov* Stochastic processes: handbook //Kiev, Naukova Dumka (in Russian), 1983, 366p.

[8] *V.B. Kurasov* Phys. Rev. E 63 (2001), 056123.

[9] *H.N.W. Lekkerkerker B. Widom* Physica A 285 (2000) 483-492.

[10] *H.N.W. Lekkerkerker, E.H.A. de Hoog* Physica A 298 (2001) 69-74.

[11] *S.M. Oversteegen, H.N.W. Lekkerkerker* Physica A 310 (2002) 181-186.

[12] *S.M. Oversteegen, H.N.W. Lekkerkerker* Physica A 341 (2004) 23-39.

[13] *Willmot G.E.* Adv. Appl. Probab. v.22 (1990) 147-159.

[14] *Zolotukhin S.I., Kurasov V.B.* Approximate account of the depletion regions overlapping in kinetics of nucleation. // St.Petersburg University Gerald ser.4 v.2 (2010) 8-15.

APPLIED METHODS OF STATISTICAL ANALYSIS. APPLICATIONS IN SURVIVAL ANALYSIS, RELIABILITY AND QUALITY CONTROL

Proceedings
of the international workshop
Novosibirsk, 25-27 september 2013

Научное издание

ПРИКЛАДНЫЕ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА. ПРИМЕНЕНИЕ В АНАЛИЗЕ ВЫЖИВАЕМОСТИ, ТЕОРИИ НАДЕЖНОСТИ И КОНТРОЛЕ КАЧЕСТВА

ТРУДЫ
Международного семинара
Новосибирск, 25-27 сентября 2013 г.

На английском языке

Подписано в печать 10.09.2013. Формат 60 × 84 1/8.
Усл. печ. л. 35,5. Уч.-изд. л. 66,03.
Тираж 50 экз. Заказ № 1107.

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20