APPLIED METHODS OF STATISTICAL ANALYSIS.

SIMULATIONS AND STATISTICAL INFERENCE

PROCEEDINGS

of the International Workshop

NOVOSIBIRSK

20-22 September 2011

UDC 519.22(063)

E d i t o r s:

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

Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Simulations and Statistical Inference" - AMSA'2011, Novosibirsk, Russia, 20-22 September, 2011. - Novosibirsk: Publishing house of NSTU, 2011. - 431 p.

ISBN 978-5-7782-1731-7
UDC 519.22(063)

Applied Methods of Statistical Analysis. Simulations and Statistical Inference

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. Vedernikova, N. Galanova, A. Gorbunova, A. Rogozhnikov

Preface

The International Workshop "Applied Methods of Statistical Analysis. Simulations and Statistical Inference" is organized by Novosibirsk State Technical University with the support of European Seminar "Mathematical Methods for Survival Analysis, Reliability and Quality of Life" (http://www.utc.fr/ nlimnios/SEMINAIRE/index.html) involving many of our friends and colleagues including N. Balakrishnan, C. Huber, H. Lauter, N. Limnios, M. Mesbah, M. Nikulin and V.Solev from European and North American universities. 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 European seminar supported many international conferences and workshops in the recent past including GOF'2000 (Goodness-Of-Fit) in Paris and MMR'2000 (Mathematical Methods in Reliability) in Bordeaux, an international workshop on "Parametric and Semi-parametric Models" organized by Mont Saint Michel in 2003, LAD'2004 in Saint Petersburg on Longevity, Aging and Degradation models, Biostat'2006 in Cyprus, ALT'2008 (Accelerated Life Testing) in Bordeaux, ALT'2010 in Clermont-Ferrand, MMR'2009 in Moscow and MMR'2011 in Beijing.

Statistical academics and students from Novosibirsk participated in many of these meetings, and now we have this nice meeting in Novosibirsk. 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 has 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.

M. Nikulin,
B. Lemeshko,
N. Balakrishnan

Contents

Simulation and Research of Probabilistic Regularities

Statistical Methods in Reliability and Survival Analysis

Testing Statistical Hypotheses

Nonparametric Methods

Application of Statistical Methods

Robust Methods of Statistical Analysis

Statistical Simulation of Natural Processes

Part I
Simulation and Research of Probabilistic Regularities

Simulation and Research of Probabilistic Regularities in Motion of Traffic Flows

M.A. Fedotkin, E.V. Kudryavtcev, M.A. Rachinskaya
N.I. Lobachevsky State University of Nizhni Novgorod
National Research University
Nizhni Novgorod, Russia
e-mail: fma5@rambler.ru

Abstract

A random number of vehicles crossed a transverse line of a highway during an arbitrary interval of time and a random number of vehicles situated on an arbitrary part of the highway at a fixed instant of time form a complicated stochastic dependence. In case of bad weather and bad road conditions the spatial intervals between the neighbouring vehicles at a fixed instant of time (a spatial characteristic of the flow) are dependent and have different probability distributions. Under these conditions, the intervals between moments of crossing the transverse line of the highway by the consecutive vehicles (a temporal characteristic of the flow) are also dependent and have different probability distributions. In this paper some probabilistic regularities in motion of this kind of traffic flows are researched.

Keywords: flow of vehicles, traffic batch, system of Kolmogorov differential equations, limiting probability distribution, parameters estimate.

Introduction

At present a great number of monographs and articles [1] deal with the mathematical theory of traffic flows in case of independent intervals between the consecutive moments when vehicles cross a certain transverse line in the highway. It is based on the assumption that the vehicles are homogeneous and uniform. At the same time in the real traffic some of similar assumptions couldn't be realized. For example, motion of a traffic flow could be complicated by the extreme weather conditions or by the heterogeneity of the vehicles. Also, their speeds are continuous random variables with the different distribution functions and the distances between the following vehicles are dependent and have different probability distributions. That is why observable traffic flow demands a mathematical model differed from those which are described in the classical queueing theory. A description of the flow of non-homogeneous vehicles is carried out in a nontraditional way, which is based on the analysis of the size distribution of so-called traffic batches and the slow vehicles flow distribution.

1 Distribution of traffic batch length

If the road surface is in a satisfactory state and the weather conditions are quiet enough, the motion of all kinds of vehicles will be quite unimpeded. Therefore, the flow could be described by

a Poisson flow. On the contrary, passing one vehicle by another one will become risky, dependent and will take considerable period in bad weather (fog, snow, icy conditions and so forth). In this situation we often observe formation of so-called traffic batches (groups of vehicles, motorcades) on intensive highways. Thus these flows could not be Poisson. The first who faced with this situation was Bartlett [2] observed a road traffic near to London in 1963. Bartlett and some other researchers did not manage to find a suitable distribution law for the dependent periods of the consecutive crossings of virtual stop line. Observing the traffic on highways close to large cities it was noticed that a traffic batch consists of one slow leading vehicle and a queue of some fast ones, which wait for an opportunity of overtaking. So it becomes necessary to construct a model of the spatial location of the vehicles in the highway. For this target an easy mechanism of traffic batches formation in case of intensive traffic and bad road conditions was offered. For a big number of highways it was found that without slow vehicles fast ones move free enough to suppose their joining to the motorcade have Poisson distribution with intensity $\lambda > 0$. This type of flows was researched in detail in [1], where a problem of unimpeded traffic was discussed. Consider basic probability space $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$. Let the symbol ω denote an elementary outcome or an element of certain event Ω. We will omit it in some cases to save space. Denote by $\xi(\omega; t, \Delta t)$ a random number of fast vehicles that join the slow one according to the Poisson law during the interval of time $[t, t + \Delta t)$.

Every slow vehicle can be interpreted as an servicing device for fast vehicles. Here the service time means the random passing time. In practice this time considerably depends on a number of vehicles in the batch (batch length). Let a random variable $\chi(\omega; t)$ measure this amount at the point of time $t \geq 0$. Denote by $\eta(\omega; t, \Delta t)$ a random number of fast vehicles which pass the slow one during the period $[t, t + \Delta t)$. We may naturally assume that the conditional probability of events generated by the random variable $\eta(\omega; t, \Delta t)$ in case of small $\Delta t > 0$ can be defined by the following relations:

$$
\begin{gathered}
\mathbf{P}(\{\omega\colon \eta(\omega; t, \Delta t) = 0\} | \{\omega\colon \chi(\omega; t) = 2, \xi(\omega; t, \Delta t) = 0\}) = 1 - \mu_1 \Delta t + o(\Delta t), \\
\mathbf{P}(\{\omega\colon \eta(\omega; t, \Delta t) = 1\} | \{\omega\colon \chi(\omega; t) = 2, \xi(\omega; t, \Delta t) = 0\}) = \mu_1 \Delta t + o(\Delta t), \\
\mathbf{P}(\{\omega\colon \eta(\omega; t, \Delta t) = 0\} | \{\omega\colon \chi(\omega; t) = 3, \xi(\omega; t, \Delta t) = 0\}) = 1 - \mu_2 \Delta t + o(\Delta t), \\
\mathbf{P}(\{\omega\colon \eta(\omega; t, \Delta t) = 1\} | \{\omega\colon \chi(\omega; t) = 3, \xi(\omega; t, \Delta t) = 0\}) = \mu_2 \Delta t - o(\Delta t), \\
\mathbf{P}(\{\omega\colon \eta(\omega; t, \Delta t) = 0\} | \{\omega\colon \chi(\omega; t) = m, \xi(\omega; t, \Delta t) = 0\}) = 1 - \mu_3 \Delta t + o(\Delta t), \\
\mathbf{P}(\{\omega\colon \eta(\omega; t, \Delta t) = 1\} | \{\omega\colon \chi(\omega; t) = m, \xi(\omega; t, \Delta t) = 0\}) = \mu_3 \Delta t - o(\Delta t), \\
m = 4, 5, \ldots
\end{gathered}
\tag{1}
$$

where $o(\Delta t)$ is an infinitesimal of higher order relative Δt as $t \to 0$. Parameters μ_1^{-1} and μ_2^{-1} in the equalities (1) specify an average passing time during which the fast vehicle can pass the slow one in cases when the traffic batch consists of two and three vehicles respectively. By analogy parameter μ_3^{-1} in the same equalities defines an average passing time when the batch consists of more than three vehicles. We will term parameters μ_1, μ_2, μ_3 passing intensities in respective cases. That is how the dependence between the average passing time and the length of traffic batch is simulated. The probabilities in (1) don't depend on time t. That's why we will omit

symbol t for the sake of simplification of record and designate variable $\eta(\omega; t, \Delta t)$ as $\eta(\omega; \Delta t)$.

The last equality from relations (1) can be interpreted in the following way. When the length of the batch is fixed, the conditional probability of an event consisted in fact that not less than two vehicles pass the slow during the period Δt (wherever it situates) is an infinitesimal of higher order relative Δt. Denote by $Q(t, m)$ the probability $\mathbf{P}(\{\omega\colon \chi(\omega; t) = m\})$ when we have fixed $t > 0$ and fixed $m = 1, 2, \ldots$ So a set of probabilities $\{Q(t, m); m = 1, 2, \ldots\}$ defines the distribution of a number of vehicles in the batch at the point of time $t \geq 0$.

With the help of reasonings analogous to those which are usually used in the queueing theory we can derive an infinite system of linear differential first-order equations for the probabilities $Q(t, m)$, when $t \geq 0$ and $m = 1, 2, \ldots$ If we fix quantity m in the random event $\{\omega\colon \chi(\omega; t+\Delta t) = m\}$, it will be easy to reveal the next equality

$$\begin{gathered} \{\omega\colon \chi(\omega; t + \Delta t) = m\} = \\ = \bigcup_{k=1}^{\infty} \bigcup_{n=0}^{\infty} \{\omega\colon \chi(\omega; \Delta t) = k, \xi(\omega; \Delta t) = n, \eta(\omega; \Delta t) = k + n - m\}. \end{gathered} \tag{2}$$

Using these relations and formulas (1) we get an infinite system of Kolmogorov linear differential equations:

$$\begin{gathered} dQ(t, 1)/dt = -\lambda Q(t, 1) + \mu_1 Q(t, 2), \\ dQ(t, 2)/dt = \lambda Q(t, 1) - (\lambda + \mu_1) Q(t, 2) + \mu_2 Q(t, 3), \\ dQ(t, 3)/dt = \lambda Q(t, 2) - (\lambda + \mu_2) Q(t, 3) + \mu_3 Q(t, 4), \\ dQ(t, m)/dt = \lambda Q(t, m - 1) - (\lambda + \mu_3) Q(t, m) + \mu_3 Q(t, m + 1), m \geq 4. \end{gathered} \tag{3}$$

Now suppose that at the moment $t = 0$ the number of vehicles in the batch is fixed and equals i. So dynamics of the batch length distribution is defined by the solution of system (3) with the initial conditions $Q(0, i) = 1$, $Q(0, m) = 0$, where $m \geq 1$ and $m \neq i$.

Note that to get an explicit solution of the system (3) we should derive and solve a partial differential equation for a generating function $\Pi_{\chi(t)}(t, z) = \sum_{m=1}^{\infty} z^m Q(t, m)$ like it is usually done in the most tasks of queuing theory. However in our case a solution process is too tedious and nontrivial. Fortunately according to our task we need only some characteristics and properties of the distribution of the traffic batch length under tending $t \to \infty$, i.e. some properties of the solution of system (3) as $t \to \infty$. Common properties of solutions of the similar systems were researched in detail in Kolmogorov's, Feller's, Ledermann's, Karlin's, Clarke's, McGregor's, Reuter's and Fedotkin's works. According to their results if $\lambda < \mu_3$, the solution of system (3) exists. Moreover, it is unique and satisfies the next conditions: $\sum_{m=1}^{\infty} Q(t, m) = 1$, $\lim_{t\to\infty} Q(t, m) = Q(m) > 0$ (for any $m = 1, 2, \ldots$) and $\sum_{m=1}^{\infty} Q(m) = 1$. We can note that the limits mentioned here don't depend on initial conditions and can be derived from a solution

of the infinite system of linear algebraic equations

$$\begin{gathered} 0 = -\lambda Q(1) + \mu_1 Q(2), \quad 0 = \lambda Q(1) - (\lambda + \mu_1)Q(2) + \mu_2 Q(3), \\ 0 = \lambda Q(2) - (\lambda + \mu_2)Q(3) + \mu_3 Q(4), \\ 0 = \lambda Q(m-1) - (\lambda + \mu_3)Q(m) + \mu_3 Q(m+1), \quad m = 4,\ 5,\ \ldots \end{gathered} \tag{4}$$

The distribution $\{Q(m); m = 1, 2, \ldots\}$ is named the limiting or the ergodic for the number $\chi(\omega)$ of all-type vehicles in the traffic batch. This distribution describes so-called stationary regime of the traffic motion. Conceptually the condition $\lambda < \mu_3$ means that intensity of joining to the slow vehicle should be smaller than passing intensity. Under this condition the system (4) is a result of passage to the limit $t \to \infty$ in all equations of the system (3) simultaneously subject to the equalities $\lim_{t\to\infty}(dQ(t,m)/dt) = 0$, $\lim_{t\to\infty} Q(t,m) = Q(m)$, $m \geq 1$. Let us remark that if we have fixed $m \geq 1$, then $\lim_{t\to\infty}(dQ(t,m)/dt)$ exist. Moreover, this limit is equal to zero for any possible m. If not, we would get an unbounded increase of absolute value of quantity $Q(t,m)$ as $t \to \infty$. Note it is impossible since quantity $Q(t,m)$ defines the probability. No we should pass to solution of the system (4).

Let's define a quantity $u_m = -\lambda Q(m) + \mu_3 Q(m+1)$ for $m = 3, 4, \ldots$ From the first three equations of the system (4) we get $Q(2) = \lambda\mu_1^{-1}Q(1)$, $Q(3) = \lambda^2\mu_1^{-1}\mu_2^{-1}Q(1)$ and $Q(4) = \lambda^3\mu_1^{-1}\mu_2^{-1}\mu_3^{-1}Q(1)$. That's why $u_3 = -\lambda Q(3) + \mu_3 Q(4) = 0$. From the system (4) we obtain $u_m - u_{m-1} = 0$, $m \geq 4$. Therefore $u_m = 0$ for $m \geq 3$. It now follows that $Q(m) = \lambda^2\mu_1^{-1}\mu_2^{-1}(\lambda\mu_3^{-1})^{m-3}Q(1)$. Substituting $\alpha = \lambda\mu_1^{-1}, \beta = \lambda\mu_2^{-1}, \gamma = \lambda\mu_3^{-1}$, subject to condition $\sum_{m=1}^{\infty} Q(m) = 1$ and inequality $\lambda < \mu_3$, we finally get

$$\begin{gathered} Q(1) = (1 + \alpha + \alpha\beta/(1-\gamma))^{-1}, Q(2) = \alpha(1 + \alpha + \alpha\beta/(1-\gamma))^{-1}, \\ Q(m) = \alpha\beta\gamma^{m-3}(1 + \alpha + \alpha\beta/(1-\gamma))^{-1}, m \geq 3. \end{gathered} \tag{5}$$

If we replace $(1 + \alpha + \alpha\beta/(1-\gamma))^{-1}$ by p, we obviously have the following expression for the generating function $\Pi_\chi(t,z) = \sum_{m=1}^{\infty} z^m Q(m) = p(z + \alpha z^2 + \alpha\beta z^3(1-\gamma z)^{-1})$. There will be no difficulty in showing that the expressions for the basic numerical characteristics satisfy the equalities

$$\begin{gathered} \mathbf{M}\chi(\omega) = p(1 + 2\alpha + \alpha\beta[2(1-\gamma)^{-1} + (1-\gamma)^{-2}]), \\ \mathbf{D}\chi(\omega) = (\alpha + \alpha\beta[(1-\gamma)^{-1} + (1-\gamma)^{-2} + 2(1-\gamma)^{-3}] + \\ +\alpha^2\beta[-(1-\gamma)^{-2} + 2(1-\gamma)^{-3}])p^2 + p^2\alpha^2\beta^2[-(1-\gamma)^{-3} + (1-\gamma)^{-4}]. \end{gathered} \tag{6}$$

The formulas (4) and equality $\gamma = \lambda\mu_3^{-1}$ make it possible to give the next easy meaning of the parameter γ when we talk about highway traffic in groups. The parameter γ specify a highway saturation rate by fast vehicles and it should be named utilization of the highway. Indeed, if $\mu_3 > \lambda$ and as $\mu_3 \to \lambda$, we clear have $p \to 0$ and $\gamma \to 1$. That's why the mathematical expectation of the number of all-type vehicles in the batch increases in the stationary regime unboundly. In practice this very fact can be observed by motorists when the weather conditions become considerably worse and there appear lengthy traffic jams. Let's take one important note.

It is easy to notice that if $\alpha = \beta = \gamma$, the given distribution coincides with the geometric distribution. From the equalities $\alpha = \lambda\mu_1^{-1}, \beta = \lambda\mu_2^{-1}, \gamma = \lambda\mu_3^{-1}$ we reveal coincidence mentioned above is possible only if $\mu_1 = \mu_2 = \mu_3$. In other words, if the average passing time doesn't depend on the number of vehicles in the traffic batch, we deal with the geometric distribution.

2 Traffic flow properties

A slow vehicles density (a number of slow vehicles per a unit length of the highway) in the real traffic flow is considerably smaller than a density of the fast ones. It can be assumed that in the stationary regime the slow vehicles move independently, so that the traffic flow of slow vehicles is supposed being the Poisson with the intensity μ. Let the maximal length of a highway section on which the average number $\mathbf{M}\chi(\omega)$ of vehicles is situated be much less than an average distance between the following slow vehicles. We can now assume that all the vehicles forming one autocade cross the virtual stop line simultaneously. For this kind of flows denote as $\varkappa(\omega;t)$ a random number of all-type vehicles crossed the fixed transverse line during the period of time $[0,t)$. For this variable introduce the generating function $\Pi_{\varkappa(t)}(t,z) = \sum_{m=0}^{\infty} z^m \mathbf{P}(\{\omega\colon \varkappa(\omega;t) = m\})$.

For the generating function $\Pi_{\varkappa(t)}(t,z)$ the following equality takes place:

$$\Pi_{\varkappa(t)}(t,z) = e^{-\mu t}\times$$
$$\times\sum_{k=0}^{\infty} z^k\{\sum_{n=0}^{[\frac{k}{2}]}\alpha^n[\frac{(\mu tp)^{k-n}}{n!(k-2n)!} + \sum_{m=1}^{min\{k-2n,n\}}\beta^m\sum_{l=0}^{k-2n-m}\gamma^l\frac{(\mu tp)^{k-n-m-l}C_{m+l-1}^l}{(n-m)!m!(k-2n-m-l)!}]\},$$

where $\left[\frac{k}{2}\right]$ means integer part of $\frac{k}{2}$.

This implies that our flow is defined by parameters μ, α, β and γ.

For the distribution $\{\mathbf{P}(\{\omega\colon \varkappa(\omega;t) = k\}), k \geq 0\}$ of the flow $\{\varkappa(\omega;t)\colon t \geq 0\}$ the following equalities take place:

$$\mathbf{P}(\{\omega\colon \varkappa(\omega;t) = 0\}) = e^{-\mu t}, \mathbf{P}(\{\omega\colon \varkappa(\omega;t) = k\}) =$$
$$= e^{-\mu t}\sum_{n=0}^{[\frac{k}{2}]}\alpha^n\left[\frac{(\mu tp)^{k-n}}{n!(k-2n)!} + \sum_{m=1}^{min\{k-2n,n\}}\beta^m\sum_{l=0}^{k-2n-m}\gamma^l\frac{(\mu tp)^{k-n-m-l}C_{m+l-1}^l}{(n-m)!m!(k-2n-m-l)!}\right].$$

Let $\mathbf{M}\varkappa(\omega;t)$ and $\mathbf{D}\varkappa(\omega;t)$ be an expectation and a variance of the random variable $\varkappa(\omega;t)$ respectively; then

$$\mathbf{M}\varkappa(\omega;t) = \mu tp\left(1 + 2\alpha + \alpha\beta\left[\frac{2}{1-\gamma} + \frac{1}{(1-\gamma)^2}\right]\right),$$
$$\mathbf{D}\varkappa(\omega;t) = \mu tp\left(1 + 4\alpha + \alpha\beta\left[\frac{4}{1-\gamma} + \frac{3}{(1-\gamma)^2} + \frac{2}{(1-\gamma)^3}\right]\right).$$

For the skewness $Sk(\varkappa(\omega;t))$ and the excess $Ex(\varkappa(\omega;t))$ we have equalities:

$$Sk(\varkappa(\omega;t)) = \left(1+8\alpha+\alpha\beta\left[\frac{8}{1-\gamma}+\frac{7}{(1-\gamma)^2}+\frac{6}{(1-\gamma)^3}+\frac{6}{(1-\gamma)^4}\right]\right)\times$$
$$\times(\mu tp)^{-1/2}\left(1+4\alpha+\alpha\beta(\frac{4}{1-\gamma}+\frac{3}{(1-\gamma)^2}+\frac{2}{(1-\gamma)^3})\right)^{-\frac{3}{2}},$$
$$Ex(\varkappa(\omega;t)) = (1+16\alpha+\alpha\beta[\frac{16}{1-\gamma}+\frac{15}{(1-\gamma)^2}+\frac{14}{(1-\gamma)^3}+\frac{12}{(1-\gamma)^4}+\frac{24}{(1-\gamma)^5}])\times$$
$$\times(\mu tp)^{-1}\left(1+4\alpha+\alpha\beta(\frac{4}{1-\gamma}+\frac{3}{(1-\gamma)^2}+\frac{2}{(1-\gamma)^3})\right)^{-2}.$$

Sum of n independent flows of this kind with parameters μ_i, α_i, β_i and γ_i $(i=\overline{1,n})$ is flow of the same kind if and only if $\gamma_1=\gamma_2=...=\gamma_n$. The summary flow has parameters: $\mu=\sum_{i=1}^n\mu_i, \alpha=\sum_{i=1}^n\mu_i\alpha_ip_i/\sum_{i=1}^n\mu_ip_i, \beta=\sum_{i=1}^n\mu_i\alpha_i\beta_ip_i/\sum_{i=1}^n\mu_i\alpha_ip_i, \gamma=\gamma_1$.

3 Case of low intensity of traffic flow

In case discussed above we assumed that intensity of the fast vehicles motion is sufficiently high. Further, we suppose that the passing intensity exceeds the intensity of the fast vehicles joining to the motorcade considerably. In this case relatively small motorcades generate indeed. Let N be their maximal length. It entails the batch joining mechanism changes in case when there are already N vehicles in the batch. If the fast vehicle catches up with the full batch consisted of N vehicles, it will join the batch, but at the same time the fast vehicle following right the slow one will pass it at the same instant. That's why we should change or add some calculations to our previous reasoning. First of all in equalities (1) parameter m varies within $4\le m\le N$. Moreover it is necessary to add the following equality:

$$\mathbf{P}(\{\omega\colon \eta(\omega;t,\Delta t)=1\}|\{\omega\colon \chi(\omega;t)=N,\xi(\omega;t,\Delta t)=1\})=1.$$

Also in (2) range of the unification index k is now $1\le k\le N$. Parameter m in the system of equations (3) has limits $4\le m\le N-1$ and there is one more equation in this system concerned with dynamics of batch in case of its fullness:

$$dQ(t,N)/dt=\lambda Q(t,N-1)-\mu_3Q(t,N).$$

We get the following stationary distribution of the vehicles number in the batch:

$$\begin{aligned} Q(1) &= (1+\alpha+\alpha\beta(1-\gamma^{N-2})/(1-\gamma))^{-1},\\ Q(2) &= \alpha(1+\alpha+\alpha\beta(1-\gamma^{N-2})/(1-\gamma))^{-1},\\ Q(m) &= \alpha\beta\gamma^{m-3}(1+\alpha+\alpha\beta(1-\gamma^{N-2})/(1-\gamma))^{-1}, 3\le m\le N. \end{aligned} \tag{7}$$

It could be easily shown that equalities (7) turn into (5) as $N\to\infty$.

Let's consider an important for practice case $N = 3$ when the fast vehicles density is rather small and they pass the slow one rather rapidly. It follows that we study a non-ordinary Poisson flow and in every calling instant one arrival comes with a probability of $p = (1 + \alpha + \alpha\beta)^{-1}$, two arrivals — with a probability of $q = \alpha(1 + \alpha + \alpha\beta)^{-1}$ and three — with a probability of $s = \alpha\beta(1 + \alpha + \alpha\beta)^{-1}$. For the one-dimensional distributions $P_k(t)$, $k \geq 0$, of the flow $\{\varkappa(t)\colon t \geq 0\}$ of the same probabilistic structure the following equalities are true

$$P_k(t) = e^{-\mu t} \sum_{i=0}^{[\frac{k}{2}]} \sum_{j=0}^{[\frac{k-2i}{3}]} \binom{k-i-2j}{i,j,k-2i-3j} p^{k-2i-3j} q^i s^j \frac{(\mu t)^{k-i-2j}}{(k-i-2j)!},\ k \geq 0.$$

The sum of m independent non-ordinary Poisson flows with parameters μ_j, p_j and q_j (where $j = 1, 2, \ldots, m$) is non-ordinary Poisson flow with parameters

$$\mu = \sum_{j=0}^{m} \mu_j, p = (\sum_{j=0}^{m} \mu_j p_j)/(\sum_{j=0}^{m} \mu_j), q = (\sum_{j=0}^{m} \mu_j q_j)/(\sum_{j=0}^{m} \mu_j).$$

For the numerical characteristics of random variable $\varkappa(\omega; t)$ we have expressions

$$\begin{aligned} \mathbf{M}\varkappa(\omega;t) &= \mu t(1+q+2s), \\ \mathbf{D}\varkappa(\omega;t) &= \mu t(1+3q+8s), \\ Sk(\varkappa(\omega;t)) &= (\mu t)^{-\frac{1}{2}}(1+7q+26s)(1+3q+8s)^{-\frac{3}{2}}, \\ Ex(\varkappa(\omega;t)) &= (\mu t)^{-1}(1+15q+80s)(1+3q+8s)^{-2}. \end{aligned} \tag{8}$$

It was revealed that the expressions for the numerical characteristics in case of unlimited batch length turn into formulas (8) if $\gamma = 0$. Conceptually this conversion means an instantaneous passing in case when the batch consists of more than three vehicles. It matches up with the physical meaning of the parameter $\gamma = \lambda\mu_3^{-1}$ because we denoted μ_3 as the passing intensity when the batch length exceeds three.

To sum it up two similar models developed and studied above harmonize with each other. Thus we could consider these models to be correct.

4 Parameters estimate

These theoretical models require to be verified in practice. We should ascertain they conform with the real data. Today there already devised a lot of criteria of this type of verification. However if we want to apply them, we should know the parameters value of the system.

Let n be an all-types batches number observed in the traffic flow. Let us introduce the following notation. Denote m_i $(i = 1, 2, \ldots, r-1)$ as a batches number of length i, and m_r as a batches number of length not less than r $(r \geq 5)$. For the model of the traffic flow with

the unlimited length of autocades the following minimum chi-square parameters estimator were derived:

$$\alpha = \frac{nm_2}{(n-m_1)m_1},$$
$$\beta = \frac{(m_3+m_4+\ldots+m_{r-1}-(r-3)m_r)}{(m_3+2m_4+3m_5+\ldots+(r-3)m_{r-1})}\frac{(n^2-2nm_1+m_1^2-nm_2)}{(n-m_1)m_1},$$
$$\gamma = \frac{m_4+2m_5+3m_6+\ldots+(r-3)m_r}{m_3+2m_4+3m_5+\ldots+(r-3)m_{r-1}}.$$

In case of low traffic intensity we have the following estimates:

$$p = \left[1+\frac{n-m_1-m_2-m_3}{Km_1}+\frac{m_3(Km_1+n-m_1-m_2-m_3)}{Km_1(m_1+m_2)}\right]^{-1},$$
$$q = \frac{\frac{n-m_1-m_2-m_3}{Km_1}}{1+\frac{n-m_1-m_2-m_3}{Km_1}+\frac{m_3(Km_1+n-m_1-m_2-m_3)}{Km_1(m_1+m_2)}},$$
$$s = \frac{\frac{m_3(Km_1+n-m_1-m_2-m_3)}{Km_1(m_1+m_2)}}{1+\frac{n-m_1-m_2-m_3}{Km_1}+\frac{m_3(Km_1+n-m_1-m_2-m_3)}{Km_1(m_1+m_2)}}.$$

In the equalities above $K > 0$ is a certain proportionality coefficient. This coefficient is defined by the traffic conditions and as a rule approximately equals one.

Conclusion

Two problems were solved. The first one is building and study of the probabilistic model of both spatial and temporal traffic flow characteristics when we observe heavy traffic with rather bad road and weather conditions. The second one is building and study of the same model but in case of relatively low fast vehicles density. The correctness of these models was showed up. Parameters of the corresponding distribution laws were estimated.

References

[1] Haight F.A. (1963). *Mathematical theories of traffic flow*. Academic press, New York London.

[2] Bartlett M.S. (1963). The spectral analysis of point processes. *J. Roy. Stat. Soc. Ser. B.* Vol. **25**, pp. 264-296.

Real-Time Studying of Statistic Distributions of Non-Parametric Goodness-of-Fit Tests when Testing Complex Hypotheses

Boris Yu. Lemeshko, Stanislav B. Lemeshko, Andrey P. Rogozhnikov
Department of Applied Mathematics, Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: Lemeshko@fpm.ami.nstu.ru

Abstract

In present work, a "real-time" ability to simulate and research the distributions of tests statistics in the course of testing the complex goodness-of-fit hypothesis (for distributions with estimated parameters) is implemented by the use of parallel computing. It makes it possible to make correct statistical inferences even in those situations when the distribution of the test statistic is unknown (before the testing procedure starts).

Keywords: goodness-of-fit test, composite hypotheses testing, Kolmogorov test, Cramer-Mises-Smirnov test, Anderson-Darling test, methods of statistical simulation.

Introduction

In composite hypotheses testing in the form $H_0 : F(x) \in \{F(x,\theta),\ \theta \in \Theta\}$, when the estimate $\hat{\theta}$ of scalar or vector distribution parameter θ is calculated by the same sample, the nonparametric goodness-of-fit Kolmogorov, ω^2 Cramer-Mises-Smirnov, and Ω^2 Anderson-Darling tests lose their distribution-free property.

The value

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

where $F_n(x)$ is the empirical distribution function, n is the sample size, is used in Kolmogorov test as a distance between the empirical and theoretical laws. When testing hypotheses, this statistic is usually used with Bolshev's correction (Bolshev, [3]) in the form (Bolshev and Smirnov, [4])

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

where $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n} - F(x_i,\theta)\right\}$, $D_n^- = \max\limits_{1\le i\le n}\left\{F(x_i,\theta) - \frac{i-1}{n}\right\}$, n is the sample size, $x_1, x_2, \ldots, x_n$ are sample values in an increasing order. The distribution of statistic (1) in testing simple hypotheses obeys the Kolmogorov distribution law $K(S) = \sum\limits_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}$.

In ω^2 Cramer-Mises-Smirnov test, one uses a statistic in the form

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

and in test of Ω^2 Anderson-Darling type (Anderson and Darling, [1, 2]), the statistic in the form

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

In testing a simple hypothesis, statistic (2) obeys the distribution (see Bolshev and Smirnov, [4]) with the CDF

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

where $I_{-\frac{1}{4}}(\cdot)$, $I_{\frac{1}{4}}(\cdot)$ are modified Bessel functions, $I_\nu(z)=\sum\limits_{k=0}^{\infty}\frac{(z/2)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)}$, $|z|<\infty$, $|\arg z|<\pi$, and statistic (3) obeys the distribution (Bolshev and Smirnov, [4]) with the CDF

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

1 Statistic distributions of the tests in testing composite hypotheses

In composite hypotheses testing, the conditional distribution law of the statistic $G(S|H_0)$ is affected by a number of factors: the form of the observed law $F(x,\theta)$ that corresponds to the true hypothesis H_0; types and number of parameters to be estimated; sometimes, it is a specific value of the parameter (e.g., in case of gamma-distribution, inverse Gaussian law, generalized Weibull distribution, beta-distribution families); the method of parameter estimation.

The paper Kac [13] was a pioneer in investigating statistic distributions of the nonparametric goodness-of-fit tests with composite hypotheses. Then, various approaches to the solution to this problem where used (Darling [6, 7], Durbin [8, 9, 10], Gihman [12], Martynov [27], Pearson and Hartley [30], Stephens [31, 32], Chandra [5], Tyurin [33], Tyurin [34], Dzhaparidze and Nikulin [11], Nikulin [28, 29]).

In our research (Lemeshko and Postovalov [14, 15, 16], Lemeshko and Maklakov [17], Lemeshko [18, 24, 25], Lemeshko and Lemeshko [19, 20, 21], Lemeshko S. [26]), statistic distributions of the nonparametric goodness-of-fit tests are investigated by the methods of statistical simulation, and approximate models of the laws are found for constructed empirical distributions. The most complete list of the constructed models of statistic distributions and tables of percentage points for nonparametric goodness-of-fit tests is provided in Lemeshko [18, 24, 25]. These models and tables are usable when testing complex hypotheses if maximum likelihood estimators were applied.

For a number of distributions often used in applications for description of random variates, distributions of statistics of nonparametric goodness-of-fit tests only have a limited set of dependences: the form of the observed law $F(x,\theta)$ that corresponds to the true hypothesis H_0; types

and number of parameters to be estimated; the method of parameter estimation. In these cases, there are no impediments for studying test statistic distributions by means of statistical simulation and further construction of approximate models for them when testing complex hypothesis (Lemeshko [18, 24, 25]).

Complications arise in case the statistic distributions $G(S\,|H_0\,)$ of nonparametric goodness-of-fit tests depend on a certain value of parameter/parameters of the distribution $F(x,\theta)$ when testing complex hypotheses (for gamma distribution, two-sided exponential law, inverse Gaussian law, generalized Weibull distribution, and beta-distribution families).

The existing dependence on parameters values should not be neglected. For example, in composite hypotheses testing subject to gamma-distribution with the density function $f(x,\theta) = \frac{x^{\theta_0-1}}{\theta_1^{\theta_0}\Gamma(\theta_0)}\exp\left(-\frac{x}{\theta_1}\right)$, limiting statistics distributions of the nonparametric goodness-of-fit tests depend on value of the form parameter θ_0. Figure 1 illustrates the dependence of the Kolmogorov statistic distribution upon the value θ_0 in testing a composite hypothesis only in the case of calculating maximum likelihood estimates (MLE) for the scale parameter of gamma-distribution.

The most serious impediment to a complete solution of the problem of testing complex hypotheses with the use of non-parametric goodness-of-fit tests is that the distributions of the test statistics depend on specific values of shape parameters of the observed laws. In papers (Lemeshko [18, 19, 20, 21, 24, 25]) models of distributions of statistics were obtained for a limited set of combinations of (integer) values of shape parameters (for gamma distribution, two-sided exponential law, inverse Gaussian law, generalized Weibull distribution, and beta-distribution families). It is unrealistic to build the models for an infinite set of combinations of the parameters values.

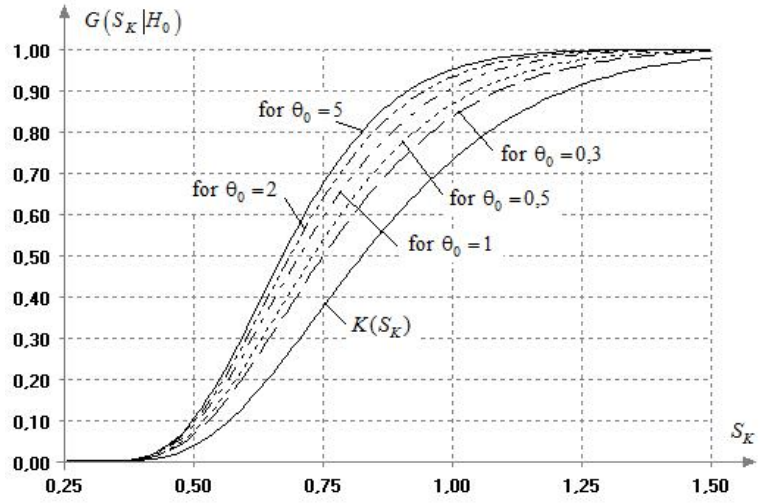


Figure 1: The Kolmogorov statistic (1) distributions for testing composite hypotheses with calculating MLE of scale parameter

In present work, a "real-time" ability to simulate and research the distributions of tests

statistics in the course of testing the complex goodness-of-fit hypothesis (for distributions with estimated parameters) is implemented by the use of parallel computing. It makes it possible to make correct statistical inferences even in those situations when the distribution of the test statistic is unknown (before the testing procedure starts).

2 Testing complex hypotheses in "real-time"

In present work, an approach is proposed and implemented that is based upon authors' evolving software and the use of simulation (Lemeshko [23]). Computational processes in the simulation of statistics of various tests can be parallelized rather easily by the use of available resources of nearby computer network. This makes it possible to dramatically reduce the time required for simulation (studying) an unknown distribution of the statistic $G(S|H_0)$. Statistical analysis is carried out by the following scheme (Fig. 2) in case of the use of nonparametric goodness-of-fit tests for testing complex hypotheses in regard to laws with characteristic dependence of statistic distribution on parameter values. Such an approach was used in Lemeshko [22]. Here the studying of $G(S|H_0)$ is carried out in "real-time" of testing the hypothesis.

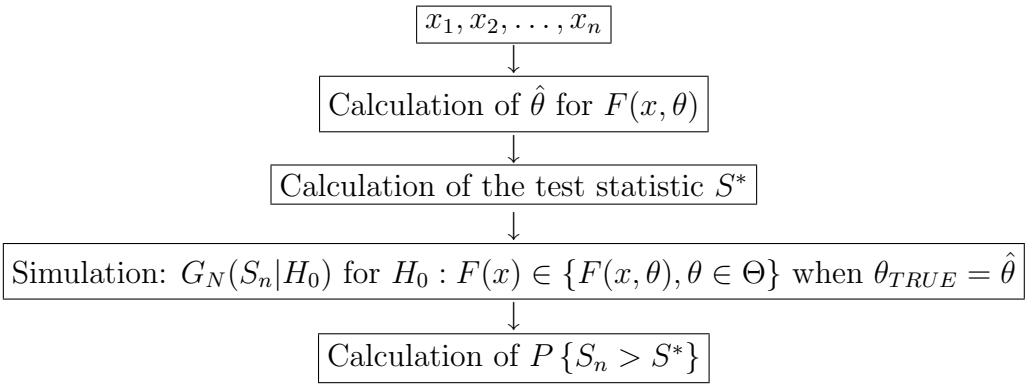


Figure 2: Testing the complex hypothesis $H_0 : F(x) \in \{F(x, \theta), \theta \in \Theta\}$

When testing complex the hypothesis $H_0 : F(x) \in \{F(x, \theta),\ \theta \in \Theta\}$ by an existing sample $x_1, x_2, \ldots, x_n$, the parameter vector estimate $\hat{\theta}$ for the law $F(x, \theta)$ is found in accordance with the selected method. Then, the value of statistic S^* of the goodness-of-fit test in use is calculated in accordance with the estimate $\hat{\theta}$ found. For making an inference on whether to reject or to accept the hypothesis H_0 under test, it's necessary to know the distribution $G(S\,|H_0)$ of the test statistic that corresponds to the parameter value $\hat{\theta}$.

After that, statistical simulation procedure is started that results in obtaining empirical distribution $G_N(S_n\,|H_0)$ of the test statistic for the corresponding sample volume n and the given number of simulations N and $F(x, \theta)$ with the parameters vector $\theta = \hat{\theta}$. One can find an estimate of an achieved significance level $P\{S_n > S^*\}$ or estimates of percentage points by the use of

empirical distribution $G_N(S_n\,|H_0)$. The hypothesis is not rejected if $P\{S_n > S^*\} > \alpha$, where α is a given type I error probability.

The value of N defines the required accuracy of simulation of $G(S_n\,|H_0)$: the greater N the better. However, time spent for simulation increases along with growth of N, therefore, one can determine N during parallelization of simulation process basing upon available computer resources (number of processors and cores) that could be used for the problem under solution.

The probability that elements of $\hat{\theta}$ are integer is zero. Thus, one should cautiously use models and percentage points of test statistic distributions for values of parameters close to integer ones provided in (Lemeshko [18, 19, 20, 21, 24, 25]) as, with interpolation applied, results obtained can be far from the true distribution $G(S\,|H_0)$ with the given $\hat{\theta}$.

Let us consider an example where a complex hypothesis is tested in regard to the inverse Gaussian law with the density function $f(x) = \left(\frac{\theta_1}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{\theta_1(x-\theta_0)^2}{2\theta_0^2 x}\right)$. In this case, distributions $G(S\,|H_0)$ of the nonparametric tests depend on specific values of θ_0 and θ_1.

The sample under analysis is presented in Table 1 ($\theta_0 = \theta_1 = 2.5$). Maximum likelihood estimates of the parameters: $\hat{\theta}_0$=2.4706, $\hat{\theta}_1$=2.5769. In Table 2, values of the tests statistics and achieved significance levels (P-values) obtained by test statistic distributions simulated (in "real time") under different values of N are given.

Table 1: 100 pseudorandom numbers from the inverse Gaussian distribution

0.278	0.633	0.928	1.078	1.334	1.937	2.297	2.630	3.554	5.674
0.312	0.686	0.933	1.080	1.497	1.965	2.362	2.919	3.593	5.989
0.358	0.716	0.936	1.089	1.612	1.991	2.364	2.995	3.948	6.284
0.361	0.776	0.938	1.113	1.671	2.012	2.417	3.002	3.996	6.863
0.362	0.777	0.956	1.119	1.680	2.026	2.467	3.120	4.053	7.580
0.374	0.789	0.996	1.159	1.687	2.027	2.566	3.149	4.141	7.644
0.403	0.796	1.038	1.165	1.731	2.069	2.577	3.166	4.363	7.874
0.590	0.805	1.053	1.166	1.735	2.146	2.599	3.224	4.597	9.236
0.597	0.822	1.060	1.192	1.763	2.210	2.621	3.278	5.022	11.704
0.599	0.849	1.066	1.245	1.898	2.213	2.628	3.528	5.201	20.069

It should be noted, that distributions of nonparametric goodness-of-fit test statistics (1)–(3) for $\hat{\theta}_0$ =2.4706, $\hat{\theta}_1$=2.5769 differ substantially from corresponding distributions under different combinations of integer values of θ_0 and θ_1.

Another example is generalized Weibull distribution with the density function

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

$\theta_0 = \theta_1 = 2.5$ (Table 3). Maximum likelihood estimates of the parameters: $\hat{\theta}_0 = 2.4718$, $\hat{\theta}_1 = 2.5187$. Values of the tests statistics and P-values obtained by simulation are given in Table 4.

Table 2: P-values of the tests for different volumes of simulations (inverse Gaussian distribution)

Test	S^*	$P\{S_n > S^*\}$				
		N=1000	N=5000	N=10000	N=100000	N=1000000
K	0.59361	0.656	0.668	0.668	0.670	0.671
ω^2	0.05380	0.562	0.576	0.574	0.578	0.578
Ω^2	0.35021	0.556	0.570	0.568	0.566	0.566

Table 3: 100 pseudorandom numbers from the generalized Weibull distribution

0.199	0.647	0.932	1.059	1.253	1.648	1.855	2.033	2.482	3.356
0.248	0.703	0.937	1.060	1.367	1.664	1.891	2.180	2.500	3.474
0.311	0.734	0.939	1.067	1.444	1.680	1.892	2.218	2.658	3.583
0.316	0.793	0.941	1.086	1.482	1.692	1.920	2.221	2.679	3.791
0.317	0.794	0.956	1.091	1.488	1.700	1.948	2.279	2.703	4.040
0.333	0.806	0.991	1.122	1.493	1.701	2.000	2.293	2.741	4.062
0.373	0.812	1.025	1.127	1.521	1.725	2.006	2.301	2.835	4.139
0.600	0.821	1.038	1.128	1.523	1.770	2.017	2.328	2.932	4.587
0.608	0.837	1.043	1.147	1.541	1.807	2.029	2.354	3.104	5.351
0.611	0.862	1.049	1.188	1.624	1.808	2.032	2.470	3.174	7.676

Table 4: P-values of the tests for different volumes of simulations (generalized Weibull distribution)

Test	S^*	$P\{S_n > S^*\}$				
		N=1000	N=5000	N=10000	N=100000	N=1000000
K	0.60473	0.670	0.672	0.670	0.673	0.675
ω^2	0.05519	0.596	0.599	0.594	0.597	0.597
Ω^2	0.35462	0.577	0.580	0.580	0.580	0.580

Conclusions

In this work, software is implemented that makes it possible to test complex hypotheses with the use of nonparametric goodness-of-fit test in cases when statistic distributions depend on specific values of the observed distributions.

This research was partially supported by the Russian Foundation for Basic Research (Project № 09-01-00056-a), the Analytical Departmental Targeted Program "Development of the Potential of Institutes of Higher Education" (Project № 2.1.2/11855), and the Federal Targeted Program of the Ministry of Education and Science of the Russian Federation "Academic and Teaching Staff of an Innovative Russia".

References

[1] Anderson T.W., Darling D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, 23, 193-212.

[2] Anderson T.W., Darling D.A. (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.*, 29, 765-769.

[3] Bolshev L.N. (1987). On the question on testing some composite statistical hypotheses. *Theory of Probability and Mathematical Statistics. Selected Works.* Nauka, Moscow, 5-63.

[4] Bolshev L.N., Smirnov N.V. (1983). *Tables of Mathematical Statistics.* Nauka, Moscow. 1983. (in Russian)

[5] Chandra M., Singpurwalla N.D., Stephens M.A. (1981). Statistics for Test of Fit for the Extreme-Value and Weibull Distributions. *J. Am. Statist. Assoc.* Vol. 76, No. 375, pp. 729-731.

[6] Darling D.A. (1955). The Cramer-Smirnov test in the parametric case. *Ann. Math. Statist.*, 26, 1-20.

[7] Darling D.A. (1957). The Cramer-Smirnov test in the parametric case. *Ann. Math. Statist.*, 28, 823-838.

[8] Durbin J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Ann. Statist.*, 1, 279-290.

[9] Durbin J. (1975). Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests of spacings. *Biometrika*, 62, 5-22.

[10] Durbin J. (1976). Kolmogorov-Smirnov Test when Parameters are Estimated. *Lect. Notes Math.* 566, 33-44.

[11] Dzhaparidze K.O., Nikulin M.S. (1982). Probability distribution of the Kolmogorov and omega-square statistics for continuous distributions with shift and scale parameters. *J. Soviet Math.*, 20, 2147-2163.

[12] Gihman I.I. (1953). Some remarks on the consistency criterion of A.N. Kolmogorov. *Dokl. Akad. Nauk SSSR*, 91(4), 715-718.

[13] Kac M., Kiefer J., and Wolfowitz J. (1955). On Tests of Normality and Other Tests of Goodness of Fit Based on Distance Methods. *Ann. Math. Stat.*, 26, 189-211.

[14] Lemeshko B.Yu., Postovalov S.N. (1998). Statistical distributions of nonparametric goodness-of-fit tests as estimated by the sample parameters of experimentally observed laws. *Industrial laboratory (Ind. lab.)*, 64, 3, 197-208. (Consultants Bureau, New York)

[15] Lemeshko B.Yu., Postovalov S.N. (2001). Application of the nonparametric goodness-of-fit Tests in testing composite hypotheses. *Optoelectronics, Instrumentation and Data Processing*, 37, 2, 76-88.

[16] Lemeshko B.Yu., Postovalov S.N. (2002). The nonparametric goodness-of-fit tests about fit with Johnson distributions in testing composite hypotheses. *News of the SB AS HS*, 1(5), 65-74. (in Russian)

[17] Lemeshko B.Yu., Maklakov A.A. (2004). Nonparametric Test in Testing Composite Hypotheses on Goodness of Fit Exponential Family Distributions. *Optoelectronics, Instrumentation and Data Processing*, 40, 3, 3-18.

[18] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N. (2010). Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses. *Comm. Stat. - Theory and Methods*, 39, 3, 460-471.

[19] Lemeshko B.Yu., Lemeshko S.B. (2007). Statistic distributions of the nonparametric goodness-of-fit tests in testing hypotheses relative to beta-distributions. *News of the SB AS HS*, 2, 9, 6-16. (in Russian)

[20] Lemeshko B.Yu., Lemeshko S.B. (2009). Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part 1. *Measurement Techniques*, 52, 6, 555-565.

[21] Lemeshko B.Yu., Lemeshko S.B. (2009). Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II. *Measurement Techniques*, 52, 8, 799-812.

[22] Lemeshko B.Yu., Lemeshko S.B., Akushkina K.A., Nikulin M.S., Saaidia N. (2011). Inverse Gaussian Model and Its Applications in Reliability and Survival Analysis. *Mathematical and Statistical Models and Methods in Reliability. Applications to Medicine, Finance, and*

Quality Control (Edited by V. Rykov, N. Balakrishnan and M. Nikulin), 433-453. Birkhäuser, Boston.

[23] Lemeshko B.Yu., Lemeshko S.B., Chimitova E.V., Postovalov S.N., Rogozhnikov A.P. (2011). Software System for Simulation and Research of Probabilistic Regularities and Statistical Data Analysis in Reliability and Quality Control. *Mathematical and Statistical Models and Methods in Reliability. Applications to Medicine, Finance, and Quality Control* (Edited by V. Rykov, N. Balakrishnan and M. Nikulin), 417-432. Birkhäuser, Boston.

[24] Lemeshko B.Yu., Lemeshko S.B., Nikulin M.S., Saaidia N. (2010). Modeling statistic distributions for nonparametric goodness-of-fit criteria for testing complex hypotheses with respect to the inverse Gaussian law. *Automation and Remote Control*, 71, 7, 1358-1373.

[25] Lemeshko B.Yu., Lemeshko S.B., Akushkina K.A. (2010). Models of statistical distributions of nonparametric goodness-of-fit tests in testing composite hypotheses of the generalized Weibull distribution. *Proceed. 3^{rd} Int. Conf. on Accel. Life Testing (ALT'2010). Clermont-Ferrand, France*, 125-132.

[26] Lemeshko S.B. (2007). Expansion of applied opportunities of some classical methods of mathematical statistics. *The dissertation on competition of a scientific degree of Cand. Tech. Sci.* Novosibirsk State Technical University. Novosibirsk. (in Russian)

[27] Martynov G.V. (1978). *Omega-Square Tests*. Nauka, Moscow. (in Russian)

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

[29] Nikulin M.S. (1992). A variant of the generalized omega-square statistic. *J. Soviet Math.*, 61, 4, 1896-1900.

[30] Pearson E.S., Hartley H.O. (1972). *Biometrica Tables for Statistics. Volume 2.* University Press, Cambridge.

[31] Stephens M.A. (1970). Use of Kolmogorov-Smirnov, Cramer - von Mises and Related Statistics - Without Extensive Table. *J. R. Stat. Soc.*, 32, 115-122.

[32] Stephens M.A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. *J. Am. Statist. Assoc.*, 69, 730-737.

[33] Tyurin Yu.N. (1984). On the Limiting Kolmogorov-Smirnov Statistic Distribution for Composite Hypothesis. *News of the AS USSR. Ser. Math.*, 48, 6, 1314-1343. (in Russian)

[34] Tyurin Yu.N., Savvushkina N.E. (1984). Goodness-of-Fit Test for Weibull-Gnedenko Distribution. *News of the AS USSR. Ser. Techn. Cybernetics*, 3, 109-112. (in Russian)

Application of Variance Homogeneity Tests Under Violation of Normality Assumption

Alisa A. Gorbunova, Boris Yu. Lemeshko
Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: gorbunova.alisa@gmail.com

Abstract

Classical tests for homogeneity of variances (Fisher's, Bartlett's, Cochran's, Hartley's, Neyman-Pearson's, Levene's, modified Levene's, Z-variance, Overall-Woodward modified Z-variance, O'Brien tests) and nonparametric tests (Ansari-Bradley's, Mood's, Siegel-Tukey's, Capon's and Klotz's tests) have been considered. Distributions of classical tests statistics have been investigated under violation of assumption that samples are normally distributed. The comparative analysis of power of classical tests with power of nonparametric tests has been carried out. Tables of percentage points for Cochran's test have been made for distributions which are different from normal. Software, that allows us to apply tests correctly, has been developed.

Keywords: homogeneity of variance test, power of test.

Introduction

Testing for samples homogeneity is frequently of interest in a number of research areas. The question can be about homogeneity of samples distributions, population means or variances. Of course, conclusions in full measure can be made in the first case. However, researcher can be interested in possible deviations in the sample mean values or differences in variances of measurements.

One of the basic assumptions to formulate classical tests for comparing variances is normal distribution of samples. It is well known, that classical tests are very sensitive to departures from normality. Therefore, the application of classical criteria always involves the question of how valid the obtained results are in this particular situation.

In this work classical Bartlett's, Cochran's, Fisher's, Hartley's, O'Brien, Neyman-Pearson's, Levene's, modified Levene's, Z-variance, Overall-Woodward modified Z-variance tests [1, 2] are compared, nonparametric (rank) Ansari-Bradley's, Mood's, Siegel-Tukey's, Capon's and Klotz's tests [1] are considered.

The purpose of our study was to:

- investigate distributions of the statistics for several tests when samples are not normally distributed;
- make a comparative analysis of the criteria power for concrete competing hypotheses;

- give a possibility to apply classical tests when the normality assumption may not be true.

A null hypothesis of equal variances for m samples is $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$ and the alternative hypothesis is $H_1 : \sigma_i^2 \neq \sigma_j^2$, where the inequality holds at least for one pair of subscripts i, j.

Statistical simulation methods and developed software were used to investigate statistics distributions, to calculate percentage points and to estimate tests power for different competing hypotheses. Each test statistic was computed $N = 10^6$ times. In this case an absolute value of the difference between the true law of statistics distribution and a simulated empirical distribution does not exceed 10^{-3}.

Distributions of the statistics were investigated using various distributions, in particular, in the case when simulated samples are in the family of distributions with the density:

$$De(\theta_0) = f(x; \theta_0, \theta_1, \theta_2) = \frac{\theta_0}{2\theta_1 \Gamma(1/\theta_0)} exp\left(-\left(\frac{|x-\theta_2|}{\theta_1}\right)^{\theta_0}\right) \tag{1}$$

using different values of the shape parameter θ_0. This family can be a good model for error distributions of many measuring systems. Special cases of the family $De(\theta_0)$ are the Laplace ($\theta_0 = 1$) and the normal ($\theta_0 = 2$) distributions. This family makes it possible to set various symmetric distributions that differ from the normal distribution. That is a smaller value of the shape parameter θ_0 leads to a "heavier" tails of the distribution.

We also consider chi-square distributions ($df = 6, df = 5$) to approximate skewed distributions where the chi-square distribution with 6 degrees of freedom is less skewed than the one with 5 degrees of freedom.

In the comparative analysis of the tests power we consider the competing hypotheses of the form $H_1 : \sigma_2 = d\sigma_1$ ($d \neq 1$). Some tests can be applied when number of samples is more then two. For these tests we consider hypotheses when different number of samples have another variance.

1 Classical tests of variances homogeneity

1.1 Bartlett's test

The test statistic is:

$$T = M\left(1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{\nu_i} - \frac{1}{N}\right)\right)^{-1}, \tag{2}$$

where $M = N \ln\left(\frac{1}{N}\sum\limits_{i=1}^{k}\nu_i S_i^2\right) - \sum\limits_{i=1}^{k}\nu_i \ln S_i^2$, k - the number of samples, $\nu_i = n_i - 1$, n_i - sample size of ith sample, $N = \sum\limits_{i=1}^{k}\nu_i$, S_i^2 - the unbiased estimate of variance for the ith sample.

If hypothesis H_0 is true, all $\nu_i > 3$ and samples are normally distributed, the statistic (2) is almost independent of the sample size and has approximately χ^2_{k-1} distribution. If samples are

not from a normal population, the distribution of the statistic depends on the sample size and differs from χ^2_{k-1}.

1.2 Cochran's test

The Cochran's test is defined as following:

$$C = \frac{S^2_{\max}}{S_1^2 + S_2^2 + \cdots + S_k^2}, \tag{3}$$

where $S^2_{\max} = \max(S_1^2, S_2^2, \ldots, S_k^2)$, S_i^2 - the unbiased estimate of variance for the ith sample, k - the number of samples.

Distribution of Cochran's test statistic depends on the sample size. The reference litetature gives tables with percentage points for limited number of values n, that are used in hypothesis testing. If the test statistic (3) exceeds the critical value, the null hypothesis is rejected.

1.3 Fisher's test

Fisher's test is used to test hypothesis of variances homogeneity for *two* samples with sample sizes n_1 and n_2. The test statistic has a simple form:

$$F = \frac{S_1^2}{S_2^2}, \tag{4}$$

where S_1^2 and S_2^2 - the unbiased sample variances.

If samples are normally distributed and hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ is true, statistic (4) has the F_{n_1-1,n_2-1}-distribution.

1.4 Hartley's test

Hartley's test is very simple to calculate. Its test statistic is just a ratio between the largest sample variance and the smallest:

$$H = \frac{S^2_{\max}}{S^2_{\min}} \tag{5}$$

where $S^2_{\max} = \max(S_1^2, S_2^2, \ldots, S_k^2)$, $S^2_{\min} = \min(S_1^2, S_2^2, \ldots, S_k^2)$, S_i^2 - the unbiased estimate of variance for the ith sample, k - the number of samples.

One can find in literature table of values created by Hartley. This table evaluates the test statistic with degrees of freedom k and $n-1$ (if $n_1 = n_2 = \cdots = n_k = n$). Reject H_0 if the test statistic (5) is more than critical value, otherwise do not reject H_0.

1.5 Neyman-Pearson's test

The test statistic is defined as ratio between arithmetic mean and geometric mean of variance estimates:

$$P = \frac{\frac{1}{k}\sum\limits_{i=1}^{k} S_i^2}{\left(\prod\limits_{i=1}^{k} S_i^2\right)^{\frac{1}{k}}}, \tag{6}$$

where k - the number of samples.

1.6 Levene's test

If X_{ij}'s represent the raw scores, Levene's test statistic is defined as:

$$L = \frac{(N-k)\sum\limits_{i=1}^{k} n_i \left(\bar{Z}_i - \bar{Z}\right)^2}{(k-1)\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i} \left(Z_{ij} - \bar{Z}_i\right)^2}, \tag{7}$$

where k - the number of samples, n_i - sample size of ith sample, $N = \sum\limits_{i=1}^{k} n_i$ - total sample size, $Z_{ij} = |X_{ij} - \bar{X}_i|$, $\bar{X}_i$ - the mean of the ith sample, $\bar{Z}_i$ - the mean of Z_{ij} for ith sample, $\bar{Z}$ - the overall mean of the Z_{ij}.

In some descriptions of this test it is said that statistic (7) has a $F_{k-1,N-k}$-distribution. Actually *distribution of Levene's test statistic is not Fisher's distribution*! If sample sizes are less than 40, the distribution of the statistic differs greatly grom Fisher's. We must take it into account when using this test.

Levene's test is less sensitive to departures from normality as compared to other classical tests. However it has less power.

1.7 Modified Levene's test

The modified Levene's test is nearly identical to the original Levene's test. Brown and Forsythe suggested using the sample median instead of the mean in computing Z_{ij}. That is $Z_{ij} = |X_{ij} - \tilde{X}_i|$, where $\tilde{X}_i$ - the median of the ith sample.

This test is more robust than original Levene's test.

1.8 Z-variance test

The test statistic is:

$$V = \frac{\sum\limits_{i=1}^{k} Z_i^2}{k-1}, \tag{8}$$

where $Z_i = \sqrt{\frac{c_i(n_i-1)S_i^2}{MSE}} - \sqrt{c_i(n_i-1) - \frac{c_i}{2}}, MSE = \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2}{N-k}$, k - the number of samples, $c_i = 2 + \frac{1}{n_i}$, n_i - sample size for ith sample, S_i^2 - the unbiased estimate of variance for ith sample, $N = \sum\limits_{i=1}^{k} n_i$ - total sample size, X_{ij} - jth observation in ith sample, $\bar{X}_i$ - the mean of ith sample.

If samples are normally distributed and null hypothesis is true, statistic (8) does not depend on sample size and has approximately $F_{k-1,\infty}$-distribution.

1.9 Overall-Woodward modified Z-variance test

As other classical tests Z-variance test is extremely sensitive to departures from normality, so Overall and Woodward conducted a series of studies to determine a c value so that variances of the Z_i would remain stable when samples are not normally distributed. Using regression, they found a c value based on sample size and kurtosis.

The new c value is evaluated as following:

$$c_i = 2.0\left(\frac{2.9 + \frac{0.2}{n_i}}{\bar{K}}\right)^{\frac{1.6\left(n_i - 1.8K + 14.7\right)}{n_i}},$$

where n_i - sample size of the ith sample, $\bar{K}$ - the mean of the kurtosis indices for all samples.

The index of kurtosis is $K_i = \frac{\sum\limits_{j=1}^{n_i} G_{ij}^4}{n_i - 2}$, where $G_{ij} = \frac{X_{ij}-\bar{X}_i}{\sqrt{\frac{n_i-1}{n_i}S_i^2}}$.

Our study has shown that this test remains stable for distributions with different kurtosis indices. However it is not true for skewness indices.

1.10 O'Brien test

Every raw score X_{ij} is transformed using the following formula:

$$V_{ij} = \frac{(n_i - 1.5)n_i(X_{ij} - \bar{X}_i)^2 - 0.5S_i^2(n_i - 1)}{(n_i - 1)(n_i - 2)},$$

where n_i - sample size for ith sample, $\bar{X}_i$ - the mean of ith sample, S_i^2 - the unbiased estimate of variance for ith sample.

After this transformation the mean of V-values will be equal to the variance for original scores, that is $\bar{V_i} = \frac{\sum_{j=1}^{n_i} V_{ij}}{n_i} = S_i^2$.

The O'Brien test statistic will be the F-value computed on applying the usual ANOVA procedure on the transformed scores V_{ij}. When null hypothesis is true, this test statistic has approximately $F_{k-1,N-k}$-distribution.

2 Nonparametric (rank) tests

Nonparametric analogues of variance homogeneity tests are used to test hypotheses that *two* samples with sample sizes n and m are from population with equal dispersion characteristics. To calculate test statistic we use ranks instead of sample values.

2.1 Ansari-Bradley's test

Tha Ansari-Bradley's test statistic is:

$$A = \sum_{i=1}^{m} \left(\frac{m+n+1}{2} - \left| R_i - \frac{m+n+1}{2} \right| \right), \tag{9}$$

where m, n - sample sizes ($m \leq n$), R_i - rank of ith value of sample with sample size m in general variational row.

Discreteness of statistics distribution can be practically neglected when $m, n > 40$.

2.2 Mood's test

The test statistic is defined as following:

$$M = \sum_{i=1}^{m} \left(R_i - \frac{n+m+1}{2} \right)^2, \tag{10}$$

where m, n - sample sizes ($m \leq n$), R_i - rank of ith value of sample with sample size m in general variational row.

Discreteness of statistics distribution can be practically neglected when $m, n > 20$.

2.3 Siegel-Tukey's test

The general variational row $X_1 \leq X_2 \leq \cdots \leq X_N$ ($N = n + m$) is transformed into sequence:

$$X_1, X_N, X_{N-1}, X_2, X_3, X_{N-2}, X_{N-3}, X_4, X_5, \ldots$$

When $m \leq n$ the test statistic is defined as:

$$W = \sum_{i=1}^{m} R_i, \tag{11}$$

where R_i - rank of ith value of sample with sample size m in transformed row.

Discreteness of statistics distribution can be practically neglected when $m, n > 30$.

2.4 Capon's test

Capon's test statistic is:

$$K = \sum_{i=1}^{m} a_{m+n}(R_i), \tag{12}$$

where m, n - sample sizes ($m \leq n$), R_i - rank of ith value of sample with sample size m in general variational row, $a_i(j)$ - the mean value of square of jth order statistic in sample with sample size i from standard normal distribution.

2.5 Klotz's test

The test statistic is defined as:

$$Q = \sum_{i=1}^{m} u^2_{\frac{R_i}{m+n+1}}, \tag{13}$$

where m, n - sample sizes ($m \leq n$), R_i - rank of ith value of sample with sample size m in general variational row, u_γ - γ-quantile of standard normal distribution.

3 Comparative analysis of power

At the given probability of a type I error α (to reject the null hypothesis when it is true) it is possible to judge about the advantages of the test by the value of power $1 - \beta$, where β - probability of type II error (not to reject the null hypothesis when alternative is true).

The study of power of classical tests for several competing hypotheses $H_1 : \sigma_2 = d\sigma_1$ ($d \neq 1$) has shown that Bartlett's, Cochran's, Hartley's, Fisher's, Neyman-Pearson's and Z-variance tests have equal power for two normal samples and Levene's test power is much less in this case.

As for non-normal distributions, for example, family of distributions with density (1), Bartlett's, Cochran's, Hartley's, Fisher's, Neyman-Pearson's and Z-variance tests remain equal in power, and Levene's test power is also much less. However, for heavy-tailed (for example, the Laplace distribution) and skewed distributions Levene's test is more powerful than the others. Furthermore modified Levene's test outperformed the original test in this case.

Bartlett's, Cochran's, Hartley's, Levene's, Neyman-Pearson's, O'Brien, Z-variance and modified Z-variance tests can be applied when number of samples $k > 2$. In such situations the power

of these tests is different. If $k > 2$ and normality assumption is true, these tests can be ordered according to the decrease of power in the following way:

$$\text{Cochran's} \succ \text{O'Brien} \succ \text{Z-variance} \succ \text{Bartlett's, Neyman-Pearson's} \succ \text{modified Z-variance} \succ \text{Hartley's} \succ \text{Levene's, modified Levene's.}$$

The preference order also remains in case of violation of a normality assumption. When samples are from heavy-tailed or skewed distributions, this preference order changes. For example, in the case of the Laplace distribution Levene's test has a greater power. Also modified Levene's test is more powerful than the original one.

However if number of samples with smaller value of variance is less than number of samples with greater value, power of Cochran's test significantly goes down. So in this case we should prefer O'Brien, Z-variance, Bartlett's or Neyman-Pearson's test.

The study of the nonparametric criteria power has shown that Mood's test power is the highest. And other nonparametric tests, as Siegel-Tukey's, Ansari-Bradley's, Capon's and Klotz's have practically equal power. But for skewed distributions all nonparametric tests are biased (power of test is less than significance level).

4 Cochran's test for non-normal distributions

The main and valid reason for using nonparametric tests is based on the fact that these test statistics are distribution-free. But this is true if both samples are from the same population. If samples are not identically distributed, *nonparametric tests depend on both sample laws and even the order in which these laws are used*.

Also classical tests have a great advantage in power over nonparametric tests. This advantage remains when samples are not normally distributed. Therefore, there is every reason to study distributions of *classical* criteria for testing variances homogeneity. To study distributions means to develop distribution models or tables of percentage points. It should be done for non-normal distributions mostly used in practice. Among the tests studied Cochran's test seems to be the most suitable for this purpose.

Tables of upper percentage points (1%, 5%, 10%) for Cochran's test were made using statistical simulation for the number of samples $m = 2 \div 5$ when simulated samples were taken from an exponential family of distributions (1) with shape parameter $\theta_0 = 1, 2, 3, 4, 5$. The results obtained can be used in situations when distribution from an exponential family (1) with an appropriate parameter θ_0 is a good model for the observed variables. Computed percentage points expand possibilities to apply Cochran's test.

5 Software for testing hypotheses

It is impossible to develop distribution models for all distributions and sample sizes. So we have developed software that allows us to correctly apply tests for comparing variances when

samples are from any distributions. We can choose any distribution from the list and simulate a distribution of the statistic. Also we can set any size of simulated samples of statistics according to required precision.

Then we define a p-value using simulated statistic distribution. Simulation process is done using parallel computing, so speed of simulation depends on number of CPU cores and takes not much time to make correct decision when testing the hypothesis of equal variances.

References

[1] Kobzar A.I. (2006). *Applied mathematical statistics. For engineers and scientists* (in Russian). FIZMATLIT Publishing House, Moscow.

[2] Lee H.B., Katz G.S., Restori A.F. (2010). A Monte Carlo study of seven homogeneity of variance tests. *Journal of Mathematics and Statistics*. Vol. **6(3)**, pp. 359-366.

On Random-Number Generator of Given Distribution

V. F. Pervushin, N. A. Sergeeva, A. V. Strelnikov
Siberian state aerospace university,
Siberian federal university, Krasnoyarsk, Russia
e-mail: sergena@list.ru

Abstract

The algorithm generating the sample of random numbers of defined distribution and numerical characteristics (expectation, dispersion, etc.) is considered. The high accuracy of algorithm working on sample modeling is shown. The general principle of sample generation allows assigning the given approach in a category of random-numbers generation algorithms that combining such algorithms in group of "precision random-numbers generator".

Introduction

Theme of development, being discussed, has been suggested by the stochastic modeling problem, which importance for practice recently increases. If the model is constructed adequately to physical process then experiments, concerned with object or process behavior studying varying the input parameters, external effects influence, may be conducted. Also it is important to study more thoroughly physical and mathematical connections in the process to find new regularities and to establish an interference of parameters and object variables, to fill up qualitative knowledge of process, to fill the lost or inexact data. It is hard and sometimes even impossible to study the real object. In such a case the importance of imitating modeling becomes extremely significant. The model gives an opportunity to implement various situations, including atypical for the given process or object, to analyze its behavior changing parameters of an input or the variables, the researcher is interested in.

1 General performance of modeling problem

The general scheme of process accepted in the identification theory is resulted lower [1].

Here designations are accepted: $x(t)$ – the output vector variable of the process, $u(t)$ – the control vector effect, $\mu(t)$ – the input vector variable of the process, $\omega^i(t): \quad i = 1, 2, \ldots, k$ – the variables of the process controlled during functioning of the object, $\xi(t)$ – the stochastic vector effect, t concluded parenthesis – the continuous time, H with sign above – the communication channels corresponding to different variables, including control devices, instruments for measurement of observable variables, the sign t in the bottom of variables (x, ω, u, μ) – means discrete time.

Control of variables (x, ω, u, μ) is carried out through time interval Δt, i.e. $x_i, \omega_i^1, \ldots, \omega_i^k, u_i, \mu_i, \; i = \overline{1, s}$ – measure sample of process variable

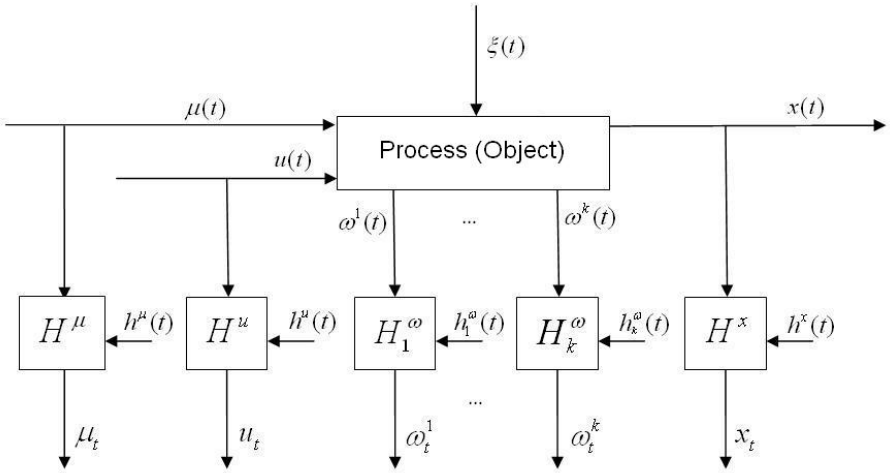


Figure 1: Scheme of process

$\left(x_1, \omega_1^1, \ldots, \omega_1^k, u_1, \mu_1\right)$, $\left(x_2, \omega_2^1, \ldots, \omega_2^k, u_2, \mu_2\right), \ldots$, $\left(x_s, \omega_s^1, \ldots, \omega_s^k, u_s, \mu_s\right)$, ... , s – amount of sampling, $h(t)$ with sign above – random noise of corresponding process variables. The researches stated in given article, have a direct bearing on values of $h(t)$. It will be a question of random noise modeling having the definite distribution law and numerical characteristics.

2 On modeling stochastic processes

In the past years there were many heuristic algorithms, first of all because fresh wordings of problems do not give in to strict mathematical statement. And it means, more often, absence of analytical synthesis procedure of those or other algorithms, proofs of corresponding convergence theorems which presence was considered earlier as a measure of the validity, legitimacy of the further actions. Thereupon it is necessary to consider a method of statistical modeling as a conclusive stage, instead of an illustration of those or other algorithms functioning. The last essentially increases requirements to carrying out of similar research. Here deciding is possibility of single cycle reiteration of numerical experiments by other researches. On this way necessity of work with the random noise, distributed under the concrete law, is important. In practice existing random-numbers generators is rather conditionally named corresponding to stated distribution laws. The deviation on samples with small amounts is especially appreciable. All classical generation algorithms [2] are right, first of all, on enough great amount of sample. The offered approach of random noises modeling, operating in communication channels of input-output variables of object, is oriented on continuous processes research.

At research of many stochastic processes assume that the probability density of random factors (noise) submits to any law. Then there is analytical a research, in particular will specify,

that the error of electric and not electric values measurement submits in the electric ways to the normal law. In problems of the reliability theory use distribution of Weibull (intensity of refusal), lognormal distribution is characteristic for many concrete physical and socio-economic situations (the size and weight of the particles formed at crushing; a salary of the worker; the sizes of space formations; durability of the product working in a mode of deterioration and ageing, etc.) [3].

3 General method of random-numbers generation

Let's consider generation trick of points set for one-dimensional random variable.

Let it is necessary to generate statistically independent sample of a random variable X with the set probability density and distribution parameters $f(x, m_x, \sigma_x, \ldots)$. Let's designate through n – quantity of generated points (amount of the future sample). Following step it is necessary to set an interval $[a, b]$, in which sample is need to generate.

Sample of generated points should cover most full all area of possible values, but for the clear reasons it is necessary to cut off distribution "tails" where the probability of values X drawing becomes small. In this case it is necessary to define this order smallness and, for example, to accept to sample reception that interval, in which $f(x, m_x, \sigma_x, \ldots)$ not less some value. Let's define this value proportionally $f_{\max}(x, m_x, \sigma_x, \ldots)$:

$$f(x = a, m_x, \sigma_{x,}, \ldots) = f(x = b, m_x, \sigma_{x,}, \ldots) = 0.01 \cdot f_{\max}(x, m_x, \sigma_{x,}, \ldots), \tag{1}$$

for distributions which begin from $x = 0$, accordingly to accept $a = 0$, and b to define from the above-stated reasons. If it is not possible to define border of intervals analytically, it is possible to apply any numerical procedures for the decision of the nonlinear equation. If it is not obviously possible for making analytically, using numerical procedures may be demanded for a finding of distribution extremum.

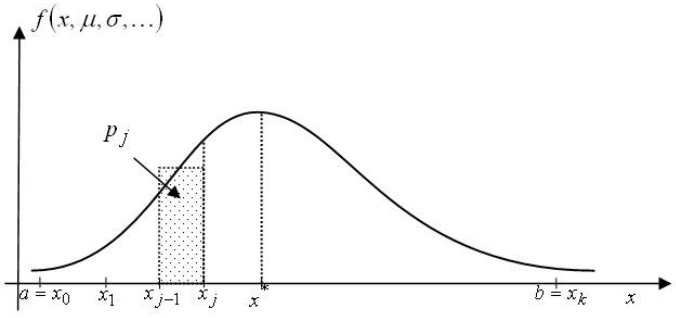


Figure 2: Graphic interpretation of algorithm

All intervals we will break on equal subinterval $[a = x_0, x_1, \ldots, x_k = b]$. The quantity of intervals will be equal k and is set by the user. In everyone subinterval average value of density

function under the formula is set:

$$\frac{f(x_{j-1}) + f(x_j)}{2} = f(\tilde{x}_j), \qquad \tilde{x}_j \in [x_{j-1}, x_j]. \tag{2}$$

The probability of hit of a random variable X in j-th an interval is the area, limited to density function on a piece $[x_{j-1}, x_j]$, we will designate it through p_j:

$$P(x_{j-1} \leq X \leq x_j) = p_j = \frac{f(x_{j-1}) + f(x_j)}{2} \cdot (x_j - x_{j-1}). \tag{3}$$

Based on the law of large numbers with growth of sample amount n random variable values, random event coming frequency approach to probability of this event then it is possible to accept:

$$p_j = \frac{n_j}{n}, \tag{4}$$

where n_j – quantity of random variable values hits X in an interval $[x_{j-1}, x_j]$ from total amount of values.

Then it is possible to define quantity of points which it is necessary to throw in an interval $[x_{j-1}, x_j]$ that the generated sample corresponded to the set distribution law. The quantity of points will be equal:

$$n_j = [p_j \cdot n], \tag{5}$$

which snatch in an interval $[x_{j-1}, x_j]$, for example, under the uniform law or any built in generation function of a concrete programming language. In this case quality of the built in random-numbers generator becomes noncritical.

Set of the points, received in the way described above on all intervals, form random variable sample X, distributed on the given law $f(x, m_x, \sigma_x, \ldots)$, and having all demanded values of parameters. The choice of points from the presented set of values can be executed also by any built in extraction function of data from array of values.

4 Numerical researches of random-numbers generators under given law

Some results of numerical modeling are resulted lower according to the offered algorithm and estimations of random variable characteristics on the generated sample.

The random variable distributed under the normal law $N(\mu, \sigma^2)$, $\mu = 0$, $\sigma^2 = 1$ has been considered. Density and distribution law of the normal law:

$$f(X, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(X-\mu)^2}{2\sigma^2}\right), F(X, \mu, \sigma) = \int_{-\infty}^{X} f(t, \mu, \sigma)\, dt \tag{6}$$

The distribution density of this random variable and also its estimation, constructed on the basis of independent measurements sample is resulted lower [4].

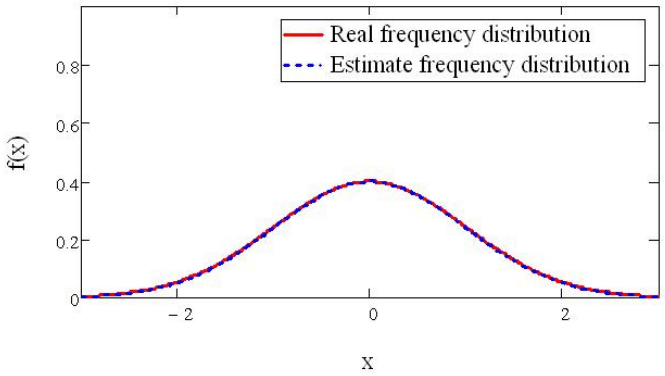


Figure 3: Distribution density of the normal law and its estimation

The estimation distribution density, resulted on fig. 3, is constructed on the basis of sample with amount $s = 500$, quantity subintervals $k = 30$. Selective estimations of distribution law's parameters, averaged on 1000 experiments, have made $\widehat{\mu} = -0.002$, $\widehat{\sigma}^2 = 0.977$. On it and the subsequent figures curves of distribution function and its estimation are not allocated because they coincide.

Also the random variable distributed on lognormal law $X \sim LogN(\mu, \sigma^2)$, $\mu = 0$, $\sigma^2 = 0.5$ has been considered. Density and distribution law of the lognormal law:

$$f(X, \mu, \sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}X\sigma} \exp\left(\frac{-(\ln X - \mu)^2}{2\sigma^2}\right), X > 0; \\ 0, X \leq 0, \end{cases} \quad F(X, \mu, \sigma) = \int_{-\infty}^{X} f(t, \mu, \sigma)\, dt \qquad (7)$$

The distribution density of this random variable, and also its estimation, constructed on the basis of independent measurements sample is resulted lower [4, 5].

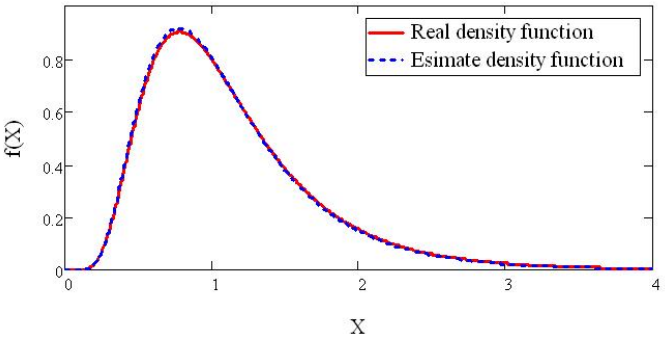


Figure 4: Distribution density of the lognormal law and its estimation

The estimation distribution density, resulted on fig. 4, is constructed on the basis of sample

with amount $s = 500$, quantity subintervals $k = 30$. Selective estimations of distribution law's parameters, averaged on 1000 experiments, have made $\widehat{\mu} = -0.011$, $\widehat{\sigma}^2 = 0.497$.

Further the random variable distributed under the two-parametrical law of Weibull [4, 5] is considered. The law and density of distribution look like:

$$F(X,a,b) = \begin{cases} 1 - e^{-\left(\frac{X}{b}\right)^a}, X \geq 0; \\ 0, X < 0, \end{cases} \quad f(X,a,b) = \begin{cases} \frac{a}{b}\left(\frac{X}{b}\right)^{a-1} e^{-\left(\frac{X}{b}\right)^a}, X \geq 0; \\ 0, X < 0, \end{cases} \tag{8}$$

where a – parameter of form, b – parameter of scale.

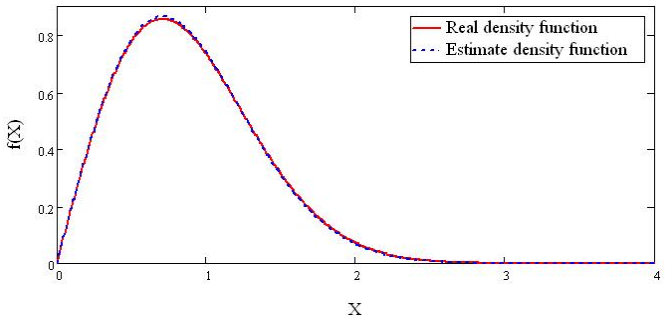


Figure 5: Distribution density of the Weibull's law and its estimation

In the experiment, which results are represented on fig. 5, the amount of generated sample is equal $s = 500$, quantity subintervals $k = 30$. True values of parameters $a = 2$, $b = 1$, and their selective values $\widehat{a} = 2.012$, $\widehat{b} = 0.994$.

Conclusion

In the present article the approach to random variables generation under the given law with the specified parameters is offered. Numerical researches for Weibull's, normal and lognormal laws of random variable distribution are conducted. In a problem of discretely-continuous processes statistical modeling there is a necessity to use methods of random variables generation under given law. Also in article results of numerical researches of the offered method are represented.

References

[1] Medvedev A.V. (2010). Nonparametric systems theory. Modeling. *Vestnik SibSAU*. Vol. **4**, pp. 303-307.

[2] Ermakov S.M., Michajlov G.A. (1976). *Statistical modeling course*. Nauka, Moscow.

[3] Fadeeva L.N. (2006). *Theory of probability and mathematical statistic*. Eksmo, Moscow.

[4] Probability and mathematical statistic. Encyclopedia. / Edit. Prohorova Ju.V. (2003). Great Russian Encyclopedia, Moscow.

[5] Pervushin V.F., Sergeeva N.A., Strelnikov A.V. (2010) Precision random-numbers generator. *Vestnik SibSAU*. Vol. **5**, pp. 86-91.

Simulation Study for the NRR Chi-Square Test of Goodness-of-Fit for Censored Data

Ekaterina V. Chimitova, Angelika O. Tsivinskaya
Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: chimitova@ngs.ru, viento.acariciador@gmail.com

Abstract

This paper presents the investigation results for the Nikulin-Rao-Robson (NRR) chi-square type test. The distributions of the NRR test statistic have been investigated by means of computer simulation technique depending on the sample size, censoring distribution, proportion of censoring and number of intervals. Simulation studies of chi-square test statistic distributions have been shown for type I, II and random censoring. Using computer simulation we have studied the power of the NRR test for close competing hypotheses.

Keywords: Nikulin-Rao-Robson chi-square test, goodness-of-fit, censored samples, test power, Monte Carlo simulations

Introduction

In research area related to reliability and survival analysis lifetimes are typically right censored [7]. The observed data are usually presented as $(X_1, \delta_1), \ldots, (X_n, \delta_n)$, where $\delta_i = 1$ if X_i is an observed lifetime, and $\delta_i = 0$ if X_i is a censoring time which means that lifetime of i-th individual is greater than X_i. There are various types of right-censoring mechanism:

- If individuals are observed at a predetermined time, then the censoring is called type I censoring.
- If a life test is terminated whenever a specified number of failures have occurred, it is called type II censoring.
- Let lifetime T and censoring time C are independent random variables from distribution functions $F(t)$ and $F^C(t)$ respectively. All lifetimes and censoring times are assumed mutually independent, and it is supposed that $F^C(t)$ does not depend on any of the parameters of $F(t)$. So, $X_i = \min(T_i, C_i)$ and $\delta_i = 1\{T_i \leq C_i\}$, it is called independent random censoring.

In this paper we consider the problem of testing composite hypothesis of the kind:

$$H_0 : F(t) \in \{F(t, \theta), \theta \in \Theta\},$$

in which parameters are unknown and should be estimated from a censored sample.

Various goodness-of-fit tests were developed for right-censored samples. Hollander and Peña in 1992 [5] developed a chi-square goodness-of-fit test for simple hypotheses with censored data.

Version of a chi-square test for censored samples was published by Habib and Thomas in 1986 [3]. Modifications of the chi-square test for a composite hypothesis with censored data were proposed by Akritas in 1988 [1] and J. H. Kim in 1993 [6]. Test statistic distributions for all these chi-square type tests for randomly censored data are influenced by distribution of censoring times. They are based on idea of comparing observed and expected numbers of failures in grouping intervals, which was also developed by Hjort [4].

The NRR chi-square type test statistic suggested by Nikulin in 2010 [2] has the limiting distribution that doesn't depend on the proportion of censoring and censoring type. The NRR chi-square test can be used for the right censored and truncated samples. No simulation studies have been published to confirm obtained asymptotical results in case of limited sample sizes yet.

The objective of this paper is to investigate the NRR statistic distributions under null hypothesis with Monte Carlo simulations for different censoring types in case of small sample sizes. Also we study the power of the NRR test for close competing hypotheses.

1 Nikulin-Rao-Robson chi-square test

For chi-square type tests the observed interval $[0, \tau]$ is divided into k grouping intervals $I_1, \ldots, I_k$, where $I_j = (a_{j-1}, a_j]$, $a_0 = 0$, $a_k = \tau$, $j = 1, \ldots, k$. As well as other chi-square type tests the NRR test is based on the vector of differences between the numbers of observed and "expected" failures in the grouping intervals. Under a null hypothesis H_0 this test is based on the statistic [2]

$$Y^2 = \sum_{j=1}^{k} \frac{(U_j - e_j)^2}{U_j} + Q,$$

where $U_j = \sum\limits_{i:X_i \in I_j}^{n} \delta_i$ is the number of observed failures in the j-th interval,
e_j is the "expected" number of failures in the j-th interval.

$$Q = W^T \hat{G}^{-} W,$$

where

$$W = \hat{C}\hat{A}^{-1}Z = (W_1, \ldots, W_m)^T,\ W_l = \sum_{j=1}^{k} \hat{C}_{lj}\hat{A}_j^{-1}Z_j.$$

$$\hat{G} = \hat{i} - \hat{C}\hat{A}^{-1}\hat{C}^T = [\hat{g}_{ll'}]_{m\times m},\ \hat{g}_{ll'} = \hat{i}_{ll'} - \sum_{j=1}^{k} \hat{C}_{lj}\hat{C}_{l'j}\hat{A}_j^{-1},$$

m is the number of estimated parameters.

$$Z = (Z_1, ..., Z_k)^T,\ Z_j = \tfrac{1}{\sqrt{n}}\,(U_j - e_j).$$

$\hat{A}$ is the diagonal $k \times k$ matrix with diagonal elements

$$\hat{A}_j = \frac{U_j}{n}.$$

$$\hat{i}_{ll'} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\partial \ln \lambda\left(X_i,\hat{\theta}\right)}{\partial \theta_l} \frac{\partial \ln \lambda\left(X_i,\hat{\theta}\right)}{\partial \theta_{l'}},$$

$$\hat{C}_{lj} = \frac{1}{n}\sum_{i:X_i \in I_j}^{n} \delta_i \frac{\partial}{\partial \theta_l} \ln \lambda\left(X_i,\hat{\theta}\right),\ j = 1,\ldots,k;\ l,l' = 1,\ldots,m,$$

where $\lambda(t)$ is the hazard rate function and $\hat{\theta}$ is the maximum likelihood estimate (MLE) of unknown parameter obtained by the original ungrouped censored sample. The limiting distribution of the NRR test statistic is χ^2_r, $r = rank(\hat{V}^-)$. $\hat{V}^-$ is calculated as follows

$$\hat{V}^- = \hat{A}^{-1} + \hat{A}^{-1}\hat{C}^T\hat{G}^-\hat{C}\hat{A}^{-1}.$$

So, the hypothesis is rejected with significance level α if $Y^2 > \chi^2_\alpha(r)$.

2 Choice of grouping intervals

In [2] it was recommended to calculate a_j as random data functions. The idea is to divide the interval $[0,\tau]$ into k intervals with equal expected numbers of failures. Thus, a_j are calculated as follows

$$E_k = \sum_{i=1}^{n} \Lambda\left(X_i,\hat{\theta}\right),$$
$$E_j = \frac{j}{k}E_k,\ j = 1,\ldots,k,$$
$$b_i = (n-i)\,\Lambda\left(X_{(i)},\hat{\theta}\right) + \sum_{l=1}^{i} \Lambda\left(X_{(i)},\hat{\theta}\right).$$

If $E_j \in [b_{i-1}, b_i]$, $j = 1,\ldots,k-1$ then

$$\hat{a}_j = \Lambda^{-1}\left(\left(E_j - \sum_{l=1}^{i} \Lambda\left(X_{(l)},\hat{\theta}\right)\right)/\left(n-i+1\right),\hat{\theta}\right),\ \hat{a}_k = X_{(n)},$$

where Λ^{-1} is the inverse of the function Λ. We have $0 < \hat{a}_1 < \hat{a}_2 < ... < \hat{a}_k = \tau$. Under this choice of the intervals $e_j = E_k/k$ for any j.

It is important to note that in the case of small sample size this grouping method can result in such a splitting into intervals, that some intervals do not contain any failure ($U_j = 0$). In simulation studies such samples were eliminated.

3 Simulation studies: NRR statistic distribution

By means of computer simulation technique we have investigated the distributions of the NRR test statistic for different sample sizes n, censoring distributions $F^C(t)$ and proportions of censoring $c = 10, 20, \ldots, 80\%$. Empirical distributions of the NRR test statistic have been simulated by the following algorithm:

1. Generate a complete sample $T_1, \ldots T_n$ from the distribution under null hypothesis H_0 being tested: $T_i = F_0^{-1}(\xi_i)$, $i = 1, \ldots, n$, where $\xi_i \sim$Uniform$(0, 1)$.

2. Transform the complete sample $T_1, \ldots T_n$ to censored sample $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ according to one of censoring schemes.

   (a) For type I censoring: set the value of censoring time t^c,
   so $X_i = \min(T_i, t^c)$ and $\delta_i = 1\{T_i \leq t^c\}$, $i = 1, \ldots, n$.

   (b) For type II censoring: set the number of complete observations r,
   so $X_i = \min(T_i, T_{(r)})$ and $\delta_i = 1\{T_i \leq T_{(r)}\}$, $i = 1, \ldots, n$, where $T_{(r)}$ is r-th ordered statistic.

   (c) For random censoring: generate a sample $C_1, \ldots C_n$ from the distribution of censoring times $F^C(t)$,
   so $X_i = \min(T_i, C_i)$ and $\delta_i = 1\{T_i \leq C_i\}$, $i = 1, \ldots, n$.

3. Estimate unknown parameters of the distribution $F_0(t)$ from the obtained censored sample $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ by the maximum likelihood method.

4. Calculate the values of boundary points $a_0, \ldots, a_k$ as suggested in Section 2 for given number of grouping intervals k .

5. Calculate the NRR test statistic S as suggested in Section 1.

6. Repeating steps 1-5 N times we obtain a sample of NRR statistics $S_1, \ldots, S_N$, for which an empirical distribution function (e.d.f) $G(S|H_0)$ is constructed.

In this Section simulation results for the NRR statistic distributions are presented for $N = 16600$. We simulated samples under null hypothesis from the Weibull distribution with the shape parameter equal to 2 and the scale parameter equal to 2. In case of random censoring we need to specify the distribution of censoring times. We have chosen two different families of distributions for censoring times: Beta-I distribution family with probability density function

$$f^C(t;\theta) = \text{Beta-I}(\theta_1, \theta_2, \theta_3) = \frac{1}{\theta_3 \mathrm{B}(\theta_1, \theta_2)} \left(\frac{t}{\theta_3}\right)^{\theta_1 - 1} \left(1 - \frac{t}{\theta_3}\right)^{\theta_1 - 1}$$

and the Weibull distribution family with

$$f^C(t;\theta) = \text{Weibull}(\theta_1, \theta_2) = \frac{\theta_1}{t} \left(\frac{t}{\theta_2}\right)^{\theta_1} \exp\left(-\left(\frac{t}{\theta_2}\right)^{\theta_1}\right).$$

Censoring distributions are given in Table 1. The distribution parameters were adjusted so that the average proportion of censoring under considered null hypothesis would be equal to $10, 20, \ldots, 80\%$. In the case of Beta-I distributions censored observations appear in the variational series of a censored sample approximately uniformly, as opposed to the Weibull distributions, for which censored observations appear generally at the end of the variational series.

Table 1: Censoring distributions

Proportion of censoring, %	Censoring distributions	
10	Beta-I(1.81,1,7)	Weibull(6.88,3.44)
20	Beta-I(1.19,1,7)	Weibull(5.74,2.87)
30	Beta-I(1,1.24,7)	Weibull(4.96,2.48)
40	Beta-I(1,1.83,7)	Weibull(4.32,2.16)
50	Beta-I(1,2.58,7)	Weibull(3.74,1.87)
60	Beta-I(1,3.58,7)	Weibull(3.18,1.59)
70	Beta-I(1,5.01,7)	Weibull(2.62,1.31)
80	Beta-I(1,7.36,7)	Weibull(2,1)

Figure 1 presents the NRR statistic distributions for randomly censored samples, when censoring distribution is from the Beta-I family and the censoring proportion $c = 10, 20, \ldots, 80\%$, according to Table 1. The sample size $n = 100$ and the number of grouping intervals $k = 5$. As you can see from the figure the test statistic distributions differ from the limiting χ^2_4 distribution, especially for censoring rate over 50%. And the greater the censoring degree, the greater the distance from empirical distributions to the limiting χ^2-distribution. Although not included, the NRR statistic distributions for I and II censoring types have similar dependency in relation to the censoring proportion.

Figure 2 illustrates the empirical statistic distributions for different sample sizes and the corresponding limiting distribution. The empirical distributions were obtained by randomly censored samples with censoring distribution from the Beta-I distribution family, the proportion of censoring $c = 40\%$ and the number of grouping intervals $k = 5$. As you can see from the figure, the obtained empirical distribution of the NRR statistic is rather close to the limit distribution χ^2_4 for sample size $n = 300$. So, if the censoring degree is large, then the larger sample size is needed in order to ensure the closeness of empirical distributions to the limiting distribution of considered statistic.

In Figure 3 you can see the empirical distributions of NRR statistic which were simulated by randomly censored samples with different distributions of censoring times for the same proportion of censoring $c = 20\%$. The sample size $n = 100$ and number of intervals $k = 4$. We may conclude that in case of small sample sizes the NRR test statistic distributions depend on not only on the censoring degree, but also on the distribution of censored observations. And the convergence of

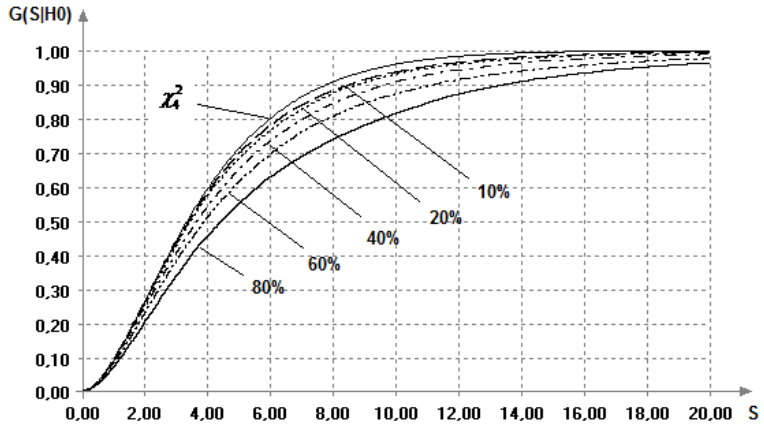


Figure 1: NRR test statistic distribution for different proportions of censoring

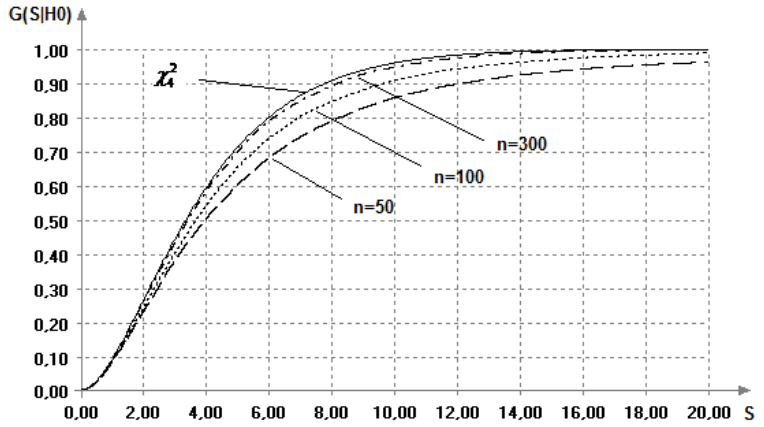


Figure 2: NRR test statistic distributions for different sample sizes

statistic distributions to the limiting χ^2-distribution in the case of censoring distribution from the Weibull family (when censored observations are mainly located at the end of variational series) turned out to be slower than in the case of the Beta-I distribution of censoring times.

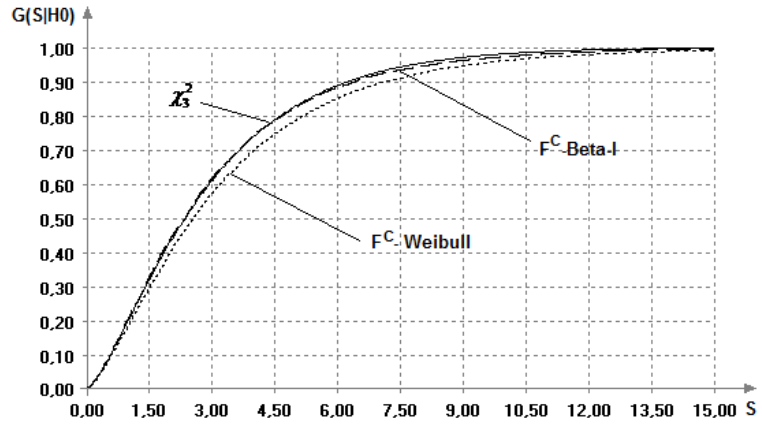


Figure 3: NRR test statistic distributions for different distribution families of censoring times

4 Power of the NRR chi-square test

It is interesting to consider the test power for close competing hypotheses and to compare the NRR test by power with other nonparametric goodness-of-fit tests. In [8] the modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for censored samples of type I and II were investigated by means of computer simulation technique. The power of these tests was estimated in particular for following pair of competing hypotheses:

$$H_0\text{: Weibull distribution - } f\left(t;\theta\right)=\frac{\theta_1}{t}\left(\frac{t}{\theta_2}\right)^{\theta_1}\exp\left(-\left(\frac{t}{\theta_2}\right)^{\theta_1}\right)$$

and

$$H_1\text{: Gamma distribution - } f\left(t;\theta\right)=\frac{1}{\theta_2\Gamma(\theta_1)}\left(\frac{t}{\theta_2}\right)^{\theta_1-1}\exp\left\{-\frac{t}{\theta_2}\right\}.$$

Let us consider the power of the NRR test on the same pair of competing hypotheses. In this study we simulated the NRR test statistic distributions $G(S|H_0)$ and $G(S|H_1)$ for type II censored samples of the size $n = 300$. The number of simulated samples $N = 10^5$. Unknown parameters were estimated with the maximum likelihood method by original censored samples. The NRR test power results for the significance level $\alpha = 0.1$ are given in Table 2. The power of Kolmogorov (K), Cramer-von Mises-Smirnov (C-M-S) and Anderson-Darling (A-D) tests, which were obtained in paper [8] for the same sample size and censoring scheme, are given for the comparison.

Table 2: Power of goodness-of-fit tests

Goodness-of-fit test	Proportion of censoring							
	10%	20%	30%	40%	50%	60%	70%	80%
K	0.51	0.26	0.16	0.14	0.17	0.26	0.43	0.38
C-M-S	0.46	0.18	0.12	0.13	0.15	0.21	0.31	0.58
A-D	0.46	0.18	0.12	0.13	0.14	0.19	0.30	0.68
NRR, k=4	0.53	0.43	0.34	0.28	0.21	0.18	0.14	0.11
NRR, k=5	0.52	0.42	0.33	0.27	0.21	0.17	0.14	0.12
NRR, k=6	0.54	0.41	0.33	0.26	0.22	0.17	0.14	0.13

As you can see from Table 2 with the censoring proportion growth the power of NRR test decreases. The NRR test for considered composite hypotheses is the most powerful in case of censoring proportion lower than 60%. For heavily censored data the NRR test have lower power than other considered tests. The values of NRR test power for different numbers of grouping intervals are very close to each other.

Conclusions

The NRR chi-square test has a number of advantages comparing with other nonparametric goodness-of-fit tests. At first, the limiting statistic distribution does not depend on the distribution of censoring times. However, simulations have shown that for small sample sizes the distributions of NRR test statistic depend on the proportion of censoring as well as the distribution of censoring times in variational series. Secondly, the power of the NRR chi-square test in the case of considered pair of competing hypotheses is higher than the power of modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for censored samples.

Acknowledgements

These researches have been supported by the Federal targeted program of the Ministry of Education and Science of the Russian Federation "Scientific and scientific-educational personnel of innovative Russia".

References

[1] Akritas M.G. (1988). Pearson-Type Goodness-of-Fit Tests: The Univariate Case. *Journal of the American Statistical Association*. Vol. **83**, No. **401**, pp. 222-230.

[2] Bagdonavicus V., Kruopis J., Nikulin M. (2010). *Nonparametric Tests for Censored Data*. John Wiley and Sons, New Jersey.

[3] Habib M.G., Thomas D.R. (1986). Chi-Square Goodness-of Fit Tests for Randomly Censored Data. *Annals of Statistics*. Vol. **14**, No. **2**, pp. 759-765.

[4] Hjort N.L. (1992). On Inference in Parametric Survival Data. *International Statistical Review*. Vol. **60**, No. **3**, pp. 355-387.

[5] Hollander M., Peña E.A. (1992). A Chi-Squared Goodness-of-Fit Test for Randomly Censored Data. *Journal of the American Statistical Association*. Vol. **87**, No. **418**, pp. 458-463.

[6] Kim J.H. (1993). Chi-Square Goodness-of-Fit Tests for Randomly Censored Data. *The Annals of Statistics*. Vol. **21**, No. **3**, pp. 1621-1639.

[7] Klein J.P., Moeschberger M.L. (2003). *Survival analysis techniques for censored and truncated data*. Springer-Verlag, New York.

[8] Lemeshko B.Yu., Chimitova E.V., Pleshkova T.A. (2010). Testing simple and composite goodness-of-fit hypotheses by censored samples. *Nauchniy vestnik NGTU*. No. **4(41)**, pp. 13-28.

Part II
Statistical Methods in Reliability and Survival analysis

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

2.1 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.

Let's 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.

Let's 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 information 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.

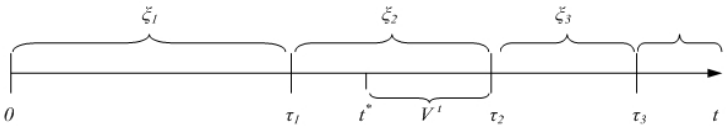


Figure 1: The strategy of the system with failure indication and a small recovery time

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.

It's 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

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

Let's 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.

2.2 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 $r_i = l_{i+1}$.

Let's consider that n is the total number of failures, m is 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

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

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 let's estimate the failure flow 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).

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

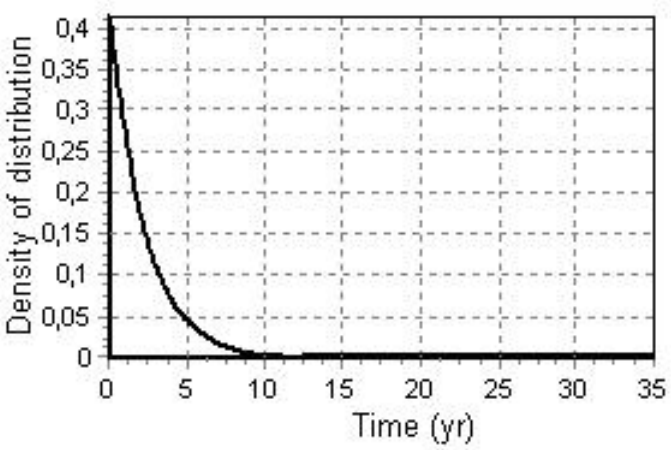


Figure 2: The kernel estimation of failure time density

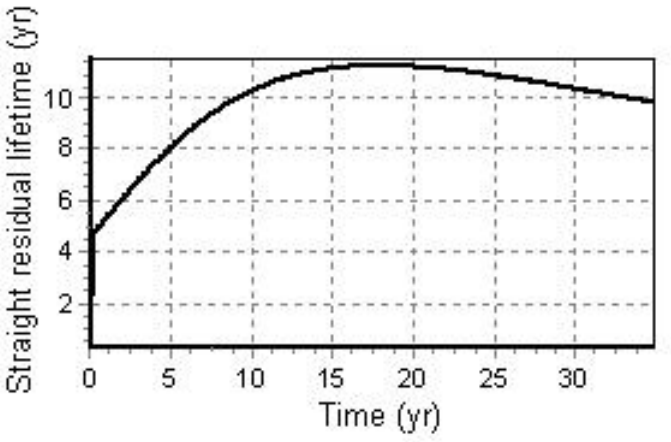


Figure 3: The mathematical expectation of average straight residual time

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. Let's 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.

Let's 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, let's 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 is the amount of bootstrap repetitions; $d(MV_i^t \leq MV_{high}^t)$ is 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.

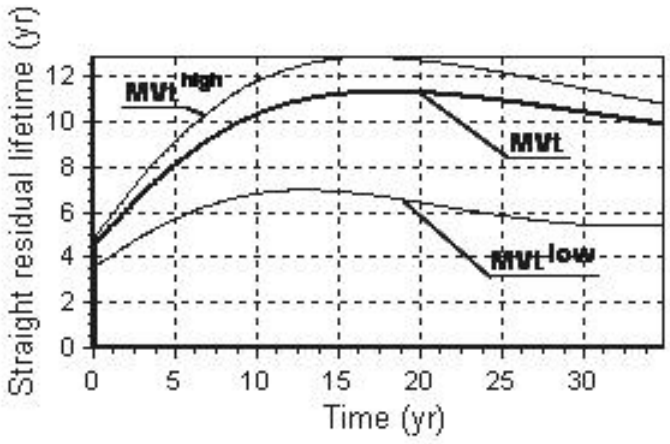


Figure 4: The construction of confidence intervals for estimating of average straight residual lifetime

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.

Statistical Analysis of Mortality-Comorbidity Links

Varvara V. Tsurko, Anatoly I. Michalski
Russian Academy of Sciences, Institute of Control Sciences
Moscow, Russian Federation
e-mail: v.tsurko@gmail.com, ipuran@yandex.ru

Abstract

In this paper we investigate dependences between associated diseases that a person has at the end of his live and the cause of death. We analyze public data about cause-specific mortality in conjunction with the problem of average risk estimation on empirical data. The use of the theory of Vapnik-Chervonenkis provides informative results about differences between distributions of associated diseases in group of people who died of cancer and group of people who died of another disease. This difference uncovers a relationship between some groups of associated diseases and risk of death of cancer.

Keywords: cancer mortality, distributions discrepancy, selection of associated diseases, the Vapnik-Chervonenkis dimension.

Introduction

Investigation of relationships between health and mortality is relevant because of longevity increase, which is observed in developed countries starting from the second part of the XX century. In order to release economic and social pressure due to ageing of population it is important to have the solution to following problems: obtaining reliable estimates for expected age structure of the population, gaining knowledge of factors responsible for "healthy aging" and understanding the impact of different diseases in cause-specific mortality. The last problem is known as mortality-comorbidity problem. This problem is especially important for old age groups in which the mortality is at the high level and several chronic diseases are presented.

Nowadays there is a great volume of statistical data for mortality and morbidity of the aged people. These data allow us to investigate the factors responsible for maintaining health in aging population, to evaluate the influence of heredity, environment and lifestyle. Recent publications explain observed increase of human life expectancy by the reduction of mortality at middle age [2, 3, 4]. There is a hypothesis that people who are down in health have a high margin of "active longevity", because enduring high morbidity risk in young and middle ages gives an advantage in survival in old age. Effective adaptation of people with chronic diseases may serve as a biological basis for this phenomenon. If this hypothesis is correct we need to focus preventive measures in young and middle ages to ensure the "healthy aging". Relationship between the cause specific mortality and chronic diseases can be an indirect confirmation of the relationship between increased morbidity and reduced mortality. In the research the Multiple Cause-of-Death Public-Use Data for 2007 by the National Center for Health Statistics USA [5] are investigated. Distribution of associated diseases presented by the ICD10 codes among people who died of cancer

(C00-C97) is compared with the same distribution among people who died of another disease. In order to select more "important" diseases associated with cancer mortality we solve a problem of contrasting the distributions. By the problem of contrasting of the distributions we mean the selection of associated diseases for which one observes the most distinguishable distributions of these diseases among people died of cancer against people died of other diseases.

We used symmetrized Kullback-Leibler divergence as a difference measure between the two distributions. For a set of associated diseases the symmetrized Kullback-Leibler divergence was estimated from the data as a half sum of mixed entropies corrected by a penalty term. This term takes into account both the amount of empirical data and the number of considered associated diseases. In this report the technique based on the Vapnik-Chervonenkis dimension was used for construction of such penalty term.

The results show that in a group of women at the age of 45-75 years, which died of cancer, not more than 9 of 126 classes of associated diseases influence on cancer mortality. Considering partition into groups of diseases we conclude that the class of chronic pulmonary and respiratory diseases is one of the most important comorbidity class of cancer and non cancer mortality differentiation. This research validates the hypothesis that presence of asthma could decrease development and mortality of cancer [3, 6].

1 Definitions

In this section we introduce some definitions and notation that will be used throughout the paper.

We consider the problem of estimation a distance between two distributions $p_1(x)$ and $p_2(x)$ on empirical data, where $p_1(x)$ is a distribution of associated diseases among people who died of cancer (let's name this group as a cancer group), $p_2(x)$ – a distribution of associated diseases among people who died of another disease (non cancer group). Associated diseases are grouped into classes according to their ICD10 classification, x is a group of associated diseases.

We used symmetrized Kullback-Leibler divergence as a distance between empirical estimates $\widehat{p}_1(x)$ and $\widehat{p}_2(x)$ of the two distributions: $D = -\frac{1}{2}\left(\sum\limits_x p_2(x)\ln\frac{\widehat{p}_1(x)}{p_2(x)} + \sum\limits_x p_1(x)\ln\frac{\widehat{p}_2(x)}{p_1(x)}\right)$

For each block of associated diseases values n_i and m_i are calculated, where n_i is a number of people of cancer group which had a disease from the ith block, m_i – a number of people of non cancer group which had a disease from the ith block. Cancer and non cancer groups have histograms of associated diseases: $g_1 = (n_{(1)}, n_{(2)}, \ldots, n_{(k)})$ and $g_2 = (m_{(1)}, m_{(2)}, \ldots, m_{(k)})$, where k – a number of blocks and blocks are sorted in descending order of the absolute difference between the values $n_i/\sum\limits_{i=1}^{k} n_i$ and $m_i/\sum\limits_{i=1}^{k} m_i$.

To find a set of associated diseases which are the most important for difference between cancer and non cancer death we consider different sets of blocks of associated diseases. Let α be a variable which labels what set of blocks we use now, Σ is a set of all possible sets α. In an experimental part we create follow sequence of variables α: $\alpha_{(1)}$ – the first block of associated disease, $\alpha_{(2)}$ – the first and the second blocks,..., $\alpha_{(k)}$ – all blocks, where the order of blocks

the same as in histograms above. Distributions of associated diseases now dependend of the variable α, and the symmetrized Kullback-Leibler divergence takes form

$$D(\alpha) = -\frac{1}{2}\left(\sum_x p_2(x,\alpha)\ln\frac{\widehat{p}_1(x,\alpha)}{p_2(x,\alpha)} + \sum_x p_1(x,\alpha)\ln\frac{\widehat{p}_2(x,\alpha)}{p_1(x,\alpha)}\right)$$

In the rest of the article we consider a functional of average risk as a criterion characterizing criterion of the distance $D(\alpha)$:

$$M(\alpha) = -\frac{1}{2}\left(\sum_x p_2(x,\alpha)\ln\widehat{p}_1(x,\alpha) + \sum_x p_1(x,\alpha)\ln\widehat{p}_2(x,\alpha)\right) \tag{1}$$

The distributions $p_1(x,\alpha)$ and $p_2(x,\alpha)$ are unknown and they are approximated by frequencies. We can use a trivial approximation by frequencies $\nu_1(x,\alpha)$ and $\nu_2(x,\alpha)$ which are equal to a portion of people who had an associated disease x and died of cancer or of another disease respectively. If x is an ith block of associated disease, α consists of k blocks, then frequencies are defined as:

$$\nu_1(x,\alpha) = \frac{n_i}{\sum\limits_{i=1}^{k} n_i}, \nu_2(x,\alpha) = \frac{m_i}{\sum\limits_{i=1}^{k} m_i},$$

To avoid zero value under logarithm in (1) we use empirical estimates $\widehat{p}_1(x)$ and $\widehat{p}_2(x)$ of distributions $p_1(x)$ and $p_2(x)$ in form:

$$\widehat{p}_1(x,\alpha) = \frac{n_i+1}{\sum\limits_{i=1}^{k} n_i + k}, \widehat{p}_2(x,\alpha) = \frac{m_i+1}{\sum\limits_{i=1}^{k} m_i + k} \tag{2}$$

These expressions are Bayes estimates of probabilities if a priori distribution of probabilities on the k-fold simplex given by $\Delta^k = \{p_1,\ldots,p_k\colon \sum\limits_{i=1}^{k} p_i = 1, p_i \geq 0, i = 1,\ldots,k\}$ is uniform.

By substitution of $\nu_1(x,\alpha)$ and $\nu_2(x,\alpha)$ instead of $p_1(x,\alpha)$ and $p_2(x,\alpha)$ in (1) we obtain so called empirical risk

$$\begin{aligned} M_e(\alpha) &= -\frac{1}{2}\left(\sum_x \nu_2(x,\alpha)\ln\widehat{p}_1(x,\alpha) + \sum_x \nu_1(x,\alpha)\ln\widehat{p}_2(x,\alpha)\right) = \\ &= -\frac{1}{2}\left(\frac{1}{\sum\limits_{j=1}^{k} m_j}\sum_{i=1}^{k} m_i \ln\frac{n_i+1}{\sum\limits_{j=1}^{k} n_j + k} + \frac{1}{\sum\limits_{j=1}^{k} n_j}\sum_{i=1}^{k} n_i \ln\frac{m_i+1}{\sum\limits_{j=1}^{k} m_j + k}\right) \end{aligned} \tag{3}$$

The deviation between the average risk and the empirical risk can be estimated in form of an inequality

$$M(\alpha) > M_e(\alpha) - d(\alpha,\eta),$$

which is valid with probability η.

By maximizing on α the right part of the inequality we determine the set of classes of associated diseases for which distribution of the associated diseases in cancer group maximal differs of the distribution of the associated diseases in non cancer group. The form of the penalty term $d(\alpha, \eta)$ and some empirical results are discussed in the rest of the article.

2 Vapnik-Chervonenkis evaluation

In this section we consider functionals of the average and the empirical risks and discuss an applicability of Vapnik-Chervonenkis evaluation as a bound of the difference between these risks.

We consider the functional of the average risk $M(\alpha)$ in the form (1) and the functional of the empirical risk $M_e(\alpha)$ in the form (3).

Let $x^\alpha_{1i}, i = 1, \ldots, L^\alpha_1$ denote a block of associated diseases which ith person form the cancer group had, where L^α_1 is a number of people who belonged to the cancer group and had an associated disease from a set α. In the same way, let $x^\alpha_{2i}, i = 1, \ldots, L^\alpha_2$ denote a block of associated diseases which ith person form the non cancer group had. Then we can obtain the following expression for the empirical risk (3):

$$M_e(\alpha) = -\frac{1}{2}\left(\frac{1}{L^\alpha_2}\sum_{i=1}^{L^\alpha_2}\ln \widehat{p}_1(x^\alpha_{2i}, \alpha) + \frac{1}{L^\alpha_1}\sum_{i=1}^{L^\alpha_1}\ln \widehat{p}_2(x^\alpha_{1i}, \alpha)\right) \tag{4}$$

We want to use the Vapnik-Chervonenkis result from [1] about the uniform convergence of means to expectations in a class of bounded functions. The result is as follows: assume $F(x, \alpha)$ – a measurable function for all $\alpha \in \Sigma$ with respect $P(x)$ in probability space X, $M(\alpha)$ is an expectation of this function for all α:

$$M(\alpha) = EF(x, \alpha) = \int F(x, \alpha)dP(x)$$

Then assume an independent sample with the distribution P(x): $X^l = x_1, \ldots, x_l$. For all α we calculate an empirical mean for $F(x, \alpha)$ for X^l

$$M_e(\alpha) = \frac{1}{l}\sum_{i=1}^{l} F(x_i, \alpha)$$

If the function $F(x, \alpha)$ is bounded: $0 \le F(x, \alpha) \le a$ then the inequality holds:

$$P\left\{\sup_{\alpha\in\Sigma}|M(\alpha) - M_e(\alpha)| > a\varepsilon\right\} \le 6M^S(2l)\exp\left[-\frac{1}{4}\varepsilon^2(l-1)\right], \tag{5}$$

where $M^S(2l)$ is a growth function of a system of events $A = \{x : F(x, \alpha) \ge c, c > 0, \alpha \in \Sigma\}$. The proof of (5) and a definition of the function $M^S(2l)$ are given in [1]. From (5) it follows that

with probability not less than $1-\eta$ for any $\alpha \in \Sigma$ it holds

$$|M(\alpha) - M_e(\alpha)| \leq 2a\sqrt{\frac{\ln 6M^s(2l) - \ln \eta}{l-1}} \tag{6}$$

To use this inequality let's prove that the function $F(x,\alpha) = -\ln \widehat{p}_i(x,\alpha)$ is bounded in our task. Hence, we should prove two inequalities:

$$a_1 \leq \ln \widehat{p}_i(x,\alpha) \leq a_2, i = 1,2$$

Indeed, we don't consider associated diseases which neither people from the cancer group nor form the non cancer groups has. Then we have $\widehat{p}_i(x,\alpha) > c > 0$ and $-\ln\left(\frac{\widehat{p}_i(x,\alpha)}{ce}\right) < 1$. By definition (2): $\widehat{p}_i(x,\alpha) < 1$ and therefore $-\ln\left(\frac{\widehat{p}_i(x,\alpha)}{ce}\right) > \ln(ce)$.

So we've proved an equality:

$$\ln(ce) < -\ln\left(\frac{\widehat{p}_i(x,\alpha)}{ce}\right) < 1$$

Functions $\ln\left(\frac{\widehat{p}_i(x,\alpha)}{ce}\right), i = 1,2$ are bounded and it is clear that:

$$-\frac{1}{2}\left(\frac{1}{L_2^\alpha}\sum_{i=1}^{L_2^\alpha}\ln\left(\frac{\widehat{p}_1(x_{2i}^\alpha,\alpha)}{ce}\right) + \frac{1}{L_1^\alpha}\sum_{i=1}^{L_1^\alpha}\ln\left(\frac{\widehat{p}_2(x_{1i}^\alpha,\alpha)}{ce}\right)\right) =$$

$$= \ln(ce) - \frac{1}{2}\left(\frac{1}{L_2^\alpha}\sum_{i=1}^{L_2^\alpha}\ln \widehat{p}_1(x_{2i}^\alpha,\alpha) + \frac{1}{L_1^\alpha}\sum_{i=1}^{L_1^\alpha}\ln \widehat{p}_2(x_{1i}^\alpha,\alpha)\right),$$

and we can use the inequality riskseval for the empirical risk estimation. With the estimation $M^S(2l) \leq (2(L_1^\alpha + L_2^\alpha))^k$ this leads to the inequality which holds with probability not less than $1-\eta$ for all sets of associated diseases composed not more than k classes

$$M(\alpha) > M_e(\alpha) - 2\sqrt{\frac{2^{k-1}\left(\ln\frac{2(L_1^\alpha+L_2^\alpha)}{2^{k-1}} + 1\right) - \ln\frac{\eta}{5}}{(L_1^\alpha + L_2^\alpha) - 1}} \tag{7}$$

2.1 Experimental results

In this section we present the analysis of the data about human comorbidity and mortality. We are interested in differences between two groups of people: people who died of cancer and people who died of another disease. Usually a person in addition to underlying disease (the cause of death) has a list of associated diseases. Therefore there are certain distributions of associated diseases in these two groups of people.

For the analysis the Multiply Cause-of-Death Public-Use data for 2007 are used. These data contain the information about people who died in 2007 year. About each person we have: age, date of death, a disease which was a cause of death, a list of associated diseases. We made our differences analysis on the data of women morbidity.

We consider the age range in which the cancer mortality is the most common. Figure 1 shows a percentage of deaths of cancer depending on age. The horizontal axis corresponds to an age of death; the vertical axis corresponds to a percentage of cancer deaths among all women death at the certain age. Figure 1 shows that the majority of cancer deaths (more than 30%) are in the age interval between 45 and 75 ages. In the rest of the article we consider the group of women who died at the age interval 45-75 years. At the first step of our analysis we consider 24

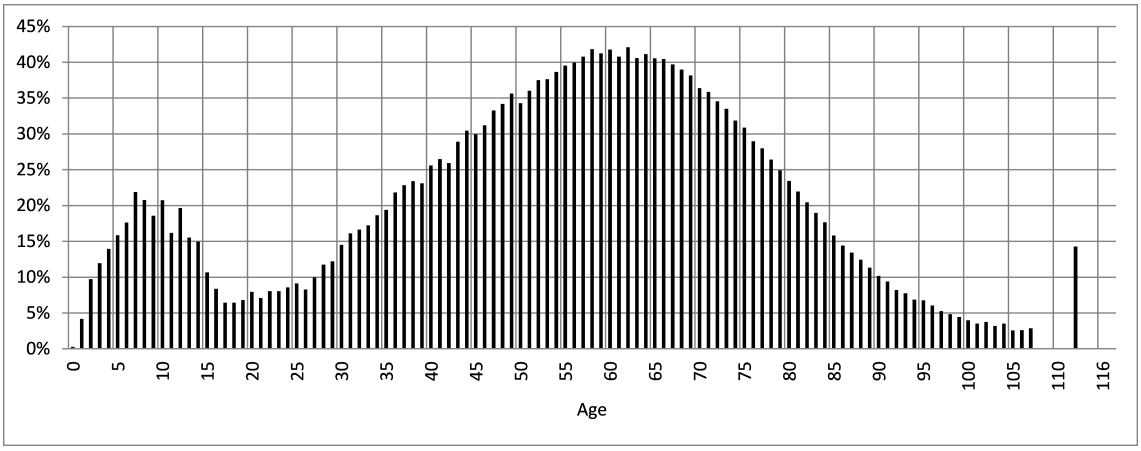


Figure 1: Histogram of proportion of women died of cancer

classes of associated diseases; these classes correspond to the first letter in the ICD10 code.

A number of women from the cancer or non cancer group who had the associated disease from fixed class is calculated (according to our definitions n_i and m_i respectively). Classes of associated diseases are sorted in descending order of the absolute difference between values $n_i/\sum_{i=1}^{k} n_i$ and $m_i/\sum_{i=1}^{k} m_i$.

To evaluate the empirical and average risks for the experimental data and to find a set of associated diseases which are the most important for difference between cancer and non cancer death we consider different sets α of classes of associated diseases. We create the following sequence of sets α: $\alpha_{(1)} = \{\mathrm{F}\}$, $\alpha_{(2)} = \{\mathrm{F}, \mathrm{T}\}$, $\alpha_{(k)}$ — all considered classes, where the order of classes is the same as defined above.

Using the mortality data we calculate values of the empirical risk functional of the form (3) for each set α. According to inequality (7) we evaluate the lower bound of the average risk. In figure 2 the empirical risk $M_e(\alpha)$ and the lower bound of the average risk $M(\alpha)$ are plotted. The

lower bound of the average risk reaches its maximum on the set $\alpha = \{\mathrm{F}, \mathrm{T}, \mathrm{J}, \mathrm{I}, \mathrm{E}, \mathrm{D}, \mathrm{N}, \mathrm{S}, \mathrm{R}, \mathrm{G}, \mathrm{K}\}$ and determines the set of classes of associated diseases for which the distribution of associated diseases in the cancer group maximally reliably differs from the distribution of associated diseases in the non cancer group.

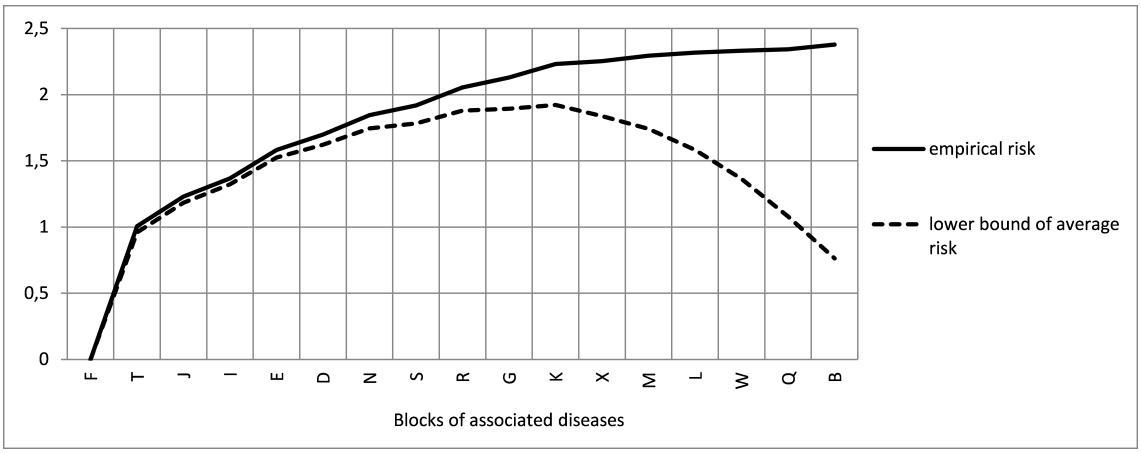


Figure 2: Empirical and average risks for classes of associated diseases

The classes of associated diseases combined by the first letter of the ICD10 code are too large and heterogeneous. At the second step of our analysis we consider more detailed composition of diseases classes: now we use blocks of diseases defined by letter and the two digits of the ICD10 code. These blocks are defined in standard classification.

For these blocks of associated diseases we perform the same comparison of the empirical and average risks. From figure 3 one can conclude that the cancer group maximally differs from the non cancer group on nine blocks of associated diseases. We emphasize such diseases as: Ischemic heart diseases (I20-I25), Hypertensive diseases (I10-I15), Other diseases of the respiratory system (J95-J99), Chronic lower respiratory diseases (J40-J47), Circulatory and respiratory systems (R00-R09), Renal failure (N17-N19), Diabetes mellitus (E10-E14), Obesity and other hyperalimentation (E65-E68), Influenza and Pneumonia (J09-J18). Some of these diseases may play protective role against cancer death, some can be artifacts. A part of the found relationships between cancer and associated diseases are well-known, some of these relationships are being discussed in professional area [6].

2.2 Conclusion

This paper is devoted to the problem of investigation of links between risk of cancer death and associated morbidity. It is mathematically formalized as the problem of contrasting the distributions of associated diseases among people died of cancer and among people died of another disease. To solve this problem we evaluate the average risk on the empirical data using the Vapnik-

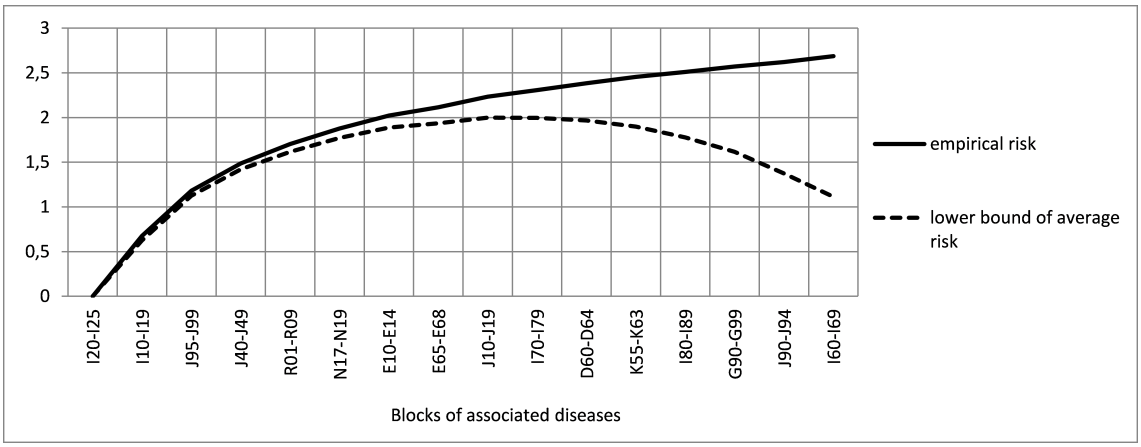


Figure 3: Empirical and average risks for blocks of associated diseases

Chervonenkis inequalities. We perform two partitions of groups of diseases and obtain the better, more interpretable results on the second partition (letter and two digits of the ICD10 code). It turns out that nine blocks of associated diseases have reliably different distributions in the cancer and non cancer groups. This allows us to discuss the role of some chronic diseases and conditions in cancer mortality. The aim of the future investigations is consideration of the more "tiny" classes of associated diseases that reduce cancer mortality. For such classes one should use more precise estimation for the average risk than estimation based on the Vapnik-Chervonenkis approach.

References

[1] Vapnik V. (1998). *Statitical Learning Theory*. Wiley Interscience.

[2] Blagosklonny M.V. (2010). Why human lifespan is rapidly increasing: solving "longevity riddle" with "revealed-slow-aging" hypothesis. *Aging*. Vol. **4**, pp. 177-182.

[3] Michalski A.I., Ukraintseva S.V.,Arbeev K.G., Yashin A.I. (2005). Investigation of old age mortality structure in the presence of comorbidity. *In European Conference On Chronic Disease Prevention. Helsinki, Finland*, p. 60.

[4] Yashin A.I. et al. (2001). THave the oldest old adults ever been frail in the past, A hypothesis that explains modern trends in survival. *Gerontol. Biol. Sci.*. Vol. **56**, pp. 432-442.

[5] Mortality Data, Multiple Cause-of-Death Public-Use Data Files

[6] Elmer W., Fisherman M.D. (1960). Does the allergic diathesis influence malignancy, *Journal of Allergy*. Vol. **31**, pp. 74-78.

Testing Goodness-of-Fit with Parametric AFT-model

Ekaterina V. Chimitova, Natalia S. Galanova
Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: natalia.galanova@gmail.com

Abstract

This paper is devoted to the problems of goodness-of-fit testing with parametric AFT-model. Modified nonparametric goodness-of-fit tests such as Kolmogorov test, Cramer-von Mises-Smirnov test and Anderson-Darling test by samples of residuals are investigated. The problem of baseline distribution selecting is considered.

Keywords: AFT-model, censored data, samples of residuals, Kolmogorov test, Creamer-von Mises-Smirnov test, Anderson-Darling test.

Introduction

There are many problems of longevity and aging data in different areas such as medicine, survival analysis, reliability studies, econometrics, etc. This is so-called time-to-event data. In medicine this event may be the time of death, time of changes in some bio-chemical indices or time of remission after some treatment. In engineering this event may be time of failure for some interesting device or technical system.

Let the nonnegative random variable ξ denote the time-to-event or failure time of an individual. The probability of an item surviving up to time t is given by the survival function.

$$S(t) = Pr\{\xi > t\} = 1 - F(t), \tag{1}$$

where $F(t)$ is cumulative distribution function of random variable ξ.

In survival analysis an individual's survival depends on some characteristics or conditions of the experiment. Usually these characteristics are coded as the so-called covariates, which could be time-dependent.

It is often necessary to obtain reliability results from experiment more quickly then it is possible with data obtained under normal conditions. In these cases experimenter may use Accelerated Failure Time Models. In AFT-models time-to-event data are obtained under some accelerated stress conditions, which shorten the life of test items. For example, covariate x is accelerated with respect to a covariate z, if:

$$S_x(t) \leq S_z(t). \tag{2}$$

The aim of such testing is to estimate survival function of an individual under the normal conditions basing on data obtained in Accelerated Life Testing.

1 AFT-model

Consider two plans of experiment [1]:

1. Experimenter divide individuals into k groups and they are tested under accelerated constant over time stresses x. Therefore n_i items are tested under x_i stress condition.

2. Individuals are tested under step stress condition:

$$x(t) = \begin{cases} x_0 & t_0 < t \text{ and } t \leq t_1 \\ x_1 & t_1 < t \text{ and } t \leq t_2 \\ ... & \\ x_{k-1} & t_{k-1} < t \text{ and } t \leq t_k \end{cases} \tag{3}$$

In addition plan of experiment may be the combination of these two plans.

Under the AFT-model survival function $S_x(t)$ is determined by baseline survival function S_0 and positive function $r(x, \beta)$:

$$S_x(t) = S_0 \left(\int_0^t \frac{ds}{r(x(s), \beta)} \right) \tag{4}$$

Stress function is usually parameterized in one of following ways:

1. Log-linear model: $r(x, \beta) = e^{\beta_0 + \beta_1 x}$;
2. Power rule model: $r(x, \beta) = e^{\beta_0 + \beta_1 ln x}$;
3. Arrhenius model: $r(x, \beta) = e^{\beta_0 + \frac{\beta_1}{x}}$;
4. Model for vector stress: $r(x, \beta) = e^{\beta_0 + \beta_1 x_1 + ... + \beta_m x_m}$.

For parametric AFT-models it is supposed that baseline survival function $S_0(t)$ belongs to some parametric family of distributions. For example: exponential model, Weibull model, Gamma model, power generalized Weibull model, inverse Gaussian model and so forth.

In survival analysis and reliability studies, time-to-event data are usually right censored. That means a time-to-event T is observed only if $T \leq T_C$, where T_C is a censoring time. There are various types of right censoring schemes [3]:

1. Type I censoring: all items are tested until a pre-specified censoring time T_C;

2. Type II censoring: only k first failure times are observed, and for remained subjects censoring time is $T_C = T_{(k)}$, where $T_{(k)}$ is failure time of k-th item;

3. Type III censoring (random censoring): the failure times $T_1, ..., T_n$ and the censoring times $C_1, ..., C_n$ are independent positive random variables.

Let denote T_i and C_i as the failure and censoring times of the i-th item respectively. Set

$$X_i = min(T_i, C_i) \tag{5}$$

Usually right censored data are presented as:

$$(X_1, \delta_1), ..., (X_n, \delta_n) \tag{6}$$

where $\delta_i = 1_{\{T_i \leq C_i\}}$ is an indicator of the event.

Estimates of AFT-model parameters can be found with maximum likelihood method, where likelihood function is:

$$L\left(T_n\right) = \prod_{i=1}^{n} f^{\delta_i}(X_i) S^{1-\delta_i}(X_i) \tag{7}$$

If the stresses are constant over time, then survival function of item which failed under x_i stress is:

$$S_{x_i}(t) = S_0\left(\frac{t}{r(x_i, \beta)}\right) \tag{8}$$

But if we have time-dependent stress (3) then survival function of item which failed under x_i stress is:

$$S_{x_i}(t) = S_0\left(\frac{t - t_i}{r(x_i, \beta)} + \sum_{j=0}^{i} \frac{t_j - t_{j-1}}{r(x_{j-1}, \beta)}\right) \tag{9}$$

2 Testing goodness-of-fit with AFT-model

It is often difficult to choose the distribution law for baseline survival function S_0 because usually there is no prior information about lifetime distribution. After estimation of model parameters one should test goodness-of-fit of obtained model to the sample of observations. So, testing goodness-of-fit is an essential part of statistical analysis. One approach to testing goodness-of-fit with parametric AFT-model is based on residuals which in case of fixed covariates can be calculated as following:

$$z_i = \frac{X_i}{r(x_i, \beta)} \tag{10}$$

But if we have time-depended covariates (3) then residuals can be calculated as follows:

$$z_i = \frac{X_i - t_i}{r(x_i, \beta)} + \sum_{j=0}^{i} \frac{t_j - t_{j-1}}{r(x_{j-1}, \beta)} \tag{11}$$

If the model (4) is appropriate the sample of residuals $Z_n = \{z_i\}_{i=1,n}$ belongs to the distribution F_0, which is standardized by the scale parameter (scale parameter is equal to 1). The hypothesis about goodness-of-fit of the sample of residuals to F_0 can be tested with the classical goodness-of-fit tests: the Kolmogorov test, the Cramer-von Mises-Smirnov test and the Anderson-Darling test. But the main problem with analysis of residual samples is that that data in reliability studies is often censored and the classical statistic inference cannot be applied.

So the hypothesis about goodness-of-fit of the sample of residuals to the baseline distribution F_0 can be tested with modified goodness-of-fit tests. These modifications are based on using the Kaplan-Meier estimator instead of empirical distribution function [5].

The Kolmogorov test statistic is given with the statement

$$D_n = \sup_{t<\infty} |\hat{F}_n(t) - F_0\left(t, \hat{\theta}\right)| \tag{12}$$

the Cramer-von Mises-Smirnov test statistic

$$W_n^2 = n \int_{-\infty}^{\infty} \left(\hat{F}_n(t) - F_0(t, \hat{\theta})\right)^2 dF_0(t, \hat{\theta}). \tag{13}$$

and the Anderson Darling test statistic

$$A_n^2 = n \int_{-\infty}^{\infty} \left(\hat{F}_n(t) - F_0(t, \hat{\theta})\right)^2 \frac{dF(t, \hat{\theta})}{F_0(t, \hat{\theta})\left(1 - F_0(t, \hat{\theta})\right)}. \tag{14}$$

where $\hat{F}_n(t)$ is the Kaplan-Meier estimator [4] by censored sample.

It should be noted that we have a composite hypothesis, for which test statistic distributions $G(S|H_0)$ are affected by a number of factors: the form of assuming lifetime distribution $F_0(t)$, the type and the number of estimated parameters, the method of parameter estimation and other factors.

In [6] the approximations of statistic distribution models and the tables of percentage points were obtained for testing composite hypotheses by the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests using the maximum likelihood estimates of unknown parameters. These approximations were obtained for complete data without covariates.

It has been shown that for testing goodness-of-fit with parametric AFT-model by complete data one can use approximations of obtained in [6]. In case of censored data statistic distributions are affected by the censoring degree. Figure 1 shows the dependence of statistic distributions from censoring degree:

So for censored data approximate p-values in testing goodness-of-fit can be obtained by simulation. It is quite possible if we have type I or type II censored data, but in case of random censoring process which often occurs in survival analysis there is a problem of ambiguity in simulating censored observations because the distribution of censoring times is unknown.

3 Motorette Failure Time Data

Let us consider the Motorette Failure Time Data [2-3].Failure times and the plan of the experiment are given in Table 1.

The parameterization of the stress function was chosen as: $r(z[x], \beta) = e^{\beta_0 + \beta_1 z[x]}$, where $z[x] = \frac{1000}{273.2+x}$.

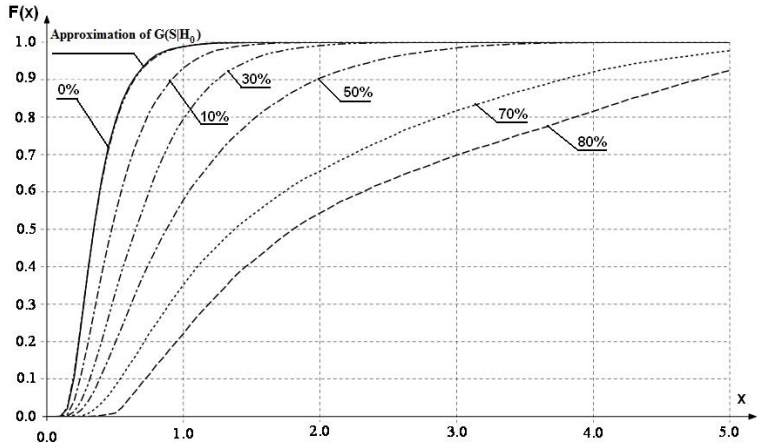


Figure 1: The Anderson-Darling test statistic distributions for censored data

Table 1: Failure Times and Plan of the Experiment

Temperature C°	Failure times
150	Censored: 10 at 8064*
170	1764, 2772, 3444, 3542, 3780, 4860, 5196 Censored: 3 at 5448*
190	408, 408, 1344, 1344, 1440 Censored: 5 at 1680*
220	408, 408, 504, 504, 504 Censored: 5 at 528*

In Table 2 MLE of the model parameters, log-likelihood function and 95% confidence intervals for different parametric models are given.

Table 2: Failure Times and Plan of the Experiment

F_o	MLE	Confidence interval		lnL
Exp	$\beta_0 = -15.28$	-24.05	-6.97	-155.36
	$\beta_1 = 10.84$	7.08	14.92	
$Weibull$	$\beta_0 = -12.96$	-15.73	-9.12	
	$\beta_1 = 9.54$	7.82	10.84	-146.28
	$\nu = 3.08$	2.15	4.90	
$Gamma$	$\beta_0 = -14.56$	-18.05	-10.05	
	$\beta_1 = 9.58$	7.64	11.18	-147.37
	$\nu = 4.46$	2.56	9.82	
PGW	$\beta_0 = -10.53$	-13.40	-7.05	
	$\beta_1 = 9.67$	7.95	10.88	-145.87
	$\nu_0 = 2.63$	1.78	4.17	
	$\nu_1 = 0.001$	0.0002	0.0096	
$LogNorm$	$\beta_0 = -13.27$	-16.99	-8.67	
	$\beta_1 = 9.65$	7.59	11.39	-148.57
	$\nu = 0.59$	0.37	0.83	

From Table 2 one cal see that by values lnL the Weibull AFT-model and the Power Generalized AFT-model fit the data better than other considered models. Figure 2 shows the survival functions for these two models under the "normal" stress conditions - temperature 130°:

The hypothesis about goodness-of fit with these parametric AFT-models was tested by modified nonparametric tests by samples of residuals. Note that in this study data are type I censored (censored degree is about 57%). So statistic distributions $G(S|H_0)$ were obtained by means of computer simulation. Table 3 presents the results of goodness-of-fit testing by modified nonparametric tests by samples of residuals.

As one can see from Table 3 both of considered models fit the data, but obtained p-values for Weibull AFT-model are bigger than for PGW AFT-model. So we can make a conclusion that the Weibull AFT-model fits the Motorette Failure Time Data better than other considered models.

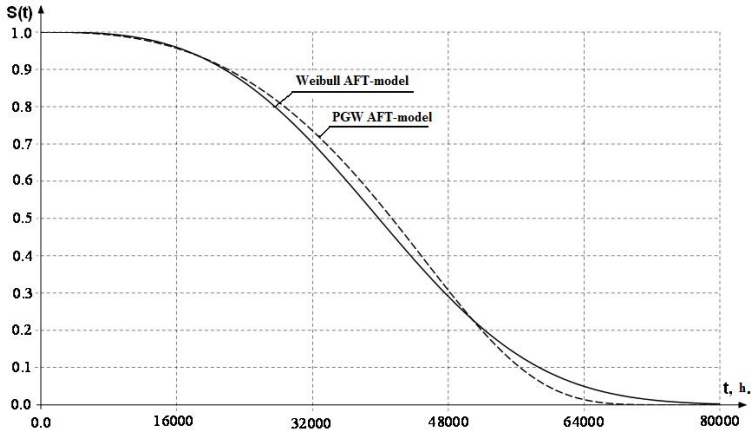


Figure 2: Weibull AFT-model and PGW-AFT-model survival functions

Table 3: Results of goodness-of-fit testing

	Kolmogorov test		Cramer-von Mises-Smirnov test		Anderson-Darling test	
	S_n	p	S_n	p	S_n	p
Weibull	1.64	0.48	0.36	0.48	1.93	0.46
PGW	1.63	0.42	0.35	0.41	1.90	0.40

Conclusions

We have briefly discussed the problem of the goodness-of-fit testing with parametric AFT-model. By means of computer simulation and developed software system we have investigated omnibus statistics distributions for testing goodness-of-fit with AFT-model basing on residuals.

It has been shown that in case of complete samples it is possible to use the approximations of statistic distributions given in [6] or to obtain p-values by simulation for data without covariates. In case of random censored samples which often occur in survival analysis there is a problem of ambiguity in simulating censored observations because the distribution of censoring times is unknown. But if we have type I or type II censored data, we don't have any problems with modeling censored samples and p-values can be obtained by simulation.

In the paper we considered the example of Motorette Failure Time Data [2-3] which include 17 complete lifetimes and 23 type I censored observations. Various parameterizations of the baseline survival function for the AFT-model have been compared for these data. Goodness-of-fit testing has been carried out by modified nonparametric tests with statistic distributions $G(S|H_0)$, obtained by simulation.

References

[1] Bagdonavicius V., Nikulin M. (2002). *Accelerated Life Models. Modeling and Statistical Analysis*. Chapman&Hall/CRC, London.

[2] Crawford D.E. (1970). Analysis of Incomplete Life Test Data on Motorettes. *Insulation/Circuits*. Vol. **16**, pp. 43-48.

[3] Kalbfleisch J.D., Prentice R.L. (2002). *The Statistical Analysis of Failure Time Data* 2nd ed. Wiley-Interscience, New Jersey.

[4] Kaplan E. Meier P. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*. Vol. **53** 282., pp. 457-481.

[5] Koziol J.A. (1980). Goodness-of-Fit Tests for Randomly Censored Data. *Biometrika*. Vol. **67** No. 3., pp. 693-696.

[6] Lemeshko B.Yu., Lemeshko S.B. (2009). Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part 1. *The Annals of Statistics*. Vol. **52**(6), pp. 555-565.

Inference for a Simple Step-Stress Model with Progressive Type II Censoring and An Extension of the Exponential Distribution

Firoozeh Haghighi, Mazaher Sohrabi Noor
Department of Mathematics, Statistics and Computer Science, University of Tehran
Tehran, Iran
e-mail: haghighi@khayam.ut.ac.ir

Abstract

In this work we consider a simple step-stress model under progressive Type-II censoring based on an extension of the exponential distribution, which provides a more flexible model than the exponential model. This new generalization of the exponential distribution has been recently introduced by Nadarajah and Haghighi (2010), and can be used for modeling lifetime data. For this simple step-stress model the maximum likelihood estimates of its parameters as well as the corresponding observed fisher information matrix are derived. A method for simulating data from an extension of the exponential distribution in the presence of progressive Type-II censoring is proposed. Using this method we conducted a simulation study for estimating the parameters of simple step-stress model and then provided asymptotic and bootstrap confidence intervals for the parameters.

Keywords: An extension of the exponential distribution, Bootstrap, Coverage probabilities, Cumulative exposure model, Fisher information matrix, Step-stress test, Type-II censoring.

Introduction

Today's many products are designed to operate without failure for years or more. Thus, few units will fail in a test of particular length at normal use conditions. A simple way to accelerate the time to failure of such products is to run the product at a higher normal stress. Generally, information from tests at high levels of stresses is extrapolated, through a reasonable statistical model, to obtain estimates lifetime at normal use conditions. Accelerated life testings are used widely in industries, particulary to obtain timely information on the reliability of products. The stress loading in an accelerated life testing can be applied various ways. A common stress loading is step stress loading wherein a unit is first subjected to a specified constant stress x_1 for a specified length of time τ_1. If it does not fail, it is subjected to a higher stress level x_2 for a specified time τ_2 and so on. The stress on a unit is thus increased step by step until it fails. If there is a single change of stress, the accelerated life test is called a simple step-stress test. For considering the cumulative effect of exposure at sequential stress levels, the cumulative exposure models are used. In this paper, we consider the problem of estimation in step-stress accelerated life tests under an extension of the exponential model and progressive Type-II censoring.

Recently a generalization of the exponential distribution has been presented by Nadarajah and Haghighi (2010), so-called an extension of the exponential distribution. The generalization always has a decreasing probability density function and yet allows for increasing, decreasing and constant hazard rates. It was shown that this new family can be used as an alternative for Weibull, gamma and exponentiated exponential distributions. An extension of the exponential distribution has a closed form expression for survival function as follows:

$$S(t) = \exp\{1 - (1 + \frac{t}{\lambda})^{\alpha}\}, \tag{1}$$

for $\alpha > 0$, $\lambda > 0$ and $t > 0$. The corresponding cumulative distribution function (cdf), probability density function (pdf) and quantile function are

$$F(t) = 1 - \exp\{1 - (1 + \frac{t}{\lambda})^{\alpha}\}, \tag{2}$$

$$f(t) = \frac{\alpha}{\lambda}(1 + \frac{t}{\lambda})^{\alpha-1} \exp\{1 - (1 + \frac{t}{\lambda})^{\alpha}\}, \tag{3}$$

and

$$Q(p) = \lambda\{(1 - \log(1-p))^{\frac{1}{\alpha}} - 1\}, \quad 0 < p < 1. \tag{4}$$

For $\alpha = 1$, the family reduces to the exponential distribution. The new family is a particular case of the power generalized Weibull family introduced by Bagdonaviçius and Nikulin (2002). They used this family as the baseline of the AFT model for FTR data. It is interesting to mention that the new distribution can be interpreted as a truncated Weibull distribution, supposing $Z = Y - 1/\lambda$, the distribution is the same as that of Z truncated at zero, when Y is a Weibull random variable. We refer the readers to Nadarajah and Haghighi (2010) for details about the properties of this family.

1 The model and assumptions

Some basic assumptions are considered as followings:

1) Two stress levels x_1 and x_2 $(x_1 < x_2)$ are used.
2) For any level of stress, the lifetime distribution of the test unit follows pdf given by (3).
3) The scale parameter of cdf $F(t)$ is a log-linear function of stress, i.e.,

$$log\lambda(x_i) = \beta_0 + \beta_1 x_i, \quad i = 0, 1, 2, \tag{5}$$

where β_0, and $\beta_1 (< 0)$ are unknown parameters depending on the nature of the product, and the method of test and x_0 is use-stress.

4) The constant α doesn't depend of the stress level.

The test is conducted as follows. An ensemble of n units is initially placed on stress x_1. At the time of the first failure, R_1 units are randomly removed from the remaining $n-1$ surviving units. At the second failure R_2 units from the remaining $n-2-R_1$ units are randomly removed and so on. The experiment continues until a pre-specified time τ. Then the stress is changed to x_2. Under stress x_2 the experiment continues until mth failure is observed and at which time all remaining $R_m = n-m-R_1-R_2-...-R_{m-1}$ units are removed. The R_i are fixed prior to the study. Let n_1 denote the random number of failures that occur before τ and $n_2 = m-n_1$ the number of failures that occur after τ. The observed data are then the ordered failure times $t_{1:m:n} < t_{2:m:n} < ... < t_{m:m:n}$. An extension of the exponential step-stress model is formulated as follows.

$$F(t) = \begin{cases} F_1(t), & 0 \le t < \tau, \\ F_2(t-\tau+\tau'), & \tau \le t < \infty. \end{cases}$$

where $F_i(t) = 1-\exp\{1-(1+\frac{t}{\lambda_i})^{\alpha}\}$, and $\tau' = (\frac{\lambda_2}{\lambda_1})\tau$ is the solution of $F_1(\tau) = F_2(\tau')$. Thus the pdf of the lifetime of the test unit is

$$f(t) = \begin{cases} f_1(t), & 0 \le t < \tau, \\ f_2(t-\tau+(\frac{\lambda_2}{\lambda_1})\tau), & \tau \le t \le \eta. \end{cases}$$

where, for $i = 1, 2$,

$$f_i(t) = \frac{\alpha}{\lambda_i}(1+\frac{t}{\lambda_i})^{\alpha-1}\exp\{1-(1+\frac{t}{\lambda_i})^{\alpha}\}.$$

The likelihood of the observed failure times given by

$$L(\beta_0,\beta_1) = C\left\{\prod_{i=1}^{n_1} f_1(t_{i:m:n})[1-F_1(t_{i:m:n})]^{R_i}\right\}\left\{\prod_{i=n_1+1}^{m} f_2(t_{i:m:n})[1-F_2(t_{i:m:n})]^{R_i}\right\},$$

It is observed that the estimation procedure, through the likelihood equations, does not result in closed-form for the maximum likelihood estimators of the parameters β_0 and β_1. Therefore, we have to solving the likelihood equations by numerical methods. The observed information matrix is derived by computing the negative of the second partial and mixed partial derivatives of log $L(\beta_0,\beta_1)$ with respect to β_0 and β_1. It is not possible to derive the exact confidence intervals for the parameters. Hence, we will consider the asymptotic and bootstrap confidence intervals for the model parameters. We also provide different bootstrap confidence intervals including bootstrap-t, percentile and adjusted percentile (BCa) CIs.

We proposed following theorem for simulating progressive Type-II data from an extension of the exponential distribution.

1.1 Theorem

Let $Z_1, Z_2, ..., Z_m$ be random variables from standard exponential distribution. Let

$$t_1 = b\left[\left(1 + \frac{Z_1}{n}\right)^{\frac{1}{\alpha}} - 1\right]$$

$$t_2 = b\left[\left(1 + \frac{Z_1}{n} + \frac{Z_2}{n-1-R_1}\right)^{\frac{1}{\alpha}} - 1\right]$$

$$t_3 = b\left[\left(1 + \frac{Z_1}{n} + \frac{Z_2}{n-1-R_1} + \frac{Z_3}{n-2-R_1-R_2}\right)^{\frac{1}{\alpha}} - 1\right]$$

$$.$$

$$.$$

$$t_m = b\left[\left(1 + \frac{Z_1}{n} + \frac{Z_2}{n-1-R_1} + ... + \frac{Z_m}{n-m+1-\sum_1^{m-1} R_i}\right)^{\frac{1}{\alpha}} - 1\right]$$

Then $t_1, t_2, ..., t_m$ is the progressive Type-II censored sample from $F(t_i) = 1 - \exp\{1 - (1 + \frac{t_i}{b})^{\alpha}\}$.

2 Simulation

A Mote Carlo simulation study is carried to evaluate the bias, mean square error and asymptotic confidence interval for the parameters. We also discussed the construction of confidence intervals for the parameters based on the different bootstrap methods and compared them in the terms of the coverage probabilities through a simulation study. Finally, a numerical example is presented to illustrate all the methods discussed here.

References

[1] V. Bagdonaviçius and M. Nikulin, *Accelerated Life Model: Modeling and Statistical Analysis*. Florida: Chapman & Hall/CRC Press, 2002.

[2] D. S. Bai, M. S. Kim and S. H. Lee, "Optimum simple step-stress accelerated life tests with censoring". *IEEE Trans. Reliability,*, vol. 38, no. 5, pp 528-532, 1989.

[3] N. Balakrishnan and Q. Xie, "Exact inference for a simple step-stress model with type-II hybrid censored data from the exponential distribution," *J. Statist. Plann. Infer*, vol. 137, pp 2543-2563, 2007a.

[4] N. Balakrishnan and Q. Xie, "Exact inference for a simple step-stress model with type-I hybrid censored data from the exponential distribution," *J. Statist. Plann. Infer*, vol. 137, pp 3268-3290, 2007b.

[5] N. Balakrishnan and D. Han, "Exact inference for a simple step-stress model with competing risks for failure from exponential distribution under type-II censoring," *J. Statist. Plann. Infer*, vol. 138, pp 4172-4186, 2008.

[6] N. Balakrishnan, L. Zhang and Q. Xie, "Inferene for a simple step-stress model with type-I censoring and lognormally distributed lifetimes," *Communications in statistics-Theory and methods*, vol. 38, pp 1690-1709, 2009.

[7] B. Efron and R. Tibshirani, *An introduction to the bootstrap* . New york: Chapman & Hall, 1993.

[8] E. Gouno, A. Sen and N. Balakrishnan, "Optimal step-stress test under progressive type-I censoring," *IEEE Trans. Reliability*, vol. 53, no. 3, pp 388-393, 2004.

[9] M. Kateri and N. Balakrishnan, "Inference for a simple step-stress model with type-II censoring, and Weibull distributed lifetimes," *IEEE Trans. Reliability*, vol. 57, no. 4, pp 616-626, 2008.

[10] W. Q. Meeker and L. A. Escobar, *Statistical methods for Reliability Data*. New York: Wiley, 1998.

[11] S. Nadarajah and F. Haghighi, "An extension of the exponntial distibution," *Statistics*, DOI:10.1080/02331881003678678 , 2010.

[12] W. B. Nelson, "Accelerated life testing - step-stress models and data analyses," *IEEE Trans. Reliability*, vol. R-29, no. 2, pp 103-108, 1980.

[13] W. B. Nelson, *Accelerated Testing Statistical Models, Test plans, and data Analysis*. New York: John Wiley & Sons, 1990.

[14] N. M. Sedyakin, "On one physical principal in reliability theory (in Russian)," *Techn. Cybernet*, vol. 3, pp 80-87, 1966.

[15] C. Xiong, "Inferences on a simple step-stress model with type-II censored exponential data," *IEEE Trans. Reliability*, vol. 47, no. 2, pp 142-146, 1998.

Comparing Prediction Performances via IDI. Application to French Alzheimer data

C. Huber-Carol, S. Toaff-Gross
University Paris Descartes, Paris, France
Bernard M. Baruch College of the City University, New York, USA
e-mail: catherine.huber@parisdescartes.fr, Shulamith.Gross@baruch.cuny.edu

1 Motivation

It may happen that among the factors that have an impact on the occurrence of a certain event we want to predict, some of them are difficult to obtain. It can be due to their high cost or else to the time spent to get them. In that case, and if the purpose is purely predictive, and not at all explanatory, it may happen that dropping such factors has a very low cost in terms of predictive ability of the model. The aim of this paper is to derive the asymptotic properties of an estimator of an index of predictive ability, the IDI, when both the full model and the reduced model are estimated on the same data set, together with their IDI. Having thus a confidence interval for their comparative predictive ability, we have elements allowing us to conclude whether we can drop or not certain pertinent factors.

2 Framework

Let $X = (Y, \mathbf{Z})$ be a random variable such that Y is binary with values in $\{0, 1\}$, $P(Y = 1|\mathbf{Z} = \mathbf{z}) = \mathbf{p}(\mathbf{z})$, and $\mathbf{Z}$ is a k-dimensional real variable, with distribution $Q(\mathbf{z})$ with density $q(\mathbf{z})$ with respect to some measure μ. We have a data set $\mathbf{X} = (X_1, \cdots, X_n)$ consisting in n i.i.d. observations of X, and two models for predicting Y on the basis of $\mathbf{Z}$ are to be compared:

$$\begin{array}{lrcl} \text{Model 1} & P(Y = 1|\mathbf{Z} = \mathbf{z}) & = & p_1(\mathbf{z}) \\ \text{Model 2} & P(Y = 1|\mathbf{Z} = \mathbf{z}) & = & p_2(\mathbf{z}) \end{array}$$

while the unknown true distribution of X is given by

$$\begin{array}{lcl} P(Y = 1|\mathbf{Z} = \mathbf{z}) & = & p(\mathbf{z}) \\ dQ(\mathbf{z}) & = & q(\mathbf{z})\mathbf{d}\mu(\mathbf{z}) \end{array}$$

This setting originates from the following special problem in epidemiology:
Y_i is the indicator of the occurrence of a certain disease for subject i. Occurrence of this event is to be predicted to happen within a fixed period of time, the prediction being based on the value $\mathbf{z_i}$ of $\mathbf{Z}$ observed on subject i. $\mathbf{Z}$ is a k-dimensional covariate, p_1 and p_2 are logistic models, denoted g_1 and g_2 in the sequel. While g_1 is including all k components of $\mathbf{Z}$, g_2 is obtained by throwing away Z_k which is a genetic feature. We consider the case when a test of fit of the full

model g_1 shows that Z_k is a pertinent covariate, that is its coefficient is significantly different from 0. It may happen that, in spite of the fact that g_1 is a better model than g_2, the improvement in prediction is not significant, due to the fact that the coefficients of the remaining covariates are modified so as to fit better the data at the cost of giving misleading false effects for the remaining covariates. The reason for avoiding the last covariate, even though it is a pertinent one, may be, as is the case for a genetic feature, the fact that it is not available for all the subjects that could be involved in the study, or else it would be too expensive or too long to obtain its values in view of the magnitude of the small benefit it would provide.
We stress that, while $q(z)$ is the true distribution of $\mathbf{Z}$, we do not assume that the full model g_1 is the true model for the data.
The aim of this work is to derive the asymptotic properties of the IDI in order to obtain confidence intervals for IDI (Integrated Discrimination Improvement) and other related measures of comparison of prediction performances. Several indexes for comparing prediction ability of two models can be found in Pencina et al.[1]. Properties of M-esdtimators as can be found in C. Huber[2] are used for derivng the asymptotic properties of IDI.

3 Application to French Alzheimer data

n = 4486 patients aged $\geq$ 65, included in a cohort between September 1999 and November 2000, are followed during several years. Covariates such as sex, age at inclusion, sociological, psychological and biological factors as well as three genetic factors are considered that could influence the occurrence of Alzheimer dsease. Among the 4486 patients, 162 became Alzheimer within 4 years. Only one of the three genetic factors is shown to be pertinent and is included in the best fitting logistic model for predicting this occurrence. Nevertheless, the IDI between the two models with and without the genetic factor is not significant. This allows us to think that the search for this costly factor could be avoided without loosing much as long as prediction only is concerned, and not structural explanation.

References

[1] Pencina M.J., D'Agostino R.B. Sr, D'Agostino R.B. Jr & Vasan R.S. (2007). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medecine*

[2] Huber C. (1985). Thorie de la Robustesse. *Lecture notes in Mathematics*, pp. 1-128, Springer Verlag.

Part III
Testing Statistical Hypotheses

Modified Chi-Squared Goodness-of-Fit Test for Birnbaum-Saunders Distribution

Mikhail Nikulin[1], N. Balakrishnan[2], Ramzan Tahir[1], Noureddine Saaidia[1]
[1]*Institut de Mathmatique de Bordeaux, University Bordeaux 1, France*
[2]*Department of Mathematics and Statistics, McMaster University, Hamilton, Ont., Canada*

Abstract

Two-parameter Birnbaum-Saunders distribution is widely used in industry for reliability. In this paper we give a modified chi-squared goodness-of-fit test for Birnbaum-Saunders distribution when the data are right censored. Random grouping intervals of the data function are used.

Introduction

Birnbaum-Saunders distribution is an important distribution in shape-scale family used for failure time data especially when the failures are due to crack. This distribution was proposed by Birnbaum and Saunders (1969a, b) with two parameters, named as shape and scale parameters. The PDF and CDF of this distribution is unimodal and is very popular in modeling fatigue failures in industry, as an alternative to other unimodal distributions such as the lognormal and inverse Gaussian.

The cumulative distribution function of two-parameter Birnbaum-Saunders distribution is

$$F(t;\alpha,\beta)=\Phi\left[\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right],\quad 0<t<\infty,\quad \alpha,\beta>0, \tag{1}$$

where α is the shape parameter, β is the scale parameter and $\Phi(x)$ is the standard normal distribution function. The probability density function can be written as

$$f(t;\alpha,\beta)=\frac{1}{2\sqrt{2\pi}\ \alpha\beta}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right\}\exp\left[-\frac{1}{2\alpha^2}\left(\frac{t}{\beta}+\frac{\beta}{t}-2\right)\right],$$
$$0<t<\infty,\quad \alpha,\beta>0.$$

We can write the survival function and the hazard function by using the following relationships;

$$S(t;\alpha,\beta)=1-F(t;\alpha,\beta),\qquad \lambda(t;\alpha,\beta)=\frac{f(t;\alpha,\beta)}{S(t;\alpha,\beta)}.$$

Two-parameter Birnbaum-Saunders distribution has been studied at large scale (see e.g. Desmond (1985,1986), Balakrishnan et al. (2007,2009), Johnson et al. (1995), Kundu et al. (2008), Lemonte et al. (2007), Volodin et al. (2000)). Some authors used this distribution for censored data (see

e.g. Lawless, (1982), Leiva et al. (2007), Zhihui et al. (2006))but there is not much work done on goodness-of-fit test for Birnbaum-Saunders distribution when the data are censored.

Here we give a modified chi-square goodness-of-fit test for this model when the data are right censored. We use the approach of Nikulin-Rao-Robson (NRR) (Bagdonavicius, et al.(2010) and Bagdonavicius, Nikulin (2011)). We give the explicit form of all the elements of the quadratic form of NRR statistic. We apply this test for the head and neck cancer data (censored) and the data analyzed by Kundu et al. (2008) (non-censored). In both case maximum likelihood estimators are used. This paper was initiated by the paper of (Kundu, D., Kannan, N. and Balakrishnan, N. (2008))

We illustrate the test with two real data examples, of which one is censored data while the second is non-censored data. The graphical comparison of the non parametric Kaplan-Meier estimation and the maximum likelihood estimation is given for each data.

1 Right censored data, composite hypothesis and ML function

We observe the right censored sample

$$(X_1, \delta_1), \ldots, (X_n, \delta_n), \tag{2}$$

where

$$X_i = T_i \wedge C_i, \quad \delta_i = \mathbf{1}_{\{T_i \le C_i\}}.$$

where $T_1, \cdots, T_n$ are the failure times which are absolutely continuous i.i.d. random variables and C_i are censoring times which are independent. The probability density function of the failure time T_1 belongs to a parametric family $\{f(\cdot, \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbf{R}^m\}$. It means that we have the parametric composite hypothesis as

$$H_0 : F(x) \in \mathcal{F}_0 = \{F(x, \boldsymbol{\theta}), \quad x \in R^1, \quad \boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T \in \Theta \subset R^m\} \subset \mathcal{F}$$

which means that the failure times T follow follow the distribution with cdf F of the parametric class $\mathcal{F}_0$ and $\boldsymbol{\theta}$ is an unknown m-dimentional parameter and F_0 is a differentiable completely specified cdf.

Let us consider the distribution of the random vector (X_i, δ_i) in the case of random censoring with absolutely continuous censoring times C_i. So we can write the loglikelihood function for non-informative censored data as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \ln \lambda(X_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} \ln S(X_i, \boldsymbol{\theta}) \{ \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}\}. \tag{3}$$

Here we have $\{\boldsymbol{\theta} = (\alpha, \beta)^t \in \boldsymbol{\Theta}\}$, as in this paper we consider that the failure time belongs to the Birnbaum-Saunders distribution. So we write the log likelihood function for this distribution as

$$\begin{aligned}\ell \quad &= \sum_{i=1}^{n} \delta_i \left[\ln\left(\tfrac{1}{2\sqrt{2\pi}}\right) - \ln\alpha - \ln\beta + \ln\left\{\left(\tfrac{\beta}{t}\right)^{\frac{1}{2}} + \left(\tfrac{\beta}{t}\right)^{\frac{3}{2}}\right\} - \tfrac{1}{2\alpha^2}\left(\tfrac{t}{\beta} + \tfrac{\beta}{t} - 2\right) - \right. \\ &\left. \ln\left\{1 - \Phi\left(\tfrac{1}{\alpha}\left\{\left(\tfrac{t}{\beta}\right)^{\frac{1}{2}} - \left(\tfrac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)\right\}\right] + \sum_{i=1}^{n} \ln\left[1 - \Phi\left(\tfrac{1}{\alpha}\left\{\left(\tfrac{t}{\beta}\right)^{\frac{1}{2}} - \left(\tfrac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)\right]\end{aligned}$$

If $\hat{\boldsymbol{\theta}}$ is the estimator of the parameter $\boldsymbol{\theta}$ then the ML estimator satisfies the equation:

$$\dot{\ell}(\hat{\boldsymbol{\theta}}) = \mathbf{0}; \tag{4}$$

here $\dot{\ell}$ is the *score vector*:

$$\dot{\ell}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\alpha}\ell(\boldsymbol{\theta}), \frac{\partial}{\partial\beta}\ell(\boldsymbol{\theta})\right)^T.$$

Fisher's information matrix is

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbf{E}_{\boldsymbol{\theta}}\ddot{\ell}(\boldsymbol{\theta}),$$

where

$$\ddot{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \frac{\partial^2}{\partial\boldsymbol{\theta}^2} \ln\lambda(X_i, \boldsymbol{\theta}) - \sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\theta}^2}\Lambda(X_i, \boldsymbol{\theta}).$$

The censored sample (2) may be written in the form of random processes

$$(N_1(t), Y_1(t), t \geq 0), \cdots, (N_1(t), Y_1(t), t \geq 0), \tag{5}$$

where

$$N_i(t) = \mathbf{1}_{X_i \leq t, \delta_i = 1}, \quad Y_i(t) = \mathbf{1}_{0 \leq t \leq X_i},$$

Using these processes we obtain two useful relations:

Using these processes under non-informative and random censoring the considered we can write the loglikelihood functions the form

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \int_0^\infty \{\ln\lambda(u, \boldsymbol{\theta}) dN_i(u) - Y_i(u)\lambda(u, \boldsymbol{\theta})\} du, \tag{6}$$

from where it follows that

$$\dot{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \int_0^\infty \frac{\partial}{\partial\boldsymbol{\theta}} \ln\lambda(u, \boldsymbol{\theta})\, dM_i(u, \boldsymbol{\theta}), \tag{7}$$

$$\ddot{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \int_0^{\infty} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln \lambda(u, \boldsymbol{\theta})\, dM_i(u, \boldsymbol{\theta}) - \sum_{i=1}^{n} \int_0^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta})\, (\frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta}))^T \lambda(u, \boldsymbol{\theta}) Y_i(u) du, \tag{8}$$

where

$$M_i(t, \boldsymbol{\theta}) = N_i(t) - A_i(t) = N_i(t) - \int_0^t Y_i(u) \lambda(u, \boldsymbol{\theta}) du.$$

is *martingale of the counting process* $N_i(t)$.

By tradition, accepted in survival analysis and reliability, we suppose that the processes N_i and Y_i are observed finite time $\tau > 0$. It means that at time τ observation of all objects is censored, so in the place of censoring time C_i, censoring time $C_i \wedge \tau$ are used. We denote them once more by C_i. The process $N(t)$ shows for any $t > 0$ the number of observed failures in the interval $[0, \tau]$ and the process $Y(t)$ shows the number of objects which are *at risk* (not failed, not truncated and not censored) just prior the time $t, t < \tau$.

So from these processes we can write the matrix of second derivatives in the following form

$$\ddot{\ell}(\boldsymbol{\theta}) = \int_0^{\tau} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln \lambda(u, \boldsymbol{\theta})\, dN(u) - \int_0^{\tau} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \lambda(u, \boldsymbol{\theta}) Y(u) du$$

$$= \int_0^{\infty} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln \lambda(u, \boldsymbol{\theta})\, dM(u, \boldsymbol{\theta}) - \int_0^{\tau} \frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta})\, (\frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta}))^T \lambda(u, \boldsymbol{\theta}) Y(u) du.$$

and the Fisher's information matrix is

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbf{E}_{\boldsymbol{\theta}} \ddot{\ell}(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \int_0^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta})\, (\frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta}))^T \lambda(u, \boldsymbol{\theta}) Y_i(u) du. \tag{9}$$

Consistency and asymptotic normality of the ML estimators $\hat{\boldsymbol{\theta}}$ holds under the some well known regularity conditions (see e.g. Hjort (1990), Bagdonavicius et al).

The score vector is

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} \delta_i \left[-\frac{1}{\alpha} + \frac{1}{\alpha^3} \left(\frac{t}{\beta} + \frac{\beta}{t} - 2 \right) + \frac{\Phi'_{\alpha} \left(\frac{1}{\alpha} \left\{ \left(\frac{t}{\beta} \right)^{\frac{1}{2}} - \left(\frac{\beta}{t} \right)^{\frac{1}{2}} \right\} \right)}{1 - \Phi \left(\frac{1}{\alpha} \left\{ \left(\frac{t}{\beta} \right)^{\frac{1}{2}} - \left(\frac{\beta}{t} \right)^{\frac{1}{2}} \right\} \right)} \right]$$

$$- \sum_{i=1}^{n} \left[\frac{\Phi'_{\alpha} \left(\frac{1}{\alpha} \left\{ \left(\frac{t}{\beta} \right)^{\frac{1}{2}} - \left(\frac{\beta}{t} \right)^{\frac{1}{2}} \right\} \right)}{1 - \Phi \left(\frac{1}{\alpha} \left\{ \left(\frac{t}{\beta} \right)^{\frac{1}{2}} - \left(\frac{\beta}{t} \right)^{\frac{1}{2}} \right\} \right)} \right],$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[-\frac{1}{\beta} + \frac{\frac{1}{2t}\left\{ \left(\frac{t}{\beta}\right)^{\frac{1}{2}} + 3\left(\frac{\beta}{t}\right)^{\frac{1}{2}} \right\}}{\left(\frac{\beta}{t}\right)^{\frac{1}{2}} + \left(\frac{\beta}{t}\right)^{\frac{3}{2}}} - \frac{1}{2\alpha^2}\left(-\frac{t}{\beta^2} + \frac{1}{t}\right) + \frac{\Phi'_\beta\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}{1 - \Phi\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)} \right] - \sum_{i=1}^{n} \left[\frac{\Phi'_\beta\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}{1 - \Phi\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)} \right],$$

where $\Phi'(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Now we have

$$\Phi'_\alpha\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right) =$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)\right]\left[-\frac{1}{\alpha^2}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right]$$

$$= -\frac{1}{\sqrt{2\pi}\ \alpha^2}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\} \exp\left[-\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)\right]$$

$$\Phi'_\beta\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right) =$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)\right]\left[\frac{1}{\alpha}\left\{-\frac{1}{2}\left(\frac{\beta}{t}\right)^{-\frac{3}{2}}\left(\frac{1}{t}\right) - \frac{1}{2}\left(\frac{\beta}{t}\right)^{-\frac{1}{2}}\left(\frac{1}{t}\right)\right\}\right]$$

$$= \frac{1}{\sqrt{2\pi}}\left[\frac{1}{\alpha}\left\{-\frac{1}{2t}\left(\frac{t}{\beta}\right)^{\frac{3}{2}} - \frac{1}{2t}\left(\frac{t}{\beta}\right)^{\frac{1}{2}}\right\}\right] \exp\left[-\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)\right]$$

$$= -\frac{1}{2\sqrt{2\pi}\ \alpha t}\left\{\left(\frac{t}{\beta}\right)^{\frac{3}{2}} + \left(\frac{t}{\beta}\right)^{\frac{1}{2}}\right\} \exp\left[-\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)\right]$$

2 NRR statistic

We consider the individual data so divide the interval $[0, \tau]$ into $k > s$ smaller intervals

$$I_j = (a_{j-1}, a_j], \quad a_0 = 0, \quad a_k = \tau,$$

and let denote by

$$U_j = N(a_j) - N(a_{j-1}) = \sum_{i:X_i \in I_j} \delta_i$$

the number of observed failures in the j-th interval, $j = 1, 2, ..., k$.

Considering the equality

$$\mathbf{E}N(t) = \mathbf{E}\int_0^t \lambda(u, \boldsymbol{\theta}_0)Y(u)du$$

we can "expect" to observe

$$e_j = \int_{a_{j-1}}^{a_j} \lambda(u, \hat{\boldsymbol{\theta}})Y(u)du \tag{10}$$

failures; here $\hat{\boldsymbol{\theta}}$ is the MLE of the parameter $\boldsymbol{\theta}$. To have a chi-square test we shall construct the NRR statistic (see Nikulin (1973b,c), Rao & Robson (1974)) based on the statistic

$$\boldsymbol{Z} = (Z_1, ..., Z_k)^T, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - e_j), \quad j = 1, ..., k. \tag{11}$$

Set for $l = 1, \cdots, m$ and $j, j' = 1, \cdots, k$

$$V_j = V(a_j) - V(a_{j-1}), \quad v_{jj'} = \mathbf{Cov}\,(V_j, V_{j'}),$$

$$A_j = A(a_j) - A(a_{j-1}), \quad \boldsymbol{C}_j = (C_{1j}, ..., C_{mj})^T = \boldsymbol{C}(a_j) - \boldsymbol{C}(a_{j-1}),$$

$$\boldsymbol{V} = [v_{jj'}]_{k\times k}, \quad \boldsymbol{C} = [C_{lj}]_{m\times k}.$$

and denote by $\boldsymbol{A}$ the $k \times k$ diagonal matrix with diagonal elements $A_1, ..., A_k$, where

$$A(t) = \int_0^t \lambda(u, \boldsymbol{\theta}_0)y(u)du, \qquad \boldsymbol{C}(t) = \int_0^t \frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta}_0)\lambda(u, \boldsymbol{\theta}_0)y(u)dt,$$

Under some regularity conditions we have

$$\boldsymbol{Z} \xrightarrow{d} \boldsymbol{Y} \sim N_k(\boldsymbol{0}, \boldsymbol{V}), \quad \text{as } n \to \infty$$

with

$$\boldsymbol{V} = \boldsymbol{A} - \boldsymbol{C}^T \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{C}.$$

where the matrix $\boldsymbol{i}(\boldsymbol{\theta}_0) = \lim_{n\to\infty} \boldsymbol{I}(\boldsymbol{\theta}_0)/n$ is positive definite and one can write

$$\boldsymbol{I}(\hat{\boldsymbol{\theta}})/n \xrightarrow{P} \boldsymbol{i}(\boldsymbol{\theta}),$$

where $\hat{\boldsymbol{\theta}}$ is the root of the equation (4).

The test statistic is

$$Y_n^2 = \boldsymbol{Z}^T \hat{\boldsymbol{V}}^- \boldsymbol{Z}. \tag{12}$$

where

$$\hat{\boldsymbol{V}}^- = \hat{\boldsymbol{A}}^{-1} + \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{C}}^T \hat{\boldsymbol{G}}^- \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^{-1},$$

is the general inverse of the matrix $\hat{\boldsymbol{V}}$, and $\hat{\boldsymbol{G}} = \hat{\boldsymbol{i}} - \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{C}}^T$.

The limit distribution of the statistic Y_n^2 is chi-square with $r = rank(\boldsymbol{V}^-) = Tr(\boldsymbol{V}^-\boldsymbol{V})$ degrees of freedom. If G is non-degenerate then $r = k$.

So the test statistic can be written in the following simple form

$$Y_n^2 = \sum_{j=1}^{k} \frac{(U_j - e_j)^2}{U_j} + Q, \tag{13}$$

where

$$Q = \boldsymbol{W}^T \hat{\boldsymbol{G}}^{-1} \boldsymbol{W}, \quad \boldsymbol{W} = \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^{-1}\boldsymbol{Z} = (W_1, ..., W_m)^T,$$

$$\hat{\boldsymbol{G}} = [\hat{g}_{ll'}]_{s\times s}, \quad \hat{g}_{ll'} = \hat{i}_{ll'} - \sum_{j=1}^{k} \hat{C}_{lj}\hat{C}_{l'j}\hat{A}_j^{-1}, \quad W_l = \sum_{j=1}^{k} \hat{C}_{lj}\hat{A}_j^{-1} Z_j.$$

$$\hat{\boldsymbol{i}}_{ll'} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\partial \ln \lambda(X_i;\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_l} \frac{\partial \ln \lambda(X_i;\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{l'}}, \quad \hat{\boldsymbol{C}}_{lj} = \frac{1}{n}\sum_{i:X_i\in I_j} \delta_i \frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(X_i, \hat{\boldsymbol{\theta}}),$$

$$\hat{A}_j = U_j/n, \quad U_j = \sum_{i:X_i\in I_j} \delta_i, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - e_j),$$

$i = 1, \cdots, n, \quad j = 1, \cdots, k, \quad l, l' = 1, \cdots, m.$
Denote by $\hat{g}^{ll'}$ the elements of $\hat{\boldsymbol{G}}^-$. The quadratic form Q can be written as follows

$$Q = \sum_{l=1}^{m}\sum_{l'=1}^{m} W_l g^{ll'} W_{l'}.$$

Chi-squared test for the hypothesis H_0: The hypothesis is rejected with approximate significance level α if $Y_n^2 > \chi_\alpha^2(k)$, where $\chi_\alpha^2(k)$ is the upper quantile of chi-square with k degree of freedom.

Remark. It is proved that under right censoring the limit distribution of the test statistic does not change (see Bagdonaviius, Kruopis and Nikulin (2010)).

Choice of intervals $\hat{a}_j$: Set

$$b_i = (n-i)\Lambda(X_{(i)}, \hat{\boldsymbol{\theta}}) + \sum_{l=1}^{i} \Lambda(X_{(l)}, \hat{\boldsymbol{\theta}}).$$

where $X_{(i)}$ is the ith element in the ordered statistics $(X_{(1)}, \cdots, X_{(n)})$.
If i is the smallest natural number verifying $E_j \in [b_{i-1}, b_i], j = 1, \cdots, k-1$ then

$$(n-i+1)\Lambda(a, \hat{\boldsymbol{\theta}}) + \sum_{l=1}^{i-1} \Lambda(X_{(l)}, \hat{\boldsymbol{\theta}}) = E_j$$

and

$$\hat{a}_j = \Lambda^{-1}\left([E_j - \sum_{l=1}^{i-1} \Lambda(X_{(l)}, \hat{\boldsymbol{\theta}})]/(n-i+1), \hat{\boldsymbol{\theta}}\right), \quad \hat{a}_k = \max(X_{(n)}, \tau)$$

where Λ^{-1} is the inverse of cumulative hazard function Λ. We have $0 < \hat{a}_1 < \hat{a}_2 \cdots < \hat{a}_k = \tau$. With this choice of intervals $ej = E_k/k$ for any j where $E_k = \sum_{i=1}^{n} \Lambda(X_i, \hat{\boldsymbol{\theta}})$. Usually in real application we fix k.

Remark: Bagdonavicius et al. (2010) give the explicit formula to estimate a_j for the shape-scale family of distributions in the form inverse hazard function. As there is no explicit form of the inverse hazard function of Birnbaum-Saunders distribution, so we estimate intervals by iterative method.

We can estimate the Fisher information matrix by using the equality

$$\hat{\boldsymbol{i}}_{ll'} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\partial \ln \lambda(X_i; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_l} \frac{\partial \ln \lambda(X_i; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{l'}}.$$

The elements of the Fisher information matrix are

$$\hat{\boldsymbol{i}}_{11} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left[-\frac{1}{\alpha} + \frac{1}{\alpha^3}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right) + \frac{\Phi'_\alpha\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}{1 - \Phi\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}\right]^2,$$

$$\hat{\boldsymbol{i}}_{22} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left[-\frac{1}{\beta} + \frac{\frac{1}{2t}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} + 3\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}}{\left(\frac{\beta}{t}\right)^{\frac{1}{2}} + \left(\frac{\beta}{t}\right)^{\frac{3}{2}}} - \frac{1}{2\alpha^2}\left(-\frac{t}{\beta^2} + \frac{1}{t}\right) + \frac{\Phi'_\beta\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}{1 - \Phi\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}\right]^2,$$

$$\hat{\boldsymbol{i}}_{12} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[-\frac{1}{\alpha}+\frac{1}{\alpha^3}\left(\frac{t}{\beta}+\frac{\beta}{t}-2\right)+\frac{\Phi'_\alpha\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}{1-\Phi\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}\right]\times$$
$$\left[-\frac{1}{\beta}+\frac{\frac{1}{2t}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}+3\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}}{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}}-\frac{1}{2\alpha^2}\left(-\frac{t}{\beta^2}+\frac{1}{t}\right)+\frac{\Phi'_\beta\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}{1-\Phi\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}\right].$$

And

$$\hat{\boldsymbol{C}}_{lj} = \frac{1}{n}\sum_{i:X_i\in I_j}\delta_i\frac{\partial}{\partial\boldsymbol{\theta}}\ln\lambda(X_i,\hat{\boldsymbol{\theta}})$$

where

$$\hat{\boldsymbol{C}}_{1j} = \frac{1}{n}\sum_{i:X_i\in I_j}\delta_i\left[-\frac{1}{\alpha}+\frac{1}{\alpha^3}\left(\frac{t}{\beta}+\frac{\beta}{t}-2\right)+\frac{\Phi'_\alpha\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}{1-\Phi\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}\right]$$

$$\hat{\boldsymbol{C}}_{2j} = \frac{1}{n}\sum_{i:X_i\in I_j}\delta_i\left[-\frac{1}{\beta}+\frac{\frac{1}{2t}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}+3\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}}{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}}-\frac{1}{2\alpha^2}\left(-\frac{t}{\beta^2}+\frac{1}{t}\right)+\right.$$
$$\left.\frac{\Phi'_\beta\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}{1-\Phi\left(\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\}\right)}\right]$$

3 Head and Neck Cancer Data

The survival times in days for the patients ($n = 51$) in arm A of the head and neck cancer trial are as below ($\delta = 42$): The data was first used by Effron (1988), and then Haghighi(2004).

7, 34, 42, 63, 64, 74*, 83, 84, 91, 108, 112, 129, 133, 133, 139, 140, 140, 146, 149, 154, 157, 160, 160, 165, 173, 176, 185*, 218, 225, 241, 248, 273, 277, 279*, 297, 319*, 405, 417, 420, 440, 523*, 523, 583, 594, 1101, 1116*, 1146, 1226*, 1349*, 1412*, 1417 .

* censoring

The maximum likelihood estimators of Birnbaum-Saunders distribution by taking into account

the survival times in months are; $\hat{\alpha} = 1.4390$, $\hat{\beta} = 7.6851$. We take 5 intervals i.e. k=5. Further results to calculate the Y^2 are shown below (using the intervals proposed by Bagdonavicius, Kruopis and Nikulin (2010)):

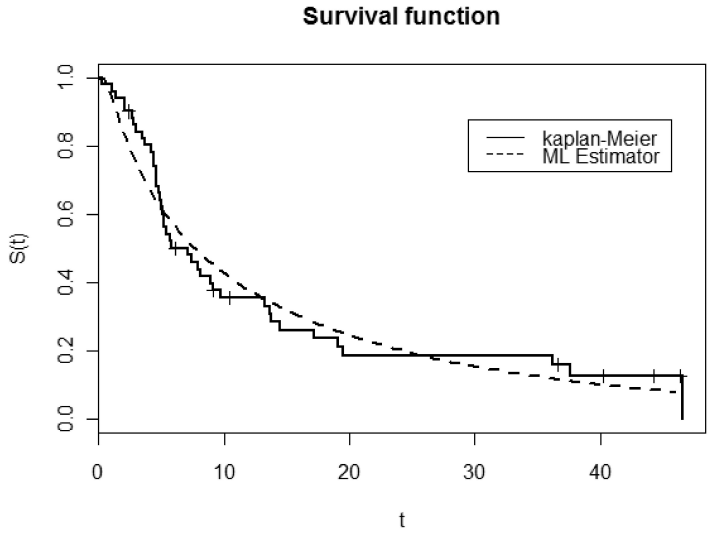


Figure 1: The empirical survival function (Kaplan-Meier) and the fitted survival functions (ML).

j	1	2	3	4	5
$\hat{a}_j$	2.0371	3.8279	7.1816	15.8970	46.5537
U_j	3	7	16	10	6
e_j	8.55876	8.55876	8.55876	8.55876	8.55876

$$\hat{G} = \begin{pmatrix} 1.062956 \cdot 10^{13} & -1.056587 \cdot 10^{12} \\ -1.056587 \cdot 10^{12} & 1.050257 \cdot 10^{11} \end{pmatrix}.$$

$$W_l = (-7398797 \quad 735422)^T$$

The value of test statistic is $Y^2 = X^2 + Q = 15.4067 + 6.7804 = 22.1871$, and the P-value is $p = P\{\chi^2_5 > 22.1871\} = 0.00048$ ($CR = 11.0705$). So from the result we can say that Birnbaum-Saunders distribution does not fits the head and neck cancer data.

Data of survival times of guinea pigs

This data by Bjerkedal (1960), was used by Gupta et al. (1997) and D. Kundu et al. (2008). The data represent the survival times of guinea pigs injected with different doses of tubercle bacilli. Here, we are primarily concerned with the animals in the same cage that were under the same

regimen. The regimen number is the common logarithm of the number of bacillary units in 0.5 ml of challenge solution; i.e., regimen 6.6 corresponds to 4.0×106 bacillary units per 0.5 ml ($\log(4.0 \times 106) = 6.6$). Corresponding to regimen 6.6, there were 72 observations listed below (no censoring):

12, 15, 22, 24, 24, 32, 32, 33, 34, 38, 38, 43, 44, 48, 52, 53, 54, 54, 55, 56,
57, 58, 58, 59, 60, 60, 60, 60, 61, 62, 63, 65, 65, 67, 68, 70, 70, 72, 73, 75,
76, 76, 81, 83, 84, 85, 87, 91, 95, 96,98, 99, 109, 110, 121, 127, 129, 131,
143,146, 146, 175, 175, 211, 233, 258, 258, 263, 297, 341, 341, 376

The maximum likelihood estimators of Birnbaum-Saunders distribution by taking into account the survival times in months are; $\hat{\alpha} = 0.76$, $\hat{\beta} = 77.5370367$. We take 8 intervals i.e. k=8. Further results to calculate the Y^2 are shown below:

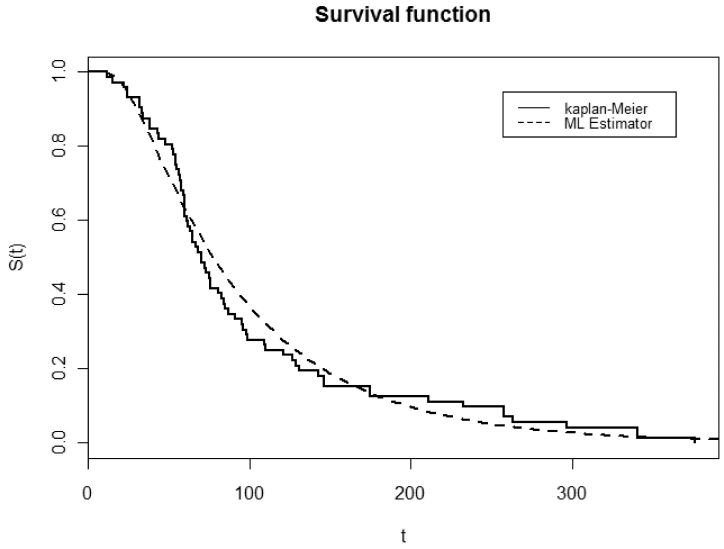


Figure 2: The empirical survival function (Kaplan-Meier) and the fitted survival functions (ML).

j	1	2	3	4	5	6	7	8
$\hat{a}_j$	33.0031	45.8625	58.4605	75.7942	101.7081	140.8752	207.7677	376
U_j	8	5	10	17	12	6	5	9
e_j	8.9549	8.9549	8.9549	8.9549	8.9549	8.9549	8.9549	8.9549

$$\hat{G} = \begin{pmatrix} 81919470537 & -548295941 \\ -548295941 & 3678042 \end{pmatrix}.$$

$$W_l = (-40424.4446 \quad 177.1083)^T$$

The value of test statistic is $Y^2 = X^2 + Q = 12.51517 + 1.08025 = 13.59542$, and the P-value is $p = P\{\chi^2_8 > 13.59542\} = 0.093$ ($CR = 15.50731$). So from the result we have no reason to reject our hypothesis of Birnbaum-Saunders distribution.

References

[1] Bagdonavicius V., Kruopis J., Nikulin M. (2010). *Nonparametric tests for Censored Data*. ISTE and J.Wiley.

[2] Bagdonavicius, V. and Nikulin, M. (2011). Chi-square goodness-of-fit test for right censored data. *Int.J.Appl.Math.Stat.*. Vol. **24**, Issue No.SI-11A, pp. 30-50.

[3] Bagdonavicius, V., Nikulin M., and Ramzan Tahir (2010). Chi-squared goodness-of-fit test for generalized Weibull distribution when data are right censored. In proceeding of, International Conference on Accelerated Life Testing, Reliability-based Analysis & Design. Clermont-Ferrand, France. No.SI-11A, pp. 30-50.

[4] Balakrishnan, N., Leiva, V., and Lopez, J. (2007). Acceptance sampling plans fromtruncated life tests based on the generalized Birnbaum-Saunders distribution. *Communications in Statistics: Simulation and Computation*. No. 36, pp. 643-656.

[5] Balakrishnan, N., Victor Leiva, Antonio Sanhueza and Filidor Vilca (2009). Estimation in the Birnbaum-Saunders distribution based on scale-mixture of normals and the EM-algorithm. *SORT*. No. **33 (2)**, pp. 171-192.

[6] Birnbaum, Z. W. and Saunders, S. C. (1969a). A new family of life distributions. *Journal of Applied Probability*. Vol. **6**, pp. 637-652.

[7] Birnbaum, Z. W. and Saunders, S. C. (1969b). Estimation for a family of life distributions with applications to fatigue. *Journal of Applied Probability*. Vol. **6**, pp. 328-347.

[8] Bjerkedal, T. (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli. *Amer. J. Hygiene*. Vol. **72**, pp. 130148.

[9] Desmond, A.F. (1985). Stochastic models of failure in random environments. *Canad. J. Statist*. Vol. **13**, pp. 171183.

[10] Desmond, A. (1986). On the Relationship Between two Fatigue-lide Models. *Transactions on Reliability*. Vol. **35**, pp. 167-169.

[11] Gupta, R.C., Kannan, N., RayChoudhuri, A. (1997). Analysis of log-normal survival data. *Math. Biosci*. Vol. **139**, pp. 103115.

[12] Haghighi, F. (2004) "Fiabilité et analyse de survie: Le modèle semi-paramétrique de dégradation. Une généralisation de la famille Weibull et ses applications." Thèse pour le Doctorat de l'Université Victor Segalen Bordeaux 2, 6 Décembre 2004, Bordeaux, France.

[13] Hjort, N.L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates. *The Annals of Statistics*. Vol. **18**, #3, pp. 1221-1258.

[14] Johnson, N.L., Kotz, S., Balakrishnan, N. (1995). *Continuous Univariate Distributions*. Vol. **2**, second ed.Wiley, NewYork.

[15] Kundu, D., Kannan, N. and Balakrishnan, N. (2008). On the hazard function of Birnbaum-Saunders distribution and associated inference. *Computational Statistics and Data Analysis*. Vol. **52**, pp. 2692-2702.

[16] Lawless, J.F. (2003). *Statistical models and methods for lifetime data*. JohnWiley & Sons, NewYork.

[17] Leiva, V., Barros, M., Paula, G. A. and Galea, M. (2007). Influence diagnostics in log-Birnbaum-Saunders regression models with censored data. *Computational Statistics and Data Analysis*. Vol. **51**, pp. 5694-5707.

[18] Lemonte, A.J., Cribari-Neto, F., Vasconcellos, K.L.P. (2007). Improved statistical inference for the two-parameter BirnbaumSaunders distribution. *Comput. Statist. Data Anal.*. Vol. **51**, pp. 46564681.

[19] Nikulin, M.S. (1973a). *Chi-square test for normality*. International Vilnius Conference on Probability Theory and Mathematical Statistics, Vol. **2**, pp. 119-122.

[20] Nikulin, M.S. (1973b). Chi-square test for continuous distribution with shift and scale parameters. *Theory of Probability and its Application*. Vol. **19**, No.3, pp. 559-568.

[21] Nikulin, M.S. (1973c). On a chi-square test for continuous distribution. *Theory of Probability and its Application*. Vol. **19**, No.3, pp. 638-639.

[22] Rao, K.C., and Robson, D.S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. *Communication in Statistics*. Vol. **3**, pp. 1139-1153.

[23] Volodin, I.N., Dzhungurova, O.A. (2000). On limit distribution emerging in the generalized Birnbaum-Saunders model. *J. Math. Sc.*. Vol. **99**, pp. 1348-1366.

[24] Zhihui Wang, A.F. Desmond, Xuewen Lu (2006). Modified censored moment estimation for the two-parameter BirnbaumSaunders distribution. *Computational Statistics & Data Analysis*. Vol. **50**, pp. 1033 1051.

On Validation of Models in Demography

Léo Gerville-Réache, Mikhail Nikulin, Ramzan Tahir
IMB, University of Bordeaux, France

Abstract

In demography, Gompertz and Makeham models have significant role in modeling and in analysis of mortality and ageing. Till the end of 20th century, researchers have been used the tables of mortalities for demographic analysis but in the end of 20th century due to the development in statistical methods of survival analysis one can treat the individuals data even with the information of censoring. Weibull model is considered the alternative for Gompertz model (Juckett and Rosenberg, (1993)). The Gompertz, Makeham, and Weibull distributions are compared with respect to the goodness-of-fit to the table of mortality and to the individuals data in presence of censoring. For data from the table of mortality, test statistic considered by Gerville-Reache and Nikulin (2000) is used. For censored individual data the test is based on the NRR-statistic where the choice of random grouping intervals is considered as given by Hjort (1990), Akritas (1988), Bagdonavicius, Kruopis and Nikulin (2010).

Keywords:Demography, Gompertz model, Makeham model, Weibull model, Composite hypothesis, ML estimators, Chi-square test, Censoring, NRR statistic.

Introduction

In demography and actuarial sciences models selection for some specific data is vital for further analysis and decision making. Testing the two-parameter Gompertz distribution (Gompertz (1825)) to model the rate of mortality has been used for a long time, where the rate of mortality increases with the age. Gompertz-Makeham (William Makeham (1860)) model with one additional parameter covers the mortality independent of age. The researchers have been used the life and mortality tables to find the force of mortality. Gerville-Reache and Nikulin (2000) gave a chi-square type goodness-of-fit test for Makeham model using the table of mortality (grouped data). In section 3 we briefly discuss their proposed statistic and also we compare Makeham model with Gompertz and Weibull models for different age groups. But now with the advanced technology and data collection techniques, one can have the individual's information (ungrouped data) also with censoring mechanism. Gompertz and Makeham models are frequently used in demography but in survival analysis many other parametric models are used (see Bagdonavicius and Nikulin (2002)).

Mostly the researcher compare Gompertz model with the Weibull model due to its flexible parameters (Gavrilov & Gavrilova (2001)). Logistic distribution can be another alternative for Gompertz (Wilson (1994)). The Gompertz function is a better choice for all causes of mortality and combined disease categories while the Weibull model has been shown a better choice

over Gompertz model for a specific cause of mortality (Juckett & Rosenberg (1993)). A little work is done for Gompertz-Makeham model in the presence of censoring (Wang et all. (1998)). Estimation of parameters and construction of test-statistic is difficult when the data are censored.

For individual data with censoring the test is based on the Nikulin-Rao-Robson statistic known as NRR statistic Y_n^2: a modified chi-squared test which is based on the differences between two estimators of the probabilities in each interval. One estimator is based on the empirical distribution function and the other is on the ML estimators of unknown parameters of the tested model from ungrouped data (See Nikulin (1973), Rao and Robson (1974), Drost (1988), LeCam et al. (1983), Van der Vaart (1998), Zhang (1999). That is we partition the total time $[0, \tau]$, where τ is the maximum time, into k intervals and we observe the number of failures U_j in each interval $I_1, \cdots, I_k$ and expected number of failures e_j by using the maximum likelihood estimators of the tested model. So the test is based on the vector $Z = (Z_1, \cdots, Z_k)^T$, where $Z_j = \frac{1}{\sqrt{n}}(U_j - e_j)$, $j = 1, \cdots, k$. Random grouping intervals are used and the end points of the interval a_j are estimated as a random data function. In literature some other modifications in the chi-square goodness-of-fit tests for censored data have been proposed (see for example Habib and Thomas (1986), Hollander and Pena (1992), Akritas (1988), Hjort (1990), Kim (1993)).

Bagdonavicius and Nikulin (2010) demonstrated the goodness-of-fit test for many parametric models but in this paper we only give for the Gompertz, Makeham and Weibull models. Some details on the Gompertz, Makeham and Weibull models are given in section 2. The NRR statistic for composite hypotheses, and application of the test for Gompertz and Weibull distributions are given in section 4, and 5 respectively. Non-parametric estimation of survival function in demography and actuaries is given in section 6.

1 Gompertz-Makeham and Weibull Models

Gompertz model of aging is widely used in demography and other scientific disciplines e.g. medical sciences, survival analysis, actuarial sciences and reliability. It was proposed by Benjamin Gompertz in 1825, now known as Gompertz law. This is the first mathematical model to explain the exponential increase in mortality rate with age (Gompertz (1825)). He explained that the law of geometric progression pervades in mortality after a certain age. Gompertz mortality rate can be presented as

$$\mu_x = \theta e^{\nu x}, \quad (\theta, \nu) > 0, \quad x > 0, \tag{1}$$

where θ is known as the baseline mortality, ν the age specific growth rate of the force of mortality. Mortality rate μ_x in demographic notation is the equivalent to the failure rate $\mu(x)$ in reliability or hazard rate in survival analysis. The Gompertz law has been the main demographic model since its discovering to fit the human mortality (see e.g. Gavrilov & Gavrilova (2001), Ricklefs & Scheuerlein (2002)).

Since Gompertz model gives the rate of mortality only related to age and does not take into account the other factors independent of age, so other authors tried to modify this model to fulfill the the requirement of real data. William Makeham (1860) modified the Gompertz model considering some other causes of death independent of age by proposing the so called ***Gompertz-Makeham*** law of mortality as

$$\mu_x = \gamma + \theta e^{\nu x}, \quad \text{where} \quad (\gamma, \theta, \nu) > 0 \quad x > 0. \tag{2}$$

Here the first term γ (Makeham parameter) is a constant and non-aging component of failure rate (e.g. accidents, independent of age) and the second term $\theta e^{\nu t}$ is the Gompertz function depending on age (aging factor).

The ***Weibull*** distribution is one of the most widely used distributions in survival analysis and reliability due to the characteristics of its shape parameter ν. The mortality rate or hazard function is

$$\mu_x = \frac{\nu}{\theta^\nu} x^{\nu-1}, \quad \text{for} \quad x \geq 0 \quad (\theta, \nu) > 0. \tag{3}$$

The hazard function of the Weibull distribution can be decreasing, constant or increasing according to the value of its shape parameter i.e. three Weibull models can make a bathtub shape, but now there are some models like generalized Weibull model which can have bathtub shape (Bagdonavicius and Nikulin (2002)). The Weibull law is more commonly applicable for technical devices while the Gompertz law is more common for biological systems (Gavrilov & Gavrilova (1991)). When the Gompertz law fails to follow some biological failure mechanism, the best alternative is Weibull law due to its basis in reliability theory. If the probability of failure at the start of the system is almost zero, the failure rate increases with the power function with age i.e. Weibull law and if the system have defects at the beginning, the failure rate increases exponentially with age i.e. Gompertz law. So to apply the Weibull law in demography, biological population should be independent of initial deaths. Logistic distribution is considered as the other alternative for Gompertz distribution (Vanfleteren et al. (1998)).

2 Test statistic for the table of mortality

Consider $x = 0$ as the origin of time for an individual of age x and T_x is a random variable for its residual life from this origin. The probability of death is

$${}_tq_x = \mathbf{P}\{0 < T_x \leq t\}, \quad t > 0, x > 0.$$

So the annual rate of mortality for the people having age x can be defined as

$$q_x = \mathbf{P}\{0 < T_x \leq 1\}, \quad x > 0.$$

A relation between the rate of mortality and the instantaneous rate of mortality μ_x is

$$q_x = 1 - \exp\left(-\int_x^{x+1} \mu_y dy\right), \quad x > 0.$$

The theoretical annual rate of mortality in the case of *Gompertz* model can be written as

$$q_x = 1 - \exp\left(-\frac{\theta}{\nu} e^{\nu x}(e^{\nu} - 1)\right), \quad \theta, \nu > 0. \tag{4}$$

In the same way we can find the theoretical annual rate of mortality for Makeham, Weibull and other parametric models.

We observe the n persons independent of mortality and we regroup them in the same age, say ω groups, where ω is the maximum age in years. The group G_x contains ℓ_x persons of age x $(x = 0, \cdots, \omega - 1)$ and q_x is the probability of death of each individual in the year. Let denote D_x the number of deaths in the group G_x.
Using the data D_x and ℓ_x from the table of mortality, we can obtain the empirical annual rate of mortality observed at age x, such that

$$Q_x = \frac{D_x}{\ell_x},$$

which follow the binomial law with parameters ℓ_x and q_x. According to the central limit theorem if $min_x(\ell_x) \to \infty$ when $n \to \infty$, then $Q = (Q_0, \cdots, Q_\omega)^t \sim^{as} N_\omega(q, P)$, where $q = (q_0, \cdots, q_{\omega-1})^t$ and P is the diagonal matrix of the elements $\frac{q_x(1-q_x)}{\ell_x}$ for $x = 0, 1, \cdots, \omega - 1$. So we can write that

$$\frac{(D_x - \ell_x q_x)^2}{\ell_x q_x (1 - q_x)} \sim^{as} \chi^2_1.$$

As it is shown in Gerville-Reache & Nikulin (2000),

$$X^2_\omega = \sum_{x=0}^{\omega-1} \frac{(D_x - \ell_x q_x)^2}{\ell_x q_x (1 - q_x)} \sim^{as} \chi^2_\omega.$$

One can use this statistic for testing simple hypotheses, as one uses the classical statistic of Pearson for testing simple hypotheses (See Greenwood and Nikulin (1996)).

Estimation of parameters in Composite hypothesis
Let consider the composite hypothesis

$$H_0 : q_x = q_x(\theta), \theta = (\theta_1, \cdots, \theta_s)^t \in \Theta \subseteq R^s, s < \omega.$$

We estimate the parameters by maximum likelihood method using the data from the table of mortality. We have the random variable D_x which follows the binomial law of parameters ℓ_x and q_x. The likelihood function is

$$L(\theta) = \prod_{x=0}^{\omega-1} \binom{D_x}{\ell_x} [q_x(\theta)]^{D_x} [1 - q_x(\theta)]^{\ell_x - D_x}.$$

We take the estimator $\hat{\theta}$ that maximizes the likelihood function, i.e.

$$\hat{\theta} = argmax L(\theta)$$

One can find the maximum likelihood estimator $\hat{\theta}$ for θ by solving the following score vector

$$\frac{\partial \ln L}{\partial \theta_i} = 0, \forall i = 1, \cdots, s.$$

Let consider the statistic

$$X^2_\omega(\hat{\theta}) = \sum_{x=0}^{\omega-1} \frac{(D_x - \ell_x q_x(\hat{\theta}))^2}{\ell_x q_x(\hat{\theta})(1 - q_x(\hat{\theta}))} \sim^{as} \chi^2_{\omega-s}.$$

Gerville-Reache & Nikulin (2000) shown that under the hypothesis H_0, $X^2_\omega(\hat{\theta})$ asymptotically follows a chi-square statistic with $\omega - s$ degrees of freedom, where s is the number of parameters to be estimated, from where it follows that we may use this statistic for testing H_0. One can see that the statistic $X^2_\omega(\hat{\theta})$ is different from the classical Pearson statistic.

Example: Analysis of data from the table of mortality (INSEE, Gironde 1990)

The data in Table 1 is from INSEE Aquitaine-France, give the number of deaths of the year 1990 in each age group of 5 year, where D_x are the numbers of deaths and ℓ_x are the numbers of habitants for each age group on 1st January 1990.

This data is used for the validity of three models i.e. *Gompertz*, *Makeham*, and *Weibull* model. The rate of mortality for these three models is adjusted with maximum likelihood estimators and then the value of chi-square is calculated. In the case of the adjustment between 5 and 84 years of age, the annual rate of mortality neither follow the Gompertz, Makeham, nor the Weibull models. But when the adjustment is between the age groups of 30 and 74 year, the the Makeham model is accepted. The Gompertz model also becomes valid along with Makeham when the model is adjusted for the age between 50 and 79 years. It means that Gompertz model is validated in the older age and it coincide the theory regarding Gompertz model. The Weibull model becomes closer but still it does not fits this data significantly. The calculated values of the test statistic with corresponding p-values are shown in table 2 and the fitted models are well presented in the following three figures.

Table 1: Table of Mortality (INSEE, Gironde 1990)

age	ℓ_x	D_x
5-9	75498	14
10-14	77284	16
15-19	90337	45
20-24	102544	91
25-29	91339	92
30-34	90769	128
35-39	93324	156
40-44	96692	226
45-49	64575	195
50-54	57974	247
55-59	61871	384
60-64	62473	622
65-69	61122	958
70-74	36425	944
75-79	37124	1341
80-84	29541	2020

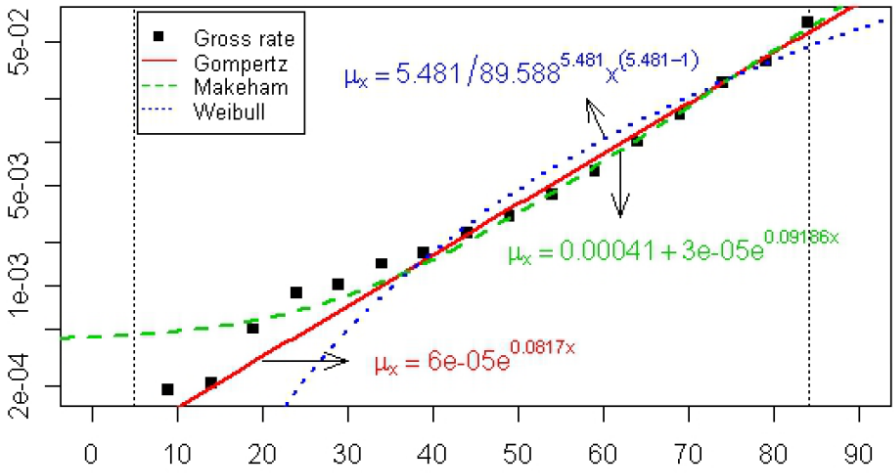


Fig1: Model fitted for age between 5 and 84 years (log scale)

3 Goodness-of-fit test for right censored data (NRR-Statistic)

Here we apply the survival analysis methods in demography where we have individual information with right censoring. For this purpose Bagdonavicius, Kruopis and Nikulin (2010) proposed a goodness of fit test based on the NRR statistic (Nikulin (1973), Rao & Robson (1974), Drost

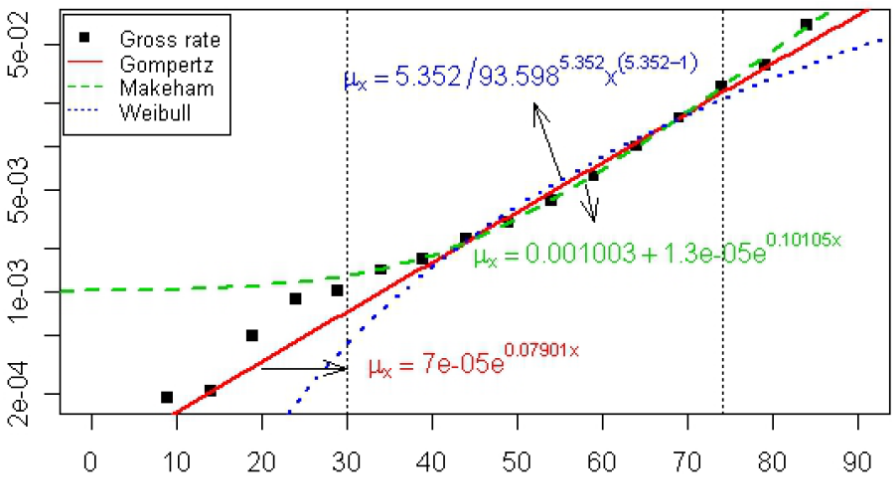


Fig2: Model fitted for age between 30 and 74 years (log scale)

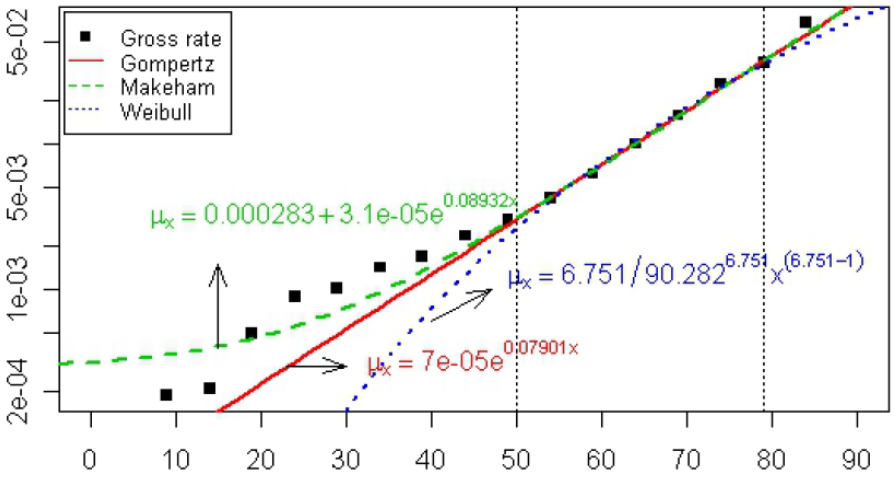


Fig3: Model fitted for age between 50 and 79 years (log scale)

Table 2: Table of Mortality (INSEE, Gironde 1990)

	Gompertz		Makeham		Weibull	
Age Groups	$X^2_\omega(\hat{\theta})$	p-value	$X^2_\omega(\hat{\theta})$	p-value	$X^2_\omega(\hat{\theta})$	p-value
5-84	214.19	≈ 0	99.98	≈ 0	2363.98	≈ 0
30-74	45.62	≈ 0	1.98	0.92	158.93	≈ 0
50-79	9.01	0.11	8.48	0.08	25.44735	0.0001

(1988), Van der Vaart (1998)). We give a chi-squared type test for testing composite parametric hypothesis when individual data are right censored.

Let us consider the composite hypothesis

$$H_0 : F(x) = F(x, \boldsymbol{\theta}), \quad x \in R^1, \quad \boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T \in \Theta \subset R^m$$

i.e. the distribution of failure times T belongs to the given parametric class. Here we consider *Gompertz*, *Makeham* and *Weibull* as parametric families.

Suppose we have right censored individual data as

$$(X_1, \delta_1), \ldots, (X_n, \delta_n), \quad X_i = T_i \wedge C_i, \quad \delta_i = \mathbf{1}_{\{T_i \le C_i\}}, \tag{5}$$

where $T_1, \cdots, T_n$ are the failure times which are absolutely continuous i.i.d. random variables and C_i are the censoring times which are independent. The probability density function of the failure times T_1 belongs to a parametric family $\{f(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbf{R}^m\}$. Denote by

$$S_t(\boldsymbol{\theta}) = \mathbf{P}_{\boldsymbol{\theta}}\{T_1 > t\}, \quad \text{and} \quad \Lambda_t(\boldsymbol{\theta}) = -lnS_t(\boldsymbol{\theta}) = \int_0^t \mu_y(\boldsymbol{\theta})dy, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

the survival function and the cumulative hazard function, respectively. With *non-informative* random censoring mechanism the loglikelihood function can be writen as,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \ln \mu_{X_i}(\boldsymbol{\theta}) + \sum_{i=1}^{n} \ln S_{X_i}(\boldsymbol{\theta}), \quad \{\boldsymbol{\theta} \in \boldsymbol{\Theta}\}. \tag{6}$$

ML estimators can be find by equating the score vector $\dot{\ell}(\hat{\boldsymbol{\theta}})$ to zero and the Fisher's information matrix is $\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbf{E}_{\boldsymbol{\theta}}\ddot{\ell}(\boldsymbol{\theta})$. Consistency and asymptotic normality of the ML estimators $\hat{\boldsymbol{\theta}}$ hold under some regularity sufficient conditions (Hjort (1990), Bagdonavicius, Kruopis and Nikulin (2010)).

To construct the test we introduce the two counting process and write the censored sample (5) as

$$(N_1(t), Y_1(t), t \ge 0), \cdots, (N_n(t), Y_n(t), t \ge 0), \tag{7}$$

where

$$N_i(t) = \mathbf{1}_{\{X_i \le t, \delta_i = 1\}}, \quad Y_i(t) = \mathbf{1}_{\{X_i \ge t\}},$$

$$N(t) = \sum_{i=1}^{n} N_i(t), \quad \text{and} \quad Y(t) = \sum_{i=1}^{n} Y_i(t).$$

Using these data one can calculate immediately the non-parametric Nelson-Aalen estimator

$$\hat{\Lambda}(t) = \int_0^t \frac{dN(u)}{Y(u)}$$

for the unknown cumulative hazard function Λ, and the famous non-parametric Kaplan-Meier estimator

$$\hat{S}(t) = \hat{S}(t-)\left(1 - \frac{\Delta N(t)}{Y(t)}\right)$$

for the unknown survival function $S(t) = 1 - F(t)$ when the censored data are ungrouped.

Suppose that the processes N_i, Y_i are observed at finite time τ. Then divide the interval $[0, \tau]$ into $k > s$ smaller intervals

$$I_j = (a_{j-1}, a_j], \quad a_0 = 0, \quad a_k = \tau.$$

Let denote the number of observed failures in the j-th interval, $j = 1, 2, ..., k$ as

$$U_j = N(a_j) - N(a_{j-1}) = \sum_{i: X_i \in I_j} \delta_i$$

Choice of random grouping intervals $\hat{a}_j$ is made to overcome the problem of very small expected number of events for some interval. This can happen in demography because the number of deaths at early age are very few. Set

$$b_i = (n-i)\Lambda_{X_{(i)}}(\hat{\boldsymbol{\theta}}) + \sum_{l=1}^{i} \Lambda_{X_{(l)}}(\hat{\boldsymbol{\theta}}),$$

where $X_{(i)}$ is the ith element in the ordered statistics $(X_{(1)}, \cdots, X_{(n)})$. If i is the smallest natural number verifying $E_j \in [b_{i-1}, b_i], j = 1, \cdots, k-1$, then

$$(n-i+1)\Lambda_a(\hat{\boldsymbol{\theta}}) + \sum_{l=1}^{i-1} \Lambda_{X_{(l)}}(\hat{\boldsymbol{\theta}}) = E_j,$$

and

$$\hat{a}_j = \Lambda^{-1}\left([E_j - \sum_{l=1}^{i-1} \Lambda_{X_{(l)}}(\hat{\boldsymbol{\theta}})]/(n-i+1), \hat{\boldsymbol{\theta}}\right), \quad \hat{a}_k = \max(X_{(n)}, \tau),$$

where Λ^{-1} is the inverse of cumulative hazard function Λ. We have $0 < \hat{a}_1 < \hat{a}_2, \cdots, \hat{a}_k = \tau$. With this choice of intervals the expected number of failures are

$$e_j = E_k/k,$$

for any j where $E_k = \sum_{i=1}^{n} \Lambda_{X_{(l)}}(\hat{\boldsymbol{\theta}})$. Usually in real application we fix k.

For testing H_0 Bagdonavicius and Nikulin (2010) considered the following statistic

$$Y_n^2 = \boldsymbol{Z}^T \hat{\boldsymbol{V}}^- \boldsymbol{Z},$$

where

$$\hat{\boldsymbol{V}}^{-} = \hat{\boldsymbol{A}}^{-1} + \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{C}}^{T}\hat{\boldsymbol{G}}^{-}\hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^{-1}$$

is a consistent estimator of a generalized inverse $\boldsymbol{V}^{-}$ of the asymptotic variance-covariance matrix $\boldsymbol{V} = \boldsymbol{V}(\theta)$ of the statistic

$$\boldsymbol{Z} = (Z_1, ..., Z_k)^T, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - e_j), \quad j = 1, ..., k, \tag{8}$$

and

$$\hat{\boldsymbol{G}} = \hat{\boldsymbol{i}} - \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{C}}^{T}.$$

So the test statistic can be written in the simple form as

$$Y_n^2 = \sum_{j=1}^{k} \frac{(U_j - e_j)^2}{U_j} + Q,$$

where

$$Q = \boldsymbol{W}^T\hat{\boldsymbol{G}}^{-1}\boldsymbol{W}, \quad \boldsymbol{W} = \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^{-1}\boldsymbol{Z} = (W_1, ..., W_m)^T,$$

$$\hat{\boldsymbol{G}} = [\hat{g}_{ll'}]_{s\times s}, \quad \hat{g}_{ll'} = \hat{i}_{ll'} - \sum_{j=1}^{k} \hat{C}_{lj}\hat{C}_{l'j}\hat{A}_j^{-1}, \quad W_l = \sum_{j=1}^{k} \hat{C}_{lj}\hat{A}_j^{-1}Z_j.$$

$$\hat{\boldsymbol{i}}_{ll'} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\partial \ln \lambda(X_i; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_l} \frac{\partial \ln \lambda(X_i; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{l'}}, \quad \hat{\boldsymbol{C}}_{lj} = \frac{1}{n}\sum_{i:X_i\in I_j} \delta_i \frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(X_i, \hat{\boldsymbol{\theta}}),$$

$$\hat{A}_j = U_j/n, \quad U_j = \sum_{i:X_i\in I_j} \delta_i, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - e_j),$$

$i = 1, \cdots, n, \quad j = 1, \cdots, k, \quad l, l' = 1, \cdots, m.$
The statistic $\hat{\boldsymbol{i}} = -\frac{1}{n}\ddot{\ell}(\hat{\boldsymbol{\theta}})$ is also a consistent estimator of $\boldsymbol{i}(\boldsymbol{\theta}_0)$ but it is prefered to use the above estimator to ensure that both components of the following RRN test statistic are non-negative for any n.
Denote by $\hat{g}^{ll'}$ the elements of $\hat{\boldsymbol{G}}^{-}$. The quadratic form Q can be written as follows:

$$Q = \sum_{l=1}^{m}\sum_{l'=1}^{m} W_l g^{ll'} W_{l'}.$$

The limit distribution of the statistic Y_n^2 is chi-square with $r = rank(\boldsymbol{V}^{-}) = Tr(\boldsymbol{V}^{-}\boldsymbol{V})$ degrees of freedom. If G is non-degenerate then $r = k$.

Statistical inference for the hypothesis H_0: The hypothesis is rejected with approximate significance level α if $Y_n^2 > \chi_\alpha^2(r)$.

4 Gooness-of-fit test for Gompertz and Weibull models

Let consider the hypothesis that under H_0 the distribution of the failure times is Gompertz with hazard function and commulative hazard function as;

$$\mu_x = \theta e^{\nu x}, \quad \Lambda_x = \frac{\theta}{\nu}(e^{\nu x} - 1) \quad x > 0, \quad (\theta, \nu) > 0,$$

The loglikelihood function is

$$\ell(\theta, \nu) = \sum_{i=1}^{n} \left\{ \delta_i [\ln \theta + \nu X_i] - \frac{\theta}{\nu}(e^{\nu X_i} - 1) \right\},$$

Let denote by $\hat{\theta}$ and $\hat{\nu}$ the ML estimator of θ and ν.
Since the matrix G is found to be degenerated so quadratic form can be written as:

$$Q = \frac{W_2^2}{\hat{g}_2 2}$$

where

$$\hat{g}_{22} = \hat{i}_{22} - \sum_{j=1}^{k} \hat{C}_{2j}^2 \hat{A}_j^{-1}, \quad \hat{i}_{22} = \frac{1}{n} \sum_{i=1}^{n} \delta_i X_i^2, \quad \hat{C}_{2j} = \frac{1}{n} \sum_{i: X_i \in I_j} \delta_i X_i,$$

$$\hat{A}_j = \frac{U_j}{n}, \quad W_2 = \sum_{j=1}^{k} \hat{C}_{2j} \hat{A}_j^{-1} Z_j, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - e_j)$$

. *Choice of* $\hat{a}_j$: Set

$$b_i = (n - i)\frac{e^{\hat{\nu} X_{(i)}} - 1}{\hat{\nu}} + \sum_{l=1}^{i} \frac{e^{\hat{\nu} X_{(l)}} - 1}{\hat{\nu}}, \quad i = 1, \cdots, n$$

If i is the smallest natural number verifying the inequalities

$$b_{i-1} \leq \frac{j}{k} b_n \leq b_i$$

then for $j = 1, \cdots, k-1$

$$\hat{a}_j = \frac{1}{\hat{\nu}} \ln \left\{ 1 + \hat{\nu} \left(\frac{j}{k} b_n - \sum_{l=1}^{i} \frac{e^{\hat{\nu} X_{(l)}} - 1}{\hat{\nu}} \right) / (n - i + 1) \right\}, \quad \hat{a}_k = \max(X_{(n)}, \tau)$$

For such choices of intervals we have $e_j = \hat{\theta} b_n / k$ for any j.

The zero hypothesis is rejected with an approximate significance level α if $Y_n^2 > \chi_\alpha^2(k)$.

Example: This example is taken from the industry. n = 120 electronic devices were observed for time $\tau = 5.54$ (years). The number of failures is δ =113. Suppose that the following failure times have a Gompertz distribution:

1.7440, 1.9172, 2.1461, 2.3079, 2.3753, 2.3858, 2.4147, 2.5404, 2.6205, 2.6471,
2.837, 2.8373, 2.8766, 2.9888, 3.0720, 3.1586, 3.1730, 3.2132, 3.2323, 3.3492,
3.3507, 3.3514, 3.3625, 3.3802, 3.3855, 3.4012, 3.4382, 3.4438, 3.4684, 3.5019,
3.5110, 3.5297, 3.5363, 3.5587, 3.5846, 3.5992, 3.654, 3.6574, 3.6674, 3.7062,
3.7157, 3.7288, 3.7502, 3.7823, 3.8848, 3.8902, 3.9113, 3.9468, 3.9551, 3.9728,
3.9787, 3.9903, 4.0078, 4.0646, 4.1301, 4.1427, 4.2300, 4.2312, 4.2525, 4.2581,
4.2885, 4.2919, 4.2970, 4.3666, 4.3918, 4.4365, 4.4919, 4.4932, 4.5388, 4.5826,
4.5992, 4.6001, 4.6324, 4.6400, 4.7164, 4.7300, 4.7881, 4.7969, 4.8009, 4.8351,
4.8406, 4.8532, 4.8619, 4.8635, 4.8679, 4.8858, 4.8928, 4.9466, 4.9846, 5.0008,
5.0144, 5.0517, 5.0898, 5.0929, 5.0951, 5.1023, 5.1219, 5.1223, 5.1710, 5.1766,
5.1816, 5.2441, 5.2546, 5.3353, 5.4291, 5.4360, 5.4633, 5.4842, 5.4860, 5.4903,
5.5199, 5.5232, 5.5335,

The maximum likelihood estimators of Gompertz model are; $\hat{\theta} = 0.0051$, $\hat{\nu} = 1.1586$. We take 10 intervals i.e. k=10. Further results to calculate the Y^2 are shown below:

j	1	2	3	4	5	6	7	8	9	10
$\hat{a_j}$	2.70	3.33	3.74	4.07	4.34	4.57	4.78	5.00	5.25	5.54
U_j	10	9	23	12	9	6	7	13	13	11
e_j	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3

$$\hat{i}_{22} = 16.7779, \quad \hat{g}_{22} = 0.0141, \quad W_2 = -0.3737.$$

The matrix G is degenerate, so $r = k - 1 = 9$. The value of test statistic is $Y^2 = X^2 + Q = 15.1130 + 9.8867 = 24.9997$, and the P-value is $pv = P\{\chi_9^2 > 24.9997\} = 0.0053$ ($CR = 16.9190$). So from the result we can say that failure times don't follow Gompertz distribution.

Weibull Distribution

Suppose that the failure times follow a Weibull model. The maximum likelihood estimators of Weibull model are; $\hat{\theta} = 4.6078$, $\hat{\nu} = 4.9554$. We take 10 intervals i.e. k=10. Further results to calculate the Y^2 are shown below:

j	1	2	3	4	5	6	7	8	9	10
$\hat{a_j}$	2.89	3.36	3.70	3.98	4.24	4.47	4.68	4.90	5.16	5.54
U_j	13	9	17	12	7	8	8	13	11	15
e_j	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3

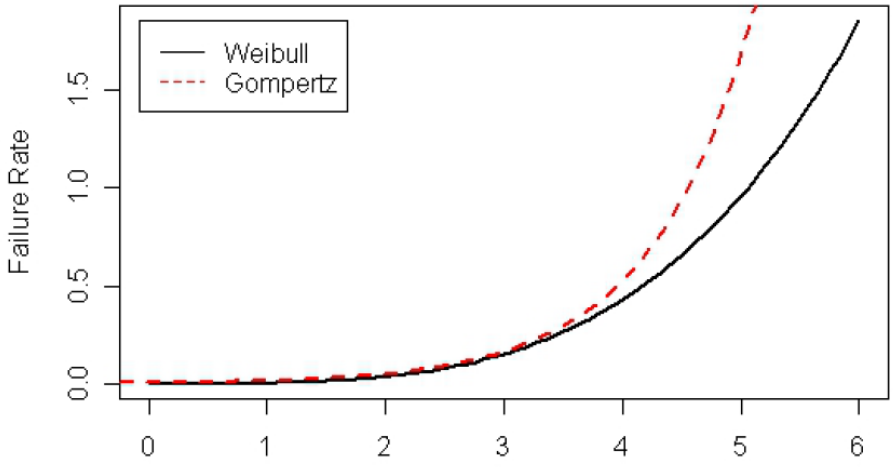


Fig4: The failure rate of electronic devices

$$\hat{i}_{22} = 0.0618, \quad \hat{g}_{22} = 0.0027, \quad W_2 = -0.0545.$$

The matrix G is degenerate, so $r = k - 1 = 9$. The value of test statistic is $Y^2 = X^2 + Q = 9.2692 + 1.0845 = 10.3536$, and the P-value is $pv = P\{\chi^2_9 > 10.3536\} = 0.3226$ ($CR = 16.9190$). So from the result we have no reason to reject that the failure times follow the Weibull distribution.

Here Weibull model gives the better fit, so it can be used as the alternative to the Gompertz model. As this data is related to the technical device and according to Gavrilov & Gavrilova (2001) technical deceives fails according to the Weibull law. Also from Figure 4 it can be seen the behavior of Gompertz model, that is, in later times the failure rate increases very fast.

5 Non-parametric estimation of survival function using the data from table of mortality

If there is no information about the model, one can estimate the survaival function by using non-parametric estimation method. In case of right censored individual sample (5), it is easy to use the well known Kaplan-Meier estimation method for estimating the survival function S_t and consequently the distribtuon function $1 - S_t$ of the failure times. One can estimate the survival function $S(t)$ from the grouped data, for example from the table of mortality with censoring in the following way.

Suppose we oberve n_0 individuals and the time sacle is divided in k intervals

$$[a_0, a_1[, \ [a_1, a_2[, \ ..., [a_{j-1}, a_j[, \ ... \ , [a_{k-1}, a_k[. \tag{9}$$

Consider the $j - th$ interval $I_j = [a_{j-1}, a_j[$ $(j = 1, ..., k)$, $a_0 = 0$, $a_k = +\infty$.

We obsereve d_j- the number of individals die in the interval I_j, $(j = 1, ..., k)$. c_j- the number of individuals censored in the interval I_j, $(j = 1, ..., k)$. and note n_j- le number of individuals at risque (not died and not censored) at time a_j, (The number of individuals who enter in the I_{j+1}-th interval). So

$$n_j = n_{j-1} - d_j - c_j. \tag{10}$$

Note that r_j- the number of individuals at death risk in the interval I_j. If all censores are at the start of the interval I_j, then

$$r_j = n_{j-1} - c_j. \tag{11}$$

If all censores are at the end of the interval I_j, then $r_j = n_{j-1}$. But the censoring times are unknown. Therefore we suppose that the censores are uniformally distributed in the interval I_j and hence we take

$$r_j = n_{j-1} - c_j/2. \tag{12}$$

Let denote

$$q_1 = P\{T > a_1\} = S(a_1),$$

and

$$q_j = P\{T > a_j | T > a_{j-1}\}$$

the conditional probability of being alive at a_j given that it was alive at a_{j-1}. So

$$q_j = \frac{P\{T > a_j\}}{P\{T > a_{j-1}\}},$$

and hence

$$P\{T > a_j\} = q_j \cdot P\{T > a_{j-1}\} =$$

$$q_j \cdot q_{j-1} P\{T > a_{j-2}\} = ... = q_j \cdot q_{j-1} \cdots q_2 \cdot q_1 = \prod_{i=1}^{j} q_i.$$

So we have

$$S_j = P\{T > a_j\} = \prod_{i=1}^{j} q_i.$$

Either

$$p_j = 1 - q_j = P\{T \leq a_j | T > a_{j-1}\},$$

the conditional probability of death in the interval I_j given that it was alive at a_{j-1}. So we can write

$$S_j = \prod_{i=1}^{j} (1 - p_i). \tag{13}$$

The number of deaths d_i in the interval I_i follow approximatally binomial law with parameters r_i and p_i, so $\mathbf{E}d_i \approx r_i p_i$ and the probability p_i is estimated by

$$\hat{p}_i = \frac{d_i}{r_i}. \tag{14}$$

So the suvival function $S_j = S(t_j)$ is estimated by

$$\hat{S}_j = \prod_{i=1}^{j}(1-\hat{p}_i) = \prod_{i=1}^{j}\left(1-\frac{d_i}{r_i}\right) = \prod_{i=1}^{j}\left(1-\frac{d_i}{n_i - c_i/2}\right). \tag{15}$$

This is the analog of Kaplan-Meier estimator for group data and it is also called product limit estimator.

References

[1] Akritas, M.G. (1988). Pearson-type goodness of fit tests: the univariate case, J. Amer.Statist.Assoc., 83, 222-230.

[2] Bagdonavicius, V. and Nikulin, M. (2002). Accelerated Life Models. Chapman and Hall/CRC, Boca Raton.

[3] Bagdonavicius V., Kruopis J., Nikulin M. (2010). *Nonparametric tests for Censored Data*, ISTE and J.Wiley.

[4] Bagdonavicius, V. and Nikulin, M. (2011) Chi-square goodness-of-fit test for right censored data. Int.J.Appl.Math.Stat.;Vol.24;Issue No.SI-11A, 30-50.

[5] Drost, F. (1988). Asymptotics for generalized chi-squared goodness-of-fit tests. CWI Tracs, Amsterdam: Centre for Mathematics and Computer Sciences, v.48.

[6] Greenwood, P. and Nikulin, M.S. (1988) Application of test of chi-square type, Journal of Soviet Mathematics, 43, No. 6, 2776-2791.

[7] Gavrilov L.A., Gavrilova N.S. (2001). The reliability theory of aging and longevity. Journal of Theoretical Biology, 213(4): 527-545.

[8] Gavrilov, L. A., Gavrilova, N. S. (1991). The Biology of Life Span: A Quantitative Approach. New York: Harwood Academic Publisher.

[9] Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, Philosophical Transactions of the Royal Society of London, Ser. A, 115, 513-585, reprint: Haberman S & Sibbett, TA (1995), Vol. II, 119-191.

[10] Habib, M.G., Thomas, D.R., Chi-squared goodness-of-fit tests for randomly censored data. The Annals of Statistics, v.14, 759-765.

[11] Haghighi, F. (2004). "Fiabilité et analyse de survie: Le modèle semi-paramétrique de dégradation. Une généralisation de la famille Weibull et ses applications." Thèse pour le Doctorat de l'Université Victor Segalen Bordeaux 2, 6 Décembre 2004, Bordeaux, France.

[12] Hjort, N.L. (1990). oodness of fit tests in models for life history data based on cumulative hazard rates. The Annals of Statistics, v.18,#3, 1221-1258.

[13] Hollander, M., Pena, E.Pena (1992). Chi-square goodness-of-fit test for randomly censored data. JASA, v.87, #417, 458-463.

[14] Izsak, J. and Gavrilov, L.A. (1995). A typical interdisciplinary topic: questions of the mortality dynamics. Archives of Gerontology and Geriatrics, 20, 283-293.

[15] Juckett DA, Rosenberg B. (1993). Comparison of the Gompertz and Weibull functions as descriptors for human mortality distributions and their intersections. Mech Ageing Dev. ;69:131.

[16] Kim, J.H. (1993). Chi-square goodness-of-fit tests for randomly censored data. The Annals of Statistics, v.21, 1621-39.

[17] LeCam, L., Mahan,C., Singh, A. (1983). An extension of a Theorem of H.Chernoff and E.L.Lehmann. In : Recent advances in statistics, Academic Press, Orlando, 303-332.

[18] Mandala, A., Huangb,W.T., Bhandaric,S.K., Basuc,A. (2011). Goodness-of-fit testing in growth curve models: A general approach based on finite differences. Computational Statistics & Data Analysis, Volume 55, Issue 2, 1086-1098.

[19] Nikulin , M.S. (2011). Modeles statistiques en demographie, gerontologie et assurance sur la vie. Lecture Notes for IMB Groupe de Fiabilité et Statistique, Université Victor Segalen Bordeaux 2, p67.

[20] Nikulin, M.S. (1973). Chi-square test for normality, International Vilnius Conference on Probability Theory and Mathematical Statistics, 2, 119-122.

[21] Nikulin, M.S. (1973). Chi-square test for continuous distribution with shift and scale parameters, Theory of Probability and its Application, 19, No.3, 559-568.

[22] Nikulin, M.S. (1973). On a chi-square test for continuous distribution, Theory of Probability and its Application, 19, No.3, 638-639.

[23] Ricklefs, R.E., A. Scheuerlein (2002). Biological implications of the Weibull and Gompertz model of aging. Journals of Gerontology: Biological Sciences 57A, 2 B69-B76.

[24] Rao, k.C., and Robson, D.S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family, Communication in Statistics, 3, 1139-1153.

[25] Van der Vaart, A.W. (1998). Asymptotic Statistics. Cambridge : UK.

[26] Vanfleteren, J. R., De Vreese, A., Braeckman, B. P. (1998). Two-parameter logistic and Weibull equations provide better fits to survival data from isogenic populations of Caenorhabditis elegans in axenic culture than does the Gompertz model. J. Gerontol. Ser. A 53, B393-403.

[27] Wang Jane-Ling, Muller Hans-Georg and Capra William B. (1998). Analysis of Oldest-Old Mortality: Lifetables Revisited. The Annals of Statistics, Vol. 26, No. 1, 126-163.

[28] Wilson, D. . (1994). The analysis of survival mortality. data: fitting Gompertz, Weibull, and logistic functions. Mechanisms of Ageing and Development, 74, 15-33.

[29] Zhang, B. (1999). A chi-squared goodness-of-fit test for logistic regression models based on case-control data. Biometrika, v.86, 3, 531-539.

Conditional Distributions and Scaling for Categorial Data

Henning Läuter
Mathematical Institute
Potsdam, Germany
e-mail: laeuter@uni-potsdam.de

Abstract

We consider continuous random variables, look at conditional distributions and find that the estimation of unknown conditions can be used for modeling categorical data. We introduce maximum likelihood estimates for the conditional parameters and so we find scaled values for the categories by maximum likelihood. With these methods the scaling for the levels of the factors in models of analysis of variance can be solved. Under normal distributions explicit solutions are given, for other distributions, e.g. for survival distributions, the scaling is described by copulas and different methods of estimates.

Keywords: Conditional distributions, modeling, categorical data, statistical scaling.

Introduction

In multivariate statistics several random variables are to be simultaneously investigated. One will find connections between variables or the response of some variables on others. These variables can be metric variables or categorical variables. One will explain the influence of the endogenous variables on the response variables. If all variables are metric variables then often a regression model will be used. If the response is metric, but the endogenous factors are categorical variables then an analysis of variance can help. For categorical response variables a discriminant analysis or logistic regression is used. In Agresti [1] and Everitt and Dunn [3] many methods for statistical analysis are described. We will concentrate on statistical scaling of categorical variables. Mostly scaling is connected with proximities or distances, we consider a statistical approach as in Läuter and Ramadan [4] or in Ramadan [5]. In this paper a scaling will base on the estimation of parameters in conditional distributions. We will see which parameters in such families are estimable and find estimates for those. The conditional parameters will be used for methods of scaling of categorical data and can be used for the interpretation of the results of the analysis of the data. Some other methods for statistical scaling bases on tests and lead to most separating scales as it was considered by Ahrens and J. Läuter in [2] and by Ramadan in [5].
Generally we consider a blocked random variable (Z_1, Z_2) where the variable Z_1 is the endogenous variable and Z_2 is the response. We are looking at the conditional variable $\tilde{U} := Z_2 \,|\, Z_1 = t$ and we will find the influence of t in this variable. We admit for Z_1 both metric and categorical variables. If Z_1 is a metric variable and t is known then often the regression

$$r(t) := \mathsf{E}(Z_2 \,|\, Z_1 = t) \tag{1}$$

is studied, nonparametric or parametric regression. If Z_1 is a categorical variable then we will find tools for representing the analyzed data. Here we consider the statistical scaling and find points in metric spaces equivalent to the categorical values.

1 Conditional Normal Distribution

We consider a $(m_1 + m_2)$-dimensional normally distributed random variable Z with

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathsf{N}_{m_1+m_2}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) . \tag{2}$$

Here we have $Z_1, \mu_1 \in R^{m_1}, Z_2, \mu_2 \in R^{m_2}$ and Σ_{11} is a $m_1 \times m_1$ matrix and the others correspondingly. The variance matrix $\Sigma = \mathsf{v}arZ$ should be nonsingular.Then the conditional distribution of Z_2 given $Z_1 = t$ is well known as

$$Z_2 \,|\, Z_1 = t \;\sim \mathsf{N}_{m_2}\Big(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(t - \mu_1), \Sigma_{2\cdot 1}\Big) \tag{3}$$

with $\Sigma_{2\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. This distribution depends on the parameters μ_1, μ_2, Σ, t. Obviously these parameters are not identifiable. Assuming for positive integers $n_1, ..., n_L$ and values $t_1, ..., t_L$ we have for $l = 1, ..., L$ variables $U_{l1}, ..., U_{ln_l}$ with

$$U_{lj} \;\sim\; \mathsf{N}_{m_2}\Big(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(t_l - \mu_1), \Sigma_{2\cdot 1}\Big), \quad j = 1, ..., n_l. \tag{4}$$

All these variables should be independent, the sample sizes n_l should be known. Then we see that the parameters μ_1, μ_2, Σ determine the distribution of Z, but the parameters $t_1, ..., t_L$ determine the localizations where the first m_1-dimensional subvector of the U_{lj} are fixed. Therefore the whole parameter $(\mu_1, \mu_2, \Sigma, t_1, ..., t_L)$ consists of global distributional parameters, global for all variables, and localized parameters. Although e.g. t_l and μ_1 influence the distribution of U_{lj} in the same way their meaning is completely different. μ_1 describes the expectation of the first subvector and the t_l describes the point where Z_1 is fixed. We see immediately that the parameters

$$\tau_l = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(t_l - \mu_1) \tag{5}$$

are estimable but not any of the parameters $\mu_1, \mu_2, t_1, ..., t_L$ separately. Moreover all the differences

$$\delta_{ls} := \Sigma_{21}\Sigma_{11}^{-1}(t_l - t_s) \tag{6}$$

are estimable. This means that we can determine relations between the points where the first subvectors are to be observed. In the special case $m_1 = 1$ we can estimate the values $t_2 - t_1, ..., t_L - t_1$ up to a factor and these estimations can be ordered on a line. For $m_2 = 1$ we represent the estimations of the real values $\Sigma_{21}\Sigma_{11}^{-1}(t_2 - t_1), ..., \Sigma_{21}\Sigma_{11}^{-1}(t_L - t_1)$.

Theorem 1. *Assuming (2) and let the variables* $\{U_{lj}, \quad j = 1, ..., n_l; l = 1, ..., L.\}$ *be independent and the variables* $\{U_{lj}, \quad j = 1, ..., n_l\}$ *are identically distributed and it is known that* U_{lj} *has a conditional distribution as in (3) and (4) for unknown* t_l*. There is no connection between the* t_l*'s. Then the best unbiased estimate for* $\delta_{ls} = \Sigma_{21}\Sigma_{11}^{-1}(t_l - t_s)$ *is given by*

$$\hat{\delta}_{ls} = \bar{U}_{l\cdot} - \bar{U}_{s\cdot} \tag{7}$$

for

$$\bar{U}_{l\cdot} = \frac{1}{n_l}\sum_{j=1}^{n_l} U_{lj} = \frac{1}{n_l}U_{l+} \quad . \tag{8}$$

The values $\nu = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1$ and $t_1, ..., t_L$ are not estimable but all differences δ_{ls}. From this follows that we can choose in an arbitrary way the origin, i.e. an arbitrary $\hat{t}_1$, say $\hat{t}_1 = 0$ and then all other $\Sigma_{21}\Sigma_{11}^{-1}\hat{t}_2, ..., \Sigma_{21}\Sigma_{11}^{-1}\hat{t}_L$ are calculable.
We see that for normal distributions the localized parameters δ_{ls} can be estimated from observations of the conditioned distributions. One can choose the origin by fixing the $\hat{t}_1 = 0$ and then the other parameters $\Sigma_{21}\Sigma_{11}^{-1}\hat{t}_l, \quad l = 1, ..., L$ are determined. It is clear from (4) that the $t_1, ..., t_L$ cannot be estimated but relations between them. We remember that the values of the localized parameters can be identified with the levels under which the observations are taken.
From the distribution it follows that Σ_{11} and Σ_{21} are not estimable. Therefore these parameters have to be assumed known or it is sufficient to estimate the δ_{ls}.

1.1 Estimation of parameters

If the conditions $t_1, ..., t_L$ are functionally completely independent then the estimates for the δ_{ls} are given in Theorem 1. But there are important cases of different $t_1, ..., t_L$ where in parts connections exist between them. We assume that $\tau = (t_1', ..., t_L')'$ varies in a p-dimensional subspace of $R^{Lm_1}, \quad p < Lm_1$ and p is as small as possible. Then we know $m_1 \times p$ matrices $C_1, ..., C_L$ with

$$t_l = C_l\beta, \quad l = 1, ..., L, \beta \in R^p \tag{9}$$

and we obtain

$$U_{lj} = \nu + \Sigma_{21}\Sigma_{11}^{-1}t_l + \varepsilon_{lj} = \nu + \Sigma_{21}\Sigma_{11}^{-1}C_l\beta + \varepsilon_{lj}. \tag{10}$$

With the Kronecker product $A \otimes B$ of matrices A, B and 1_k the k-dimensional vector of ones, I_k the $k \times k$ identity matrix and

$$E_n = \begin{pmatrix} 1_{n_1} & 0 & 0 & \dots & 0 \\ 0 & 1_{n_2} & 0 & \dots & \vdots \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & & 0 & 1_{n_L} \end{pmatrix}, \tag{11}$$

we get

$$U = (1_n \otimes I_{m_2})\nu + (E_n \otimes \Sigma_{21}\Sigma_{11}^{-1}) \begin{pmatrix} C_1 \\ \vdots \\ C_L \end{pmatrix} \beta + \varepsilon = X \begin{pmatrix} \nu \\ \beta \end{pmatrix} + \varepsilon \tag{12}$$

with $U = (U_{11}', ..., U_{Ln_L}')', \quad \varepsilon = (\varepsilon_{11}', ..., \varepsilon_{Ln_L}')'$ and

$$X = \left(1_n \otimes I_{m_2} \vdots (E_n \otimes \Sigma_{21}\Sigma_{11}^{-1}) \begin{pmatrix} C_1 \\ \vdots \\ C_L \end{pmatrix}\right). \tag{13}$$

Then for a matrix F with m_2+p columns the best unbiased estimation for an estimable $F(\nu',\beta')'$ is

$$F\begin{pmatrix}\hat{\nu}\\ \hat{\beta}\end{pmatrix} = F\left(X'(I_n\otimes\Sigma_{2\cdot1}^{-1})X\right)^{+}X'(I_n\otimes\Sigma_{2\cdot1}^{-1})u \tag{14}$$

for a realization u from (12). As a consequence the optimal estimate for δ_{ls} is

$$\hat{\delta}_{ls} = \Sigma_{21}\Sigma_{11}^{-1}(C_l - C_s)\,\hat{\beta}, \quad l = 1,...,L. \tag{15}$$

Theorem 2. *Assuming (2) and let the variables*

$$U_{lj}, \quad j = 1,...,n_l; l = 1,...,L. \tag{16}$$

be independently distributed. The variables U_{lj}, $j = 1,...,n_l$ are identically distributed and it is known that U_{lj} has a conditional distribution as in (3) and (4) for unknown t_l. There exist such $m_1\times p$ matrices $C_1,...,C_L$ that (9) is fulfilled. Then $\hat{\delta}_{ls}$ from (15) is the maximum likelihood estimator for δ_{ls}.

1.1.1 Cross-classified designs

We consider a two-way cross-classified design. This means that we have observations as in Table 1. We consider the observations $u_{(i,k)j}$ as a realization of U_{lj} with (4) for $t_l = (s_{1i}', s_{2k}')'$. The

Table 1: Two-way table

	Factor B				
Factor A	s_{21}	s_{22}	s_{23}	$\cdots$	s_{2b}
s_{11}	$u_{(11)j}$ $j=1,...,n_{11}$	$u_{(12)j}$ $j=1,...,n_{12}$	$u_{(13)j}$ $j=1,...,n_{13}$	$\cdots$	$u_{(1b)j}$ $j=1,...,n_{1b}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s_{1a}	$u_{(a1)j}$ $j=1,...,n_{a1}$	$u_{(a2)j}$ $j=1,...,n_{a2}$	$u_{(a3)j}$ $j=1,...,n_{a3}$	$\cdots$	$u_{(ab)j}$ $j=1,...,n_{ab}$

conditions for Z_1 are the values

$$t_1 = (s_{11}', s_{21}')', t_2 = (s_{11}', s_{22}')', t_3 = (s_{11}', s_{23}')', ..., t_b = (s_{11}', s_{2b}')', \tag{17}$$

$$t_{b+1} = (s_{12}', s_{21}')', ..., t_L = (s_{1a}', s_{2b}')' \tag{18}$$

with $s_{1i}\in R^{m_{11}}, s_{2k}\in R^{m_{12}}, m_{11}+m_{12} = m_1 \geq 2$. This means we have a two-way classification, altogether with $L = ab$ categories and any value of the factor A (first m_{11} components) is combined

with any value of factor B (second m_{12} components). For such a design we have (9) for $p = am_{11} + bm_{12} < abm_1 = Lm_1$. If we define

$$\beta = (s'_{11}, s'_{12}, ..., s'_{1a}, s'_{21}, ..., s'_{2b})' \tag{19}$$

then we find C_l with (9). These are block matrices and especially we have, with $0_{(11)} := 0_{m_{11}\times m_{11}}, 0_{(22)} := 0_{m_{12}\times m_{12}}, 0_{(12)} := 0_{m_{11}\times m_{12}}$, $0_{(21)} := 0_{m_{12}\times m_{11}}$

$$C_1 = \begin{pmatrix} I_{m_{11}} & 0_{(11)} & \cdots & 0_{(11)} & 0_{(12)} & \cdots & \cdots & 0_{(12)} \\ 0_{(21)} & \cdots & \cdots & 0_{(21)} & I_{m_{12}} & 0_{(22)} & \cdots & 0_{(22)} \end{pmatrix}. \tag{20}$$

Here $I_{m_{11}}$ stands in the first column position, $I_{m_{12}}$ stands in the column position $a+1$. All other matrices C_l are of the same structure, $I_{m_{11}}$ vary in the first row and the first a positions and $I_{m_{12}}$ vary in the second row and the positions $a+1$ till $a+b$. Then we calculate the estimate $\hat{\beta}$ in (14) and the estimates $\hat{\delta}_{ls}$ from (15).
Summarizing in this standard model where the random variables are normally distributed and the conditional variable will be observed at unknown points we see that the expectations of the variables Z_1 and Z_2 are not estimable, the conditional points $t_1, ..., t_L$ are not estimable, the variance of Z_1 and the correlation between Z_1 and Z_2 are not estimable. The contrasts $\delta_{ls} = \Sigma_{21}\Sigma_{11}^{-1}(C_l - C_s)\beta$ are estimable.
We cannot expect that in general distributional families the possibilities for estimated parameters are completely better. We consider general continuous random variables in the next section, here the principle will be repeated in an obvious way.

2 Continuous conditional distributions

Assuming that the $m_1 + m_2$-dimensional variable $Z = (Z'_1, Z'_2)'$ has any continuous distribution. The one-dimensional marginal densities of the components of Z_1 and Z_2 are denoted by $f_{11}, ..., f_{1m_1}$ and $f_{21}, ..., f_{2m_2}$ and the distribution functions by F_{1i} and F_{2j} respectively. The dependence between Z_1 and Z_2 is described by a copula function C. We denote $F_1(\xi_1) = \Big(F_{11}(\xi_{11}), ..., F_{1m_1}(\xi_{1m_1})\Big)$ and $F_2(\xi_2) = \Big(F_{21}(\xi_{21}), ..., F_{2m_2}(\xi_{2m_2})\Big)$. Then the joint distribution function of (Z_1, Z_2) is

$$F(\xi_1, \xi_2) = C\Big(F_1(\xi_1), F_2(\xi_2)\Big). \tag{21}$$

For any distribution F with the marginal distributions F_{1i}, F_{2j} such a copula function exists and is uniquely determined. Assuming that the densities of F, F_{1i}, F_{2j} and of the copula C exist then we have for the joint density f the representation

$$f(\xi_1, \xi_2) = c\Big(F_1(\xi_1), F_2(\xi_2)\Big) f_{11}(\xi_{11}) \cdot ... \cdot f_{1m_1}(\xi_{1m_1}) \, f_{21}(\xi_{21}) \cdot ... \cdot f_{2m_2}(\xi_{2m_2}). \tag{22}$$

Consequently the density of $Z_2 \,|\, Z_1 = t$ with $t = (t_{(1)}, ..., t_{(m_1)})'$ is

$$f^{Z_1=t}(\xi_2) = c\Big(F_1(t), F_2(\xi_2)\Big) f_{11}(t_{(1)}) \cdot ... \cdot f_{1m_1}(t_{(m_1)})) \, f_{21}(\xi_{21}) \cdot ... \cdot f_{2m_2}(\xi_{2m_2}) / f^{Z_1}(t). \tag{23}$$

Therefore the copula function determines the dependence of Z_1 and Z_2 and hence the copula function is an essential part of the statistical model.
Now we assume that for positive integers $n_1, ..., n_L$ and values $t_1, ..., t_L$ and for all $l = 1, ..., L$ the variables $U_{l1}, ..., U_{ln_l}$ are identically distributed with the density $f^{Z_1=t_l}$ from (23). All these variables should be independent, the sample sizes are known. We assume that the distribution function F_1 of Z_1 depends on a parameter $\vartheta_1 \in R^{p_1}$, the F_2 depends on $\vartheta_2 \in R^{p_2}$. Then the global parameter for the distribution of the sample is $\vartheta = (\vartheta_1, \vartheta_2)$. We observe $u = (u_{11}, ..., u_{Ln_L})$ with $u_{lj} = (u_{lj,1}, ..., u_{lj,m_2})'$. The likelihood function for the whole sample is with $t_l = (t_{l1}, ..., t_{lm_1})'$

$$\mathcal{L}(\vartheta, t_1, ..., t_L; u) = \prod_{l=1}^{L} \prod_{j=1}^{n_l} c\Big(F_1(t_l; \vartheta_1), F_2(u_{lj}; \vartheta_2)\Big) f_{11}(t_{l1}; \vartheta_1) \cdot ...$$

$$... \cdot f_{1m_1}(t_{lm_1}; \vartheta_1)\, f_{21}(u_{lj,1}, \vartheta_2) \cdot ... \cdot f_{2m_2}(u_{lj,m_2}, \vartheta_2)/f^{Z_1}(t_l; \vartheta_1). \tag{24}$$

Now we can define the maximum likelihood estimation of the parameters. Of course maximization of the likelihood goes over the possible parameter set. If the $t_1, ..., t_L$ are unrestricted then we maximize $\mathcal{L}$ without additional conditions. If there are restrictions as in (9) Then we substitute $t_l = C_l\, \beta$ and consider $\mathcal{L}$ as a function of ϑ, β. Then we proceed as in the preceding section.

2.1 Normal copula

Obviously the maximum likelihood estimation depends on the copula, i.e. it depends clearly on the model which we have assumed. In the special case of a normal copula we will look at the properties. Let Φ be the distribution of $\mathsf{N}(0, 1)$ and Φ_R be the distribution function of $\mathsf{N}_m(0, R)$ where the variance matrix R is at the same time a correlation matrix. Then the normal copula has the form

$$C_R^N(u_1, ..., u_m) = \Phi_R(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_m)) \tag{25}$$

$$= \frac{1}{(2\pi)^{m/2}|R|^{1/2}} \int_{-\infty}^{\Phi^{-1}(u_1)} \mathrm{d}ots \int_{-\infty}^{\Phi^{-1}(u_m)} \exp(-\frac{1}{2}\, \xi' R^{-1} \xi)\, \mathrm{d}\xi \tag{26}$$

Denoting the copula density of C_R^N by c_R^N then the conditional density in (23) has the form

$$f^{Z_1=t}(\xi_2) = c_R^N\Big(F_1(t; \vartheta_1), F_2(\xi_2; \vartheta_2)\Big) f_{11}(t_{(1)}; \vartheta_1) \cdot ... \cdot f_{1m_1}(t_{(m_1)}; \vartheta_1) \cdot ...$$

$$... \cdot f_{21}(\xi_{21}, \vartheta_2) \cdot ... \cdot f_{2m_2}(\xi_{2m_2}, \vartheta_2)/f^{Z_1}(t; \vartheta_1). \tag{27}$$

The copula density has the form

$$c_R^N(u_1, ..., u_m) = \frac{1}{|R|^{1/2}} \exp\Big(-\frac{1}{2}[Q_R(u) - \Phi^{-1}(u_1)^2 - ... - \Phi^{-1}(u_m)^2]\Big) \tag{28}$$

with the quadratic form

$$Q_R(u) = \Big(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_m)\Big)\, R^{-1}\, \Big(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_m)\Big)'. \tag{29}$$

Under $m_2 = 1$ the correlation matrix R in a special standardized form is of interest. This is

$$R = R_0 := \begin{pmatrix} & & & \rho_1 \\ & I_{m_1} & & \vdots \\ & & & \rho_{m_1} \\ \rho_1 & \rho_2 & \cdots & \rho_{m_1} \quad 1 \end{pmatrix} \tag{30}$$

and we have

$$|R_0| = 1 - \sum_{j=1}^{m_1} \rho_j^2 \tag{31}$$

and R^{-1} can be calculated.

2.1.1 One-dimensional case

Consider the special case $m = 2$, i.e. Z_1 and Z_2 are one-dimensional. Then Φ_R is the distribution function of $\mathsf{N}_2(0, R)$ with the variance matrix

$$R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{32}$$

and we obtain the copula density

$$c_R^N(u_1, u_2) = \frac{1}{(1-\rho^2)^{1/2}} \quad \exp\Big(- \tfrac{1}{2(1-\rho^2)} [\Phi^{-1}(u_1)^2 - 2\rho\Phi^{-1}(u_1)\Phi^{-1}(u_2) + \Phi^{-1}(u_2)^2] + \\ + \tfrac{1}{2}(\Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)^2)\Big). \tag{33}$$

3 Scaling of categorical data

Some technical or medical products, some behaviors or properties are described by categorical data. The perception of the objects should be represented in a metric space in such a way that distances are expressive. A geometric nearness should reflect a statistical proximity or similarity of objects.

In a statistical background scaling bases on probabilities. Conditions should be interpreted as near if observed probabilities under these conditions are near. A statistical model for categorical data should describe the important observed properties. For instance, some levels of factors and/or the response variables in models of analysis of variance are categorically. In the sense of scaling one will find metric values for the levels (categories) and the methods for finding a scaling base on statistical decisions. Here we will use the results from the preceding sections. The idea of scaling includes the fact that levels which lead to almost the same or at least lying close together probabilities should have similar scaling values. And if under some levels the observed

probabilities are very different from those in other levels then the corresponding scaling values should have a large distance. Most models for categorical data should have related properties as the monotone likelihood ratio in one-parametric families or some ordering of the distributions is possible. Hence very often the normal distribution will be taken in the model. But of course other families are possible if there is a good motivation for another choice.

Definition 1. *Given a blocked random variable (Z_1, Z_2). Let $\mathcal{P}$ be a statistical model for Z. Let L be the number of categories which are described by $t_1, ..., t_L$ as values of Z_1. Assuming for positive integers $n_1, ..., n_L$ for each l there are identically distributed variables $U_{l1}, ..., U_{ln_l}$ with the density $f^{Z_1=t_l}$. All these variables for $l = 1, ..., L$ should be independent. These variables $U_{l1}, ..., U_{ln_l}, \quad l = 1, ..., L$ are to be observed and we denote these observations by $u = (u_{11}, ...u_{1n_1}, u_{21}, ..., u_{Ln_L})$. Any estimation of $t_1, ..., t_L$ in $\mathcal{P}$ is called a scale of the categories $t_1, ..., t_L$.*

There is a variety of different scales, depending on the model and the principle of estimating. We can apply the results from the preceding sections. If the observed data are assumed to be realizations from normally distributed variables and we use the results from section 1 then the model says that we observed realizations from $Z_2 \mid Z_1 = t$ for different unknown t which are to be estimated.

Example: *We measure electrical resistances under different conditions. These conditions are determined by two factors (temperatures, geometrical forms of the sticks). We have $L = 8$ categories. As a sake of modeling we choose $\Sigma_{11} = I_2, \Sigma_{21} = \rho(1, 1)$ and $\rho = 0.5$. Using an appropriate shift translation we obtain the scales for the levels of both factors. The results have a strong technological meaning.*

Table 2: Resistances

	Factor B			
Factor A	level 1 (s_{21})	level 2 (s_{22})	level 3 (s_{23})	level 4 (s_{24})
level 1 (s_{11})	64.8 63.2	65.7 63.1	63.9 62.5	65.2 63.2
	61.7 65.2	62.8 66.3	59.9 64.9	62.1 65.9
level 2 (s_{12})	60.2 58.9	64.5 62.9	61.9 60.5	62.5 62.1
	58.3 60.6	60.1 66.0	59.2 60.1	60.1 62.6

factor A levels	1	2
scaled categories	0	2.60

Table 3: Scaled categories

factor B levels	1	2	3	4
scaled categories	0	2.562	0.287	1.60

References

[1] Agresti A.(2002). *Categorical Data Analysis*. Wiley, New York.

[2] Ahrens H., Läuter J. (1981). *Mehrdimensionale Varianzanalyse*. Akademie-Verlag, Berlin.

[3] Everitt B.S., Dunn G. (2001) *Applied Multivariate Data Analysis*. Hodder Education, London.

[4] Läuter H., Ramadan A.M. (2010). Modeling and scaling of categorical data. *Preprint 03/2010 Univ. Potsdam*.

[5] Ramadan A.M. (2010) Statistical Model for Categorical Data. *Doct. Thesis Univ. Potsdam*.

Goodness-of-Fit Testing, Smoothing and Resampling Under Censoring

Hannelore Liero
Institute of Mathematics, University of Potsdam
Potsdam, Germany
e-mail: liero@uni-potsdam.de

Abstract

Goodness-of-fit tests using kernel estimators for the density and the hazard rate are proposed. The aim is to investigate these tests for data where censoring is present. A third test statistic, the continuous analogue of the χ^2-test statistic based on the counts of uncensored observations, is considered. Resampling methods for the realization of the test procedures are discussed.

Keywords: censoring; goodness of fit; kernel estimators; resampling

Introduction

We consider the problem of testing whether the distribution of life times of interest belongs to a certain class of parametric distributions. Often in applications it occurs, that the observations are subject to random right censoring. Thus, we will consider the problem of goodness-of-fit testing under censoring.

Several proposals for test procedures based on chi-square statistics are investigated in [2],[4], [5] and [1]. Here following up the idea proposed in [6] the method of kernel smoothing to estimate the density and the hazard rate is applied to derive asymptotic α-tests of L_2-type. Moreover, a continuous analogue of the χ^2-test based on the number of uncensored observations is considered. The characterizing function is denoted by ξ.

The properly standardized test statistic converges in distribution to a normal distribution. It will turn out, that this limit distribution does not always provide a good approximation of the distribution of the test statistic. Thus, the application of resampling methods should be discussed. In the case without censoring it is obvious how one should resample. In the case of censored data it is not clear how one can generate the "hypothetical situation". The unknown distribution of the censoring variables can be considered as an unknown parameter – unfortunately as an infinite dimensional parameter. Since this problem occurs also in other contexts some ideas about resampling under censoring are discussed.

We start with some notation: Let $Y_1, \ldots, Y_n$ be a sequence of i.i.d. survival times with absolutely continuous distribution function F. As often occurs in applications the Y_i's are subject to random right censoring, i.e. the observations are

$$T_i = \min(Y_i, C_i) \quad \text{and} \quad \Delta_i = 1(Y_i \leq C_i)$$

where $C_1, \ldots, C_n$ are i.i.d. random censoring times which are independent of the Y- sequence. The Δ_i indicates whether Y_i has been censored or not.

We wish to test whether F lies in a parametric class of functions, i.e.

$$\mathcal{H}: \quad F \in \mathcal{F} \,=\, \{F(\cdot,\theta)\,|\,\theta \in \Theta \subseteq \mathbb{R}^k\} \quad \text{versus} \quad \mathcal{K}: \quad F \notin \mathcal{F}.$$

1 Nonparametric Estimators of the Density, the Hazard Rate and of a function ξ

We assume that the underlying random variables (r.v.'s) are absolutely continuous with density f. The hazard rate is defined by

$$\lambda(t) \,=\, \lim_{s \downarrow 0} \frac{1}{s} \mathsf{P}(t \le Y_i < t+s | Y_i \ge t) \,=\, \frac{f(t)}{1-F(t)}.$$

Since no parametric form of the alternative is assumed we will use a nonparametric estimators. To describe the estimation procedure we introduce the distribution function of the observations T_i and the subdistribution function of the pairs $(T_i, \Delta_i = 1)$

$$H(t) := \mathsf{P}(T_i \le t) \quad \text{and} \quad H^U(t) := \mathsf{P}(T_i \le t, \Delta_i = 1).$$

Since

$$1 \,-\, H(t) \,=\, (1 \,-\, G(t))\,(1 \,-\, F(t))$$

and

$$H^U(t) \,=\, \int_0^t (1 - G(s_-))\,\mathrm{d}F(s),$$

where G is the distribution function of the censoring times C_i, the cumulative hazard function

$$\Lambda(t) := \int_0^t \lambda(s)\,\mathrm{d}s$$

can be written for all $t < \tau_H$ with $\tau_H = \inf\{t:\, H(t) = 1\}$ as

$$\Lambda(t) \,=\, \int_0^t \frac{\mathrm{d}F(s)}{1-F(s_-)} \,=\, \int_0^t \frac{\mathrm{d}H^U(s)}{1-H(s_-)}.$$

Further, set $\xi(t) = (1-G(t_-))f(t)$. Now, for estimating Λ we replace H^U and H by their empirical versions, that is by

$$\widehat{H}_n^U(t) \,=\, \frac{1}{n}\sum_{i=1}^n \mathsf{1}(T_i \le t, \Delta_i = 1) \quad \text{and} \quad \widehat{H}_n(t) \,=\, \frac{1}{n}\sum_{i=1}^n \mathsf{1}(T_i \le t).$$

The resulting estimator

$$\widehat{\Lambda}_n(t) \;=\; \int_0^t \frac{\mathrm{d}\widehat{H}_n^U(s)}{1-\widehat{H}_n(s_-)} \;=\; \sum_{i=1}^{n} \frac{\mathbf{1}(T_{[i]} \le t)\,\Delta_{(i)}}{n-i+1}$$

is the Nelson–Aalen estimator of Λ. Here $T_{[1]} \le \cdots \le T_{[n]}$ are the ordered observations and $\Delta_{(i)} = \Delta_j$ iff $T_j = T_{[i]}$.

From the relationship between integrated hazard function and survival function we obtain the Kaplan–Meier estimator for the survival function

$$\widehat{S}_n(t) \;=\; \prod_{s\le t}(1 \;-\; \Delta\widehat{\Lambda}_n(s)) \;=\; \prod_{\substack{T_{[i]}\le t\\ \Delta_{(i)}=1}} \left(\frac{n-i}{n-i+1}\right), \qquad \widehat{F}_n(z) = 1-\widehat{S}_n(z).$$

As estimator of the derivative of and F and Λ we define the kernel smoothed versions of the negative Kaplan–Meier estimator and of the Nelson–Aalen estimator, respectively:

$$\widehat{f}_n(t) = \frac{1}{b_n}\int K\left(\frac{t-z}{b_n}\right)\,\mathrm{d}\widehat{F}_n(z)$$

and

$$\widehat{\lambda}_n(t) = \frac{1}{b_n}\int K\left(\frac{t-s}{b_n}\right)\,\mathrm{d}\widehat{\Lambda}_n(s).$$

The function ξ is estimated by

$$\widehat{\xi}_n(t) \;=\; \frac{1}{b_n}\int K\left(\frac{t-s}{b_n}\right)\,\mathrm{d}\widehat{H}_n^U(s).$$

Here K is a kernel function and $\{b_n\}$ is a sequence of bandwidths tending to zero with an appropriate rate. For simplicity we will use the notation $K_b(x) = K(x/b)/b$ and $b = b_n$.

We can write these estimators in the following equivalent form:

$$\begin{aligned}
\widehat{f}_n(t) &= \frac{1}{n}\sum_{i=1}^{n} K_b(t-T_i)\,\Delta_i \frac{1-\widehat{F}_n(T_{i-})}{1-\widehat{H}_n(T_{i-})},\\
\widehat{\lambda}_n(t) &= \sum_{i=1}^{n} \frac{K_b(t-T_{[i]})\,\Delta_{(i)}}{n-i+1}\\
&\text{and}\\
\widehat{\xi}_n(t) &= \frac{1}{n}\sum_{i=1}^{n} K_b(t-T_i)\Delta_i.
\end{aligned}$$

2 Asymptotic Normality of the Weighted Integrated Squared Error

As test statistics we will consider a properly weighted integrated squared distance between the estimators proposed above and the smoothed versions of the corresponding hypothetical functions. Define

$$\tilde{f}_n(t) \;=\; \int K_b(t-s)\,\mathrm{d}F(s), \qquad \tilde{\lambda}_n(t) \;=\; \int K_b(t-s)\,\mathrm{d}\Lambda(s)$$

and

$$\tilde{\xi}_n(t) \;=\; \int K_b(t-s)(1-\widehat{H}_n(s_-))\,\mathrm{d}\Lambda(s)$$

We consider

$$Q_n^f = \int \Big(\widehat{f}_n(t) \;-\; \tilde{f}_n(t)\Big)^2 a(t)\,\mathrm{d}t, \qquad Q_n^\lambda = \int \Big(\widehat{\lambda}_n(t) \;-\; \tilde{\lambda}_n(t)\Big)^2 a(t)\,\mathrm{d}t$$

and

$$Q_n^\xi = \int \Big(\widehat{\xi}_n(t) \;-\; \tilde{\xi}_n(t)\Big)^2 a(t)\,\mathrm{d}t,$$

where a is a continuous weight function vanishing for all $t > \tau_H$. Further set $\kappa_1 = \int K(x)^2\,\mathrm{d}x$ and $\kappa_2 = \int (K*K)^2(x)\,\mathrm{d}x$ To formulate the desired limit theorem let us introduce the following terms for $j = f,\, \lambda,\, \xi$

$$m_n^j = (nb_n)^{-1}\,\kappa_1 \int v_j(t)\,a(t)\,\mathrm{d}t$$

and

$$\sigma_j^2 = 2\,\kappa_2 \int v_j^2(t)\,a^2(t)\,\mathrm{d}t.$$

The functions v_j will be given later. Extending methods from [3] for deriving the limit distribution of quadratic functionals of kernel estimators one can prove the following statements:

Theorem 1. *Suppose that*

(i) *K is a continuous density function vanishing outside the interval $[-L, L]$ for some $L > 0$.*

(ii) *f, λ and ξ, respectively, and H are Lipschitz continuous.*

(iii) *The function a is continuous and $a(t) \equiv 0$ for all $t > T_H$.*

(iv) *$b_n \to 0$ and $nb_n^2 \to \infty$.*

Then for $n \to \infty$

$$nb_n^{1/2}\left(Q_n^j \;-\; m_n^j\right) \;\xrightarrow{\mathcal{D}}\; \mathrm{N}(0, \sigma_j^2) \qquad j = f,\, \lambda,\, \xi$$

with

$$\begin{aligned} v_f(t) &= \frac{\mathrm{d}\left(S(t)\int_0^t \frac{\mathrm{d}\Lambda(s)}{1-H(s)}\right)}{\mathrm{d}t} \\ v_\lambda(t) &= \frac{\lambda(t)}{1-H(t)} \\ v_\xi(t) &= \xi(t) = (1-H(t_-))\lambda(t) \end{aligned}$$

Remark: In the case that there are no censored observations we have $v_f = v_\xi$.

3 Formulation of the Test Procedures

Suppose that the underlying model is regular. Then the maximum likelihood estimator for the parameter θ is $\sqrt{n}$-consistent. Thus, the limit statements formulated above remain true if the parameter θ is substituted by its maximum likelihood estimator $\hat{\theta}$ (constructed in the hypothetical model). Moreover, the unknown quantities in the standardizing terms can be replaced by estimators:

$$\begin{aligned} \widehat{v_{f}}_n(t) &= \frac{\mathrm{d}\left(S(t;\hat{\theta})\int_0^t \frac{\mathrm{d}\Lambda(s;\hat{\theta})}{1-\widehat{H}_n(s_-)}\right)}{\mathrm{d}t} \\ \widehat{v_{\lambda}}_n(t) &= \frac{\lambda(t;\hat{\theta})}{1-\widehat{H}_n(t_-)} \\ \widehat{v_{\xi}}_n(t) &= (1-\widehat{H}_n(t_-))\lambda(t;\hat{\theta}) \end{aligned}$$

So finally we obtain as test procedures the following three rules for $j = f, \lambda, \xi$: The hypothesis $\mathcal{H}$ is rejected, iff

$$\widehat{Q}_n^j \geq \frac{z_\alpha \widehat{\sigma}_{jn}}{n b_n^{1/2}} + \widehat{m}_n^j$$

where

$$\widehat{m}_n^j = (nb_n)^{-1}\,\kappa_1 \int \widehat{v}_{j_n}(t)\,a(t)\,\mathrm{d}t, \qquad \widehat{\sigma}_{jn}^2 = 2\,\kappa_2 \int \widehat{v}_{j_n}^{\,2}(t)\,a^2(t)\,\mathrm{d}t$$

and z_α is the $(1-\alpha)$-quantile of the standard normal distribution.

Remark: The behavior of the power of these three test procedures are considered in the talk.

4 Resampling Methods

For a good approximation of the distribution of the test statistics by the limit distribution the sample size and the behavior of the smoothing parameter is essential. Roughly speaking, two limiting processes are important: Pointwise the estimators are asymptotically normally distributed, and the integral over the squared differences behaves like a sum of chi-squared variables. To ensure their asymptotical independence the bandwidth has to converge to zero fast enough. To overcome insufficient accuracy of approximation resampling methods can be applied. For data without censoring one can apply parametric bootstrap: One estimates the unknown parameter by the original data and generates samples of from the hypothetical distribution with the estimated parameter. Note, this parameter is finite-dimensional. Then the computation of the test statistics for these samples provide empirical quantiles or an empirical p-value. The situation becomes more complicated in the case of censoring. If the censoring distribution G is known then there is no additional problem to generate the hypothetical situation. The same is true, if G is known up to a finite-dimensional parameter. However, in general G is an unknown infinite-dimensional parameter. Let us discuss two proposals for this case: For simplicity of presentation we consider now the simple hypothesis, that is we check

$$\mathcal{H}: \quad F = F_0 \quad \text{versus} \quad \mathcal{K}: \quad F \neq F_0.$$

Both proposals are based on estimators for the unknown distribution G. The first estimate includes the hypothesis $\mathcal{H}$. Estimate G by

$$1 - G_1^*(t) \;=\; \frac{1 - \widehat{H}_n(t)}{1 - F_0(t)}.$$

In the second we do not take into account $\mathcal{H}$, and we estimate

$$1 - G_2^*(t) \;=\; \frac{1 - \widehat{H}_n(t)}{1 - \widehat{F}_n(t)}.$$

In other words, here the estimator for G is the Kaplan–Meier estimator. The resampling procedures are now defined as follows:

1. For $r = 1, \ldots, R$ generate observations

$$\begin{aligned} &Y_{ri}^* \;\text{ according to }\; F_0 \\ &C_{ri}^* \;\text{ according to }\; G_1^* \;\text{ (or }\; G_2^*) \\ &T_{ri}^* = \min(Y_{ri}^*, C_{ri}^*), \qquad \Delta_{ri}^* = \mathbf{1}(Y_{ri}^* \leq C_{ri}^*) \end{aligned}$$

2. Compute for each $r = 1, \ldots, R$ the statistics

$$Q_{rn}^f, \quad Q_{rn}^\lambda, \quad Q_{rn}^\xi$$

3. The empirical p-values are given by

$$p^j = \frac{\#\{Q^j_{rn} \geq Q^j_n\}}{R} \qquad j = f,\, \lambda,\, \xi$$

where Q^j_n is the test statistic based on the original data.

References

[1] Bagdonavicius V., Krupois J., Nikulin M. (2010). *Nonparametric Tests for Censored Data*. Wiley-VCH

[2] Habib P., Thomas D.R. (1986). Chi-Square Goodness-of-fit tests for randomly censored data. *Ann. Statist.*. Vol. **14**, pp. 759–765.

[3] Hall P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Analysis*. Vol. **14**, pp. 1–16.

[4] Hjort N.L. (1990). Goodness of fit testss in models for life history data based on cumulative hazard rates. *Ann. Statist.*. Vol. **18**, pp. 1221–1258.

[5] Kim J.H. (1993). Chi-Square goodness-of-fit tests for randomly censored data. *Ann. Statist.*. Vol. **21**, pp. 1621–1639.

[6] Liero, H. (2003). Testing the hazard rate. *Preprint of the Mathematical Institute of the University of Potsdam*. 2003/ 17.

Bayesian Model Specification: Some Problems Related to Model Choice and Calibration

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

In the development of Bayesian model specification for inference and prediction we focus on the conditional distributions $p(\theta|\mathcal{B})$ and $p(D|\theta,\mathcal{B})$, with data D and background assumptions $\mathcal{B}$, and consider *calibration* (an assessment of how often we get the right answers) as an important integral step of the model development. We compare several predictive model-choice criteria and present related calibration results. In particular, we have implemented a simulation study to compare predictive model-choice criteria LS_{CV}, a *log-score* based on cross-validation, LS_{FS}, a full-sample log score, with deviance information criterion, DIC. We show that for several classes of models DIC and LS_{CV} are (strongly) negatively correlated; that LS_{FS} has better small-sample model discrimination performance than either DIC, or LS_{CV}; we further demonstrate that when validating the model-choice results, a standard use of *posterior predictive tail-area* for hypothesis testing can be poorly calibrated and present a method for its proper calibration.

Keywords: log-score, deviance information criterion, posterior predictive tail areas, hypothesis testing.

Introduction

Bayesian approach to modeling comprises *inference, prediction* and *decision-making* and considers three main objects: θ, a model parameter vector; D, an information (data) source about θ; and $\mathcal{B}$, a set of propositions summarizing background assumptions about of θ and D, for example, that $\theta > 0$ if θ represents the mean remission time for a specified set of patients with a given disease; or that the data set arose as the result of a randomized controlled trial with the specified design). From the results of Cox (1946) and Ramsey (1926) each of these three basic Bayesian statistical activities is governed concptually by a single equation and requires a series of specification tasks:

- (inference) $p(\theta|D,\mathcal{B}) = c\,p(\theta|\mathcal{B})\,p(D|\theta,\mathcal{B})$, where $c > 0$, and $p(\theta|D,\mathcal{B})$ *posterior distribution*, quantifies the information about θ, both internal and external to D;
- (prediction) $p(D^*|D,\mathcal{B}) = \int_\Theta p(D^*|\theta,\mathcal{B})\,p(\theta|D,\mathcal{B})\,d\theta$, where D^* is future data;
- (decision) The optimal action is given by $a^* = \text{argmax}_{a\in\mathcal{A}}\,E_{(\theta|D,\mathcal{B})}U(a,\theta)$.

In problems of realistic complexity it is uncertain how to specify $p(D|\theta, \mathcal{B})$. In our view, a leading principle governing this specification should be *calibration*, which consists of checking how often one obtains the right answers. For example, a statement such as "$p(a < \theta < b|D, \mathcal{B}) = 0.9$" should be verifiably correct about 90% of the time. To address the uncertainty in specifying a model, $p(D|\theta, \mathcal{B})$, we search for an ensemble, $\mathcal{M}$, of such specifications in a well calibrated manner, carefully avoiding a double use of data (to specify priors on model space and again to update this prior when carrying out inference and prediction). In the paper we present calibration results related to the following basic questions in Bayesian model specification, "Q_1: Is model M_j better than M_k?" and "Q_2: Is model M_j good enough?" These questions are not complete without a clear reference to the purpose of the models. However, once the purpose is made explicit, the inferential task transforms into a decision problem, best solved by maximizing expectation of utility (MEU) specific for the model's purpose.

A standard way to answer question Q_1 is to use *Bayes factors* and related criteria, for example (for reasons of space we comment on this very briefly). A well known problem with Bayes factors is a possibly extreme sensitivity to the way diffuse priors are specified on the model parameters, see e.g. Bernardo & Smith (1994). The consequence of this instability is that the evidence in favor of one model over the other may be made arbitrarily large, based on a range of plausible parameter values, even regardless of the data set. Motivated in part by this well known problem, we focus on stable model-choice criteria based on the *posterior predictive* distribution (of the future data, D^*, given the observed sample, D), which has a sound basis as a utility for model comparison and is entirely stable relative to the specification of diffuse priors: $p(D^*|D, M_j, \mathcal{B}) = E_{(\eta_j|D,M_j,\mathcal{B})} p(D^*|\eta_j, M_j, \mathcal{B})$.

We argue that the quality of model prediction is also a solid basis for a useful generic utility in model comparison and hence we focus on working with posterior predictive distributions. In order to compare a predictive distribution with the actual data point, y^*, we use two *log-score* criteria, LS_{CV}, based on cross validation and defined as $n\, LS_{CV}(M_j|y, \mathcal{B}) = \sum_{i=1}^{n} \log p(y_i|y_{-i}, M_j, \mathcal{B})$, and LS_{FS}, the *full-sample log-score* defined as $n\, LS_{FS}(M_j|y, \mathcal{B}) = \sum_{i=1}^{n} \log p(y_i|y, M_j, \mathcal{B})$, and which uses all data in the sample only once (see, for example, Gelfand & Dey (1994) and Laud & Ibrahim (1995). Considering how to address the question Q_1, we contrast deviance information criterion, DIC with the *log-score* rules, LS_{CV} and LS_FS.

The plan of the paper is as follows: In Sections 1 and 2 we present aspects of some answers to question Q_1 whereas Section 3 addresses a calibration issue arising from Q_2. Specifically, in Section 1, we consider how to obtain answers to Q_1, explore similarities and differences between DIC and LS_{CV} in Gaussian and Poisson models, and show results on the small-sample performance of DIC, LS_{CV} and LS_{FS} in discriminating between nested models. In Section 3 we show that the *posterior predictive tail areas* (Gelman et al. (1996)), a standard method for answering "could model M_j have generated the data?" (a question related to Q_2) can be poorly calibrated, and we document an approach to calibrating the answer.

1 LS_{CV} and DIC

In order to show the relationship between LS_{CV} and DIC as model-comparison criteria let us consider a simple parametric model, M_0, for continuous outcomes, where: $(y_i|\mu, \mathcal{B}) \overset{\text{IID}}{\sim} N(\mu, s^2)$ and $(\mu|\mathcal{B}) \sim N(a, b^2)$.

With (s^2, a, b^2) known and a diffuse prior on μ (large b^2), the posterior for μ is: $p(\mu|y, \mathcal{B}) \doteq N(\bar{y}, s^2/n)$, where $\bar{y}$ is the sample mean of $y = (y_1, \ldots, y_n)$. The predictive distribution for the next observation is then $p(y_{n+1}|y, \mathcal{B}) \doteq N(\bar{y}, s^2/(1 + 1/n))$. Similarly, $p(y_i|y_{-i}, \mathcal{B}) \doteq N(\bar{y}_{-i}, s_n^2)$, where $\bar{y}_{-i}$ is the sample mean with observation i omitted and $s_n^2 = s^2\,(1 + 1/(n-1))$, so that $LS_{CV}(M_0|y, \mathcal{B}) \doteq c_1 - c_2 \sum_{i=1}^n (y_i - \bar{y}_{-i})^2$ for some constants c_1 and c_2 with $c_2 > 0$.

With a bit of algebra it can be shown that $LS_{CV}(M_0|y, \mathcal{B}) \doteq c_1 - c_2 \sum_{i=1}^n (y_i - \bar{y})^2$, $(c_2 > 0)$, meaning that for M_0 with a diffuse prior the LS_{CV} is almost perfectly *negatively correlated* with the sample variance.

In model M_0 the *deviance* is $D(\mu) = -2 \ln l(\mu|y, \mathcal{B}) = c_0 + c_3 \sum_{i=1}^n (y_i - \mu)^2$ for some $c_3 > 0$. Given a parametric model $p(y|\theta)$, Spiegelhalter et al. (2002) define the *deviance information criterion* (DIC) as: $DIC(M|y, \mathcal{B}) = D(\bar{\theta}) + 2\,\hat{p}_D$, where p_D is the effective number of model parameters, and $\bar{\theta}$ is the posterior mean of θ, so that models with low DIC values are to be preferred over those with higher values. When p_D is difficult to read directly from the model (e.g., in complex hierarchical settings with random effects), it can be estimated from standard MCMC output as $\hat{p}_D = \overline{D(\theta)} - D(\bar{\theta})$, where $\overline{D(\theta)}$ is the posterior mean of the deviance and $D(\bar{\theta})$ is the deviance evaluated at the posterior mean of θ.

Model M_0 has just one parameter ($p_D = 1$), a diffuse prior for which implies $\bar{\theta} \doteq \bar{y}$, so that we get $DIC(M_0|y, \mathcal{B}) \doteq c_0 + c_3 \sum_{j=1}^n (y_j - \bar{y})^2 + 2$ concluding that

$$-DIC(M_0|y, \mathcal{B}) \doteq c_1 + c_2 LS_{CV}(M_0|y, \mathcal{B}) \tag{1}$$

for $c_2 > 0$. In other words, in this simple setting, choosing a model by *maximizing* LS_{CV} and by *minimizing* DIC are approximately equivalent behaviors. This argument readily generalizes to any situation in which the predictive distribution is approximately Gaussian.

As a second example of the relationship between LS_{CV} and DIC we consider two models for count data a fixed-effects Poisson (FEP), model M_1 where $(y_i|\lambda, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Poisson}(\lambda)$ and $(\lambda|\mathcal{B}) \sim p(\lambda|\mathcal{B})$, and random-effects Poisson (REP), model M_2:

$$\begin{array}{ccc} (y_i|\lambda_i, \mathcal{B}) & \overset{\text{indep}}{\sim} & \text{Poisson}(\lambda_i) \\ \log(\lambda_i) & = & \beta_0 + e_i \\ (\beta_0, \sigma^2|\mathcal{B}) & \sim & p(\beta_0, \sigma^2|\mathcal{B}) \\ (e_i|\sigma^2, \mathcal{B}) & \overset{\text{IID}}{\sim} & N(0, \sigma^2). \end{array} \tag{2}$$

where $i = 1, ..., n$. M_1 is of course a special case of M_2 with $\left(\sigma^2 = 0, \lambda = e^{\beta_0}\right)$; the likelihood in M_2 is a Lognormal mixture of Poissons.

We conducted a partial-factorial simulation study with factors $\{n = 18, 32, 42, 56, 100\}$, $\{\beta_0 = 0.0, 1.0, 2.0\}$, and $\{\sigma^2 = 0.0, 0.5, 1.0, 1.5, 2.0\}$ in which {(data-generating mechanism,

assumed model)} = $\{(M_1, M_1), (M_1, M_2), (M_2, M_1), (M_2, M_2)\}$; in each cell of this grid we used 100 simulation replications. Here we summarize only a small part of the results of this simulation (see Krnjajić (2005) for additional details).

When both the data-generating model and the assumed model were M_1 (the fixed-effects Poisson), LS_{CV} and DIC are almost perfectly negatively correlated (graph not shown); By contrast, the Figure 1 shows that when the data-generating and assumed models were M_2 (the random-effects Poisson), LS_{CV} and DIC are less strongly negatively correlated, although the correlation increases with n (graph not shown).

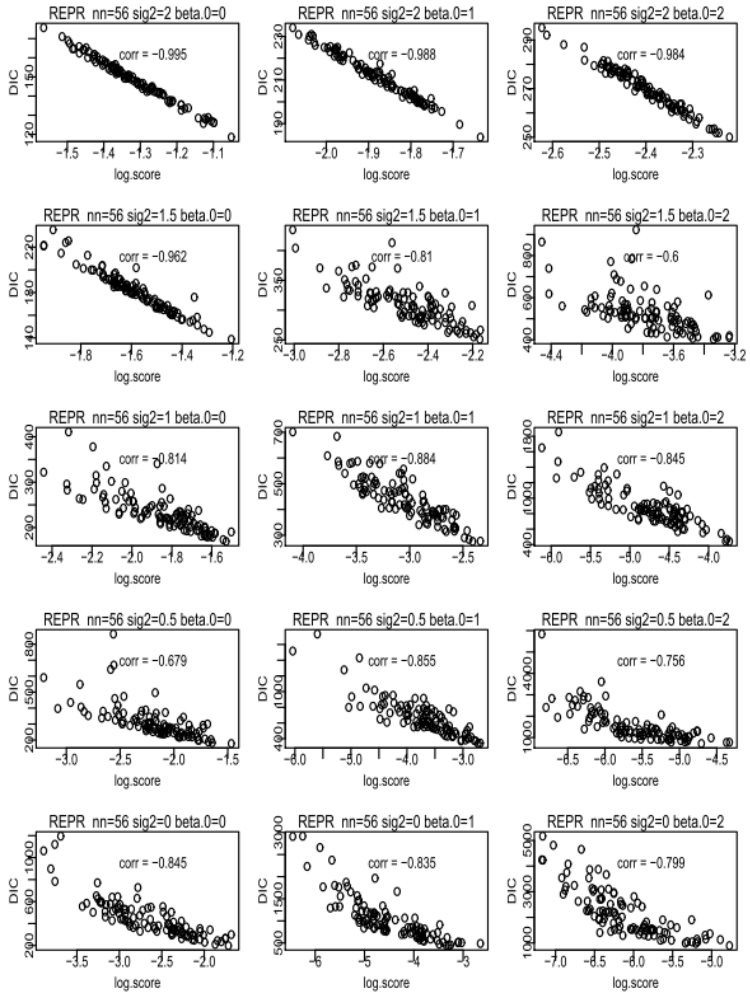


Figure 1: *DIC versus LS_{CV} with $n = 56$; the data-generating and assumed models were both M_2 (random-effects Poisson).*

Table 1: Percentages of correct model choice and mean absolute difference in LS_{CV} between M_1 and M_2 when the right model is M_2, for $n = 32$.

% Correct Decision			Mean Absolute Difference in LS_{CV}		
	β_0			β_0	
σ^2	0	1	σ^2	0	1
0.10	31	47	0.10	0.001	0.002
0.25	49	85	0.25	0.002	0.013
0.50	76	95	0.50	0.017	0.221
1.00	97	100	1.00	0.237	4.07
1.50	98	100	1.50	1.44	17.4
2.00	100	100	2.00	12.8	63.9

2 Model-comparison criteria and small data samples

In addition to LS_{CV}, which requires n model fitting exercises, our interest was drawn to another version of the log-score idea in which no cross-validation is employed. Instead, in the one-sample situation, for instance, it suffices to compute only a single predictive distribution $p(\cdot\,|y, M_j)$ for future data, for each model M_j under consideration and based on the entire data set y. Thus, we define the *full-sample log score* $n\,LS_{FS}(M_j|y, \mathcal{B}) = \sum_{i=1}^{n} \log p(y_i|y, M_j, \mathcal{B})$ (cf. Laud & Ibrahim (1995)). Remark. This appears to use the data twice, but (a) all LS_{FS} is actually doing is evaluating the posterior predictive distribution for the *next* data value at the observed data, and (b) when n is even moderate in size, any effect this may induce is small. The calculation of LS_{FS}, as opposed to Bayes factors, is entirely stable and does not have any difficulties related to the way diffuse priors may be specified.

Here we examine three model-choice rules: {maximize LS_{CV}, maximize LS_{FS}, minimize DIC}. and consider two models M_1 and M_2 to choose between. Our objective is to find out how accurately do these rules discriminate between M_1 and M_2?

As an extension of the previous simulation study, we generated data from the random-effects Poisson model M_2 (equation (2)) and computed LS_{CV}, LS_{FS}, and DIC for models M_1 (the fixed-effects Poisson, FEP) and M_2 (the random-effects Poisson, REP) in the full-factorial grid $\{n = 32, 42, 56, 100\}$, $\{\beta_0 = 0.0, 1.0\}$, $\{\sigma^2 = 0.1, 0.25, 0.5, 1.0, 1.5, 2.0\}$, with 1000 simulation replications in each cell, and we monitored the percentages of correct choice for each model specification method (in this simulation M_2 is always correct).

Table 1 gives examples of the results of this simulation, using LS_{CV} for illustration. Even with a sample size of only 32, LS_{CV} makes the right model choice more than 90% of the time when $\sigma^2 > 0.5$ for $\beta_0 = 1$ and when $\sigma^2 > 1.0$ for $\beta_0 = 0$ (these are parameter ranges that lead to large enough amounts of extra-Poisson variability that random-effects models would be contemplated). The right part of the table shows that even rather small differences in LS_{CV}

can separate correct and incorrect model choice, which begs the question "When a difference on the log score scale is big enough?" (we return to this point in Section 3). Based on model discrimination results for LS_{CV}, LS_{FS}, and DIC we created a series of performance graphs (not shown) and observed (expectedly) that increasing σ^2 makes it easier for all three methods to conclude that random effects model (the nesting model) is needed to accommodate the Poisson over-dispersion. Interestingly, in this simulation environment LS_{FS} was more accurate, with small samples of data, at identifying the correct model than LS_{CV} or DIC; for this reason, we focus on LS_{FS} in what follows.

3 Calibration of posterior predictive tail areas

Section 2 shows that full-sample log scores can stably and reliably decide between two models by choosing one with higher LS_{FS} (or LS_{CV}) value. However, this still leaves open model specification question Q_2: Is M_1 good enough?

In our view, a full judgment of adequacy requires real-world input about the purpose of the model, so it does not seem possible to propose generic methodology to answer Q_2. Instead, the somewhat related question "Q_2': Could model M_j have generated the data?" can be answered in a general way by simulating from M_j many times, developing a distribution of (e.g.) LS_{FS} values, and seeing how unusual the actual data set's log score is in this distribution.

This is related to the *posterior predictive model-checking* method of Gelman et al. (1996). However, this kind of simulation needs to be done carefully (see Draper (1996)), or the result will be poor calibration; indeed, Bayarri & Berger (2000) and Robins et al. (2000) have demonstrated that the procedure in Gelman et al. (1996) may be (sharply) conservative. Using a modification of an idea suggested by Robins et al., we have developed a method for accurately calibrating the log score scale.

The inputs to our procedure are: a data set and a model (which may be parametric or non-parametric). For simplicity, consider a one-sample data set, D, of counts and suppose the goal is to quantify whether this data set could have come from the model $(y_i|\lambda, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Poisson}(\lambda)$, and $(\lambda|\mathcal{B}) \sim$ diffuse (call it model $(*)$). Now, consider the following procedure:

Step 1: Calculate LS_{FS} for data set D and call it the *actual log score* (ALS). Obtain the posterior for λ given y based on data set D; call this the *actual posterior*. **Step 2:**

```
for ( i in 1:m1 ) {
  Let lambda[ i ] be a draw from the actual posterior.
  Sample n data points from model (*) above, using lambda = lambda[ i ].
  Compute the full-sample log-score, LS.FS[ i ], for this data set.
}
```

The output of this loop is a vector of log scores; call this *V.LS*. Locate the ALS in the distribution of LS_{FS} values by computing the percentage of LS_{FS} values in V.LS that are no greater than ALS; call this percentage the *unadjusted actual tail area* (suppose, e.g., that this comes out 0.22).

So far this is similar to Gelman et al. with LS_{FS} as the *discrepancy function*. We know from our own simulations (summarized below) and the literature such as Bayarri & Berger (2000), Robins et al. (2000) that this tail area (a P-value for a composite null hypothesis, e.g., Poisson(λ) with λ unspecified) is conservative, i.e., with the 0.22 example above an adjusted version of it that is well calibrated would be smaller (and might be much smaller, e.g., 0.02). We have modified and implemented one of the ways suggested by Robins et al. for improving calibration, and we have shown that it does indeed work even in rather small-sample situations, although implementing the basic idea can be computationally intensive. **Step 3:**

```
for ( j in 1:m2 ){
  Let lambda* be a draw from the actual posterior.
  Generate a data set of size n from the model (*) above,
    using lambda = lambda*; call this the simulated data set.
  Repeat Steps 1 and 2 above on this simulated data set.
}
```

The result will be a vector of unadjusted tail areas; call this *V.P.* Compute the percentage of tail areas in V.P that are no greater than the unadjusted actual tail area; this is the *adjusted actual tail area.*

The claim is that the 3–step procedure above is well-calibrated: if the sampling part of model $(*)$ really did generate the observed data, the distribution of adjusted actual tail areas would be approximately uniform, since $X \sim F_X$ implies $F_X(X) \sim U(0,1)$. This claim of calibration can be verified by building a further loop around steps 1–3 as follows:

```
Choose a lambda value of interest; call it lambda.sim.
for ( k in 1:m3 ) {
  Generate a data set of size n from the model (*) above,
     using lambda = lambda.sim; call this the validation data set.
  Repeat Steps 1-3 on the validation data set.
}
```

The result here is a vector, *V.TA*, of *adjusted tail areas.* We have verified (via simulation, performed on a cluster of 100 Linux-based CPUs) in several settings that the distribution of values in V.TA is (very) close to $U(0,1)$ indeed.

Figure 2 summarizes a set of histograms of the uncalibrated actual tail areas from one-sample Poisson model, indicating that in many cases the tail areas (p-values) are far from the target (uniform) distribution.

Consider, for example, the case ($n = 100, \lambda = 0.14$) in the fourth row and first column of the Figure 2: if the uncalibrated tail area came out 0.35 in this situation, it would be natural to conclude that the data could very well have come from the Poisson model, but this part of Figure 2 demonstrates clearly that in fact an uncalibrated tail area of 0.35 with ($n = 100, \lambda = 0.14$) is highly unusual under the Poisson model. Our procedure solves the calibration problem by asking "How often would one get 0.35 or less for an uncalibrated tail area in this situation?", and it is

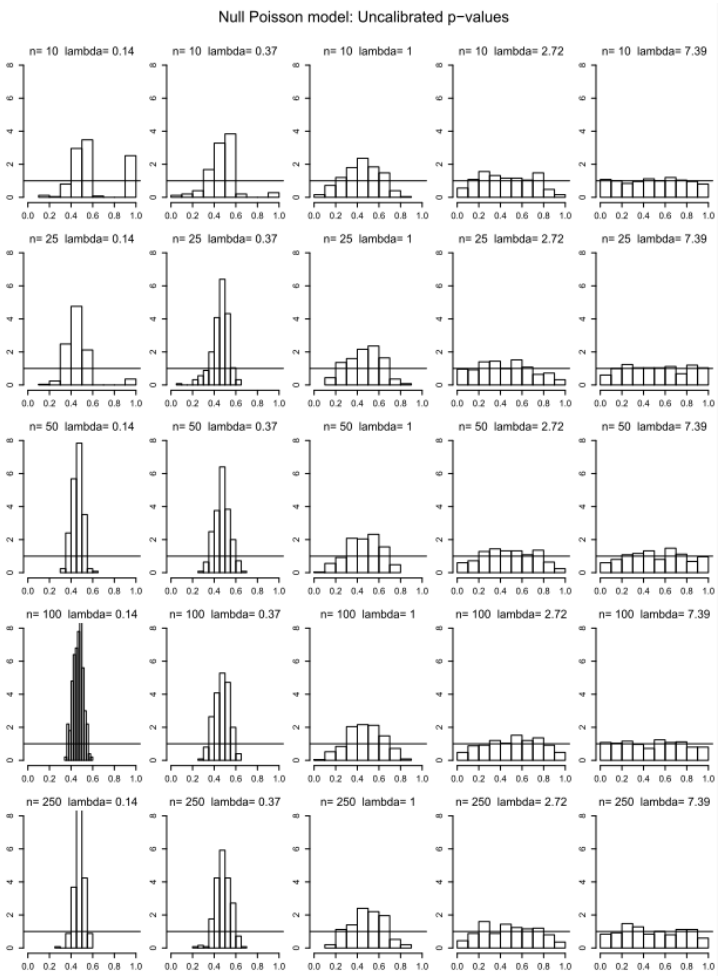


Figure 2: *Poisson model: uncalibrated tail-area values.*

evident from Figure 2 that the answer is not very often (in fact, only about 0.035 of the time, i.e., in this case the calibrated version of the uncalibrated Gelman et al. tail area is 10 times smaller).

Figure 2 shows also that the calibration of the unadjusted approach improves in the one-sample Poisson setting, for increasing λ (even for small n), but in case of the Gaussian model with both μ and σ^2 unknown, the unadjusted approach remains poorly calibrated across the entire subset $\{-1 \leq \mu \leq +1\} \times \{0.1 \leq \sigma^2 \leq 10\}$ of parameter space we examined, and things actually seem to get worse as n increases (not shown). However, the adjusted results, for the Gaussian model, are nearly perfectly calibrated, having distributions close to $U(0, 1)$ for all examind parameter values and sample sizes, (again, not shown). Note that for the reason of limited space here we could show only a small fraction of results and graphs.

Conclusions

We have argued that *calibration* (checking how often one obtains the right answer) is an important principle that that arises naturally in good Bayesian modeling; the question "Q_1: Is model M_j better than M_k?" is central to the process of well-calibrated Bayesian model specification; and it is not well formed unless the purpose of the model is considered. Once the the purpose of the model is explicitly stated, the task of Bayesian model specification turns into a decision problem of maximizing expected utility (MEU), with a purpose-specific utility function (which may be computationally intensive).

LS_{FS} appears as a useful improvement upon DIC, with three advantages: LS_{FS} may well have better small-sample model discrimination behavior (as in the simulation of Section 3.1); LS_{FS} is insensitive to model parameterization; and LS_{FS} can be used both in Bayesian nonparametric and parametric settings; To decide when to stop looking for a better model, the question "Q_2': Could model M_j have generated the data?" can be answered in a well-calibrated manner, using LS_{FS} as a model choice criterion, as shown in the last section.

References

[1] Berger, M. J. B. . J. (2000). P-values for composite null models. *JASA*, 95: 1127–1170.

[2] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

[3] Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14: 1–13.

[4] Draper, D. (1996). Utility, sensitivity analysis, and cross-validation in Bayesian model checking. Discussion of Posterior predictive assessment of model fitness via realized discrepancies, by A. Gelman, X.-L. Meng, and H. Stern. *Statistica Sinica*, 6: 760–767.

[5] Gelfand, A. E. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *JRSS (Series B)*, 56: 501–514.

[6] Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6: 733–807.

[7] Krnjajic, M. (2005). Contributions to Bayesian statistical analysis: model specification and nonparametric inference. Department of Applied Mathematics and Statistics, UCSC.

[8] Laud, P. and Ibrahim, J. (1995). Predictive model selection. *JRSS/B*, 57: 247–262.

[9] Ramsey, F. P. (1926). Truth and probability. In Braithwaite, R. B. (ed.), *The Foundations of Mathematics and Other Logical Essays*, 156–198. London: Kegan Paul.

[10] Robins, J. M., van der Vaart, A., and Ventura, V. (2000). Asymptotic distribution of P-values in composite null models. *JASA*, 95: 1143–1156.

[11] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *JRSS/B*, 64: 583–639.

Statistical Ananlysis of Markov Chains of Conditional Order

M.V. Maltsew, Yu.S. Kharin
Research Institute for Applied Problems of Mathematics and Informatics
Belarusian State University
Minsk, Belarus
e-mail: maltsew@mail.ru, kharin@bsu.by

Abstract

A new mathematical model — Markov chain of conditional order — is proposed for statistical analysis of discrete time series with "long memory". Statistical estimators for parameters of this model are constructed and their properties are analyzed. A statistical test for the values of parameters is proposed. Results of computer experiments are presented.

Keywords: Markov Chain, Markov Chain of Conditional Order, Maximum Likelihood, Hypothesis Testing.

Introduction

Markov chain [3] is a well known model to analyze data sequences in many applications. It is used to solve problems of statistical data analysis in signal processing [7], information protection [4], genetics [9], and other fields. Fitting of an adequate mathematical model is an important step in data analysis.

The Markov chain of the order $s, s \geq 1$, is a widely used model in analysis of discrete time series. Unfortunately, the number of parameters of this universal model increases exponentially when the order s grows. One needs to have the data set of huge size to estimate this model. Thus development and analysis of high-order Markov chain models with a small number of parameters is an actual problem. Give some examples of such models: Markov chain of the order s with r partial connections [5], Raftery model [8], the variable length Markov chain [1], and the Markov chain of conditional order that is proposed and analyzed in this paper.

1 Mathematical Model

Introduce the following notation: $\mathbf{N}$ is the set of natural numbers, $2 \leq N < \infty$; $A = \{0, 1, \ldots, N-1\}$ is the state space of N elements; $J_m^n = (j_m, \ldots, j_n) \in A^{n-m+1}$, $n \geq m$ is the multiindex; $x_t \in A, t \in \mathbf{N}$, is a homogeneous Markov chain of the order s ($2 \leq s < \infty$) defined at some probability space (Ω, F, P); $P = (p_{J_1^{s+1}})$ is the $(s+1)$-dimensional one-step transition probability matrix, $p_{J_1^{s+1}} = \mathrm{P}\{x_{t+s} = j_{s+1} | x_{t+s-1} = j_s, \ldots, x_t = j_1\}, \forall t \in \mathbf{N}$; $B_* \in \{1, 2, \ldots, s-1\}$, $K = N^{B_*} - 1$ are natural numbers; $Q^{(1)}, \ldots, Q^{(M)}$ are M ($1 \leq M \leq K+1$) different stochastic

matrices of the order N: $Q^{(m)} = (q_{i,j}^{(m)}), 0 \le q_{i,j}^{(m)} \le 1, \sum\limits_{j \in A} q_{i,j}^{(m)} \equiv 1, i, j \in A,\ 1 \le m \le M$; $< J_m^n >= \sum\limits_{k=m}^{n} N^{k-1} j_k \in \{0, 1, \ldots, N^{n-m+1}\},\ 1 \le m \le n \le s$ is the numeric representation of the multiindex J_m^n; $\delta_{J_m^n, I_m^n} = \prod\limits_{k=m}^{n} \delta_{j_k, i_k}$ is the Kronecker symbol for multiindices J_m^n, I_m^n.

Define the Markov chain $x_t \in A$ of conditional order if its one-step transition probabilities have the following form:

$$p_{J_1^{s+1}} = \sum_{k=0}^{K} \delta_{<J_{s-B_*+1}^s>, k} \cdot q_{j_{b_k}, j_{s+1}}^{(m_k)}, \tag{1}$$

where $1 \le m_k \le M$, $1 \le b_k \le s - B_*$, $0 \le k \le K$, $\min\limits_{0 \le k \le K} b_k = 1$. The sequence of elements $J_{s-B_*+1}^s$, that determines conditions in the formula (1), is called the base memory fragment (BMF) of the random sequence. We can see that the state of the model x_t at time t doesn't depend on all previous states, but depends only on $B_* + 1$ states $(j_{b_k}, J_{s-B_*+1}^s)$; the value of BMF $J_{s-B_*+1}^s$ determines not only the state j_{b_k}, but it also determines the transition matrix. The transition matrix $P = (p_{J_1^{s+1}})$ is defined by $D = 2(N^{B_*} + 1) + MN(N-1)$ independent parameters.

Note that if $B_* = s - 1, b_0 = \ldots = b_K = 1$, we have fully connected Markov chain of the order s; similarly if $b_0 = \ldots = b_K = s - B_*$, we have fully connected Markov chain of the order $B_* + 1$. If $M = K + 1$, then all the parameters $\{m_k\}$ are different and each value k of BMF has its own transition matrix $Q^{(k)}$.

2 Estimators and Their Properties

At first, let us give ergodicity conditions for the Markov chain of conditional order

Theorem 1. Markov chain of conditional order is ergodic if and only if there is a natural number $m \in \mathbf{N}, s \le m < \infty$ such that the inequality holds:

$$\min_{J_1^s, J_{1+m}^{s+m} \in A^s} \sum_{J_{s+1}^m \in A^{m-s}} \prod_{i=1}^{m} \sum_{k=0}^{K} \delta_{<J_{i+s-B_*}^{i+s-1}>, k} \cdot q_{j_{b_k+i-1}, j_{i+s}}^{(m_k)} > 0.$$

In the sequel, we will consider ergodic Markov chains; denote its stationary probability distribution by $\pi_{J_1^s} = \mathrm{P}\{x_{t+s-1} = j_s, \ldots, x_t = j_1\}, J_1^s \in A^s, \forall t \in \mathbf{N}$. The stationary probability distribution $\Pi = (\pi_{<J_1^s>})$ can be found by solving the system of linear algebraic equations [6]: $\bar{P}'\Pi = \Pi$, $\sum\limits_{J_1^s \in A^s} \pi_{J_1^s} = 1$; here $\bar{P} = (\bar{p}_{<I_1^s>, <J_1^s>})$ is the one-step transition probability matrix for the s-dimensional Markov chain of the first order $X^{(t)} = (x_t, x_{t+1}, \ldots, x_{t+s-1}), t \in \mathbf{N}$, with extended state space, $I_1^s, J_1^s \in A^s$, $\bar{p}_{<I_1^s>, <J_1^s>} = \delta_{I_2^s, J_1^{s-1}} p_{I_1^s, j_s}$.

Construct the maximum likelihood estimators (MLE) of the matrices $Q^{(1)}, \ldots, Q^{(M)}$ using an observed realization X_1^n of length n. We need the following notation:

$$n^- = n - s,\ 1 \le l \le s,\ 0 \le l_0 \le s - l;\ A^{s+1}(J_1^l) = \{I_1^{s+1} \in A^{s+1} : I_1^l = J_1^l\},$$

$$A^{1+l_0+l}(j_0^{l_0}, J_1^l) = \{I_1^{1+l_0+l} \in A^{1+l_0+l} : i_1 = j_0, I_{2+l_0}^{1+l_0+l} = J_1^l\};$$

$$A_j = \{I_1^{s+1} \in A^{s+1} : i_1 = j_0, I_{l+2}^{s+1} = J_1^{B_*+1}\},$$

$$\nu_{J_1^{s+1}}(n) = \sum_{t=1}^{n^-} \delta_{X_t^{t+s}, J_1^{s+1}},\ \nu_{J_1^l}(n) = \sum_{I_1^{s+1} \in A^{s+1}(J_1^l)} \nu_{I_1^{s+1}}(n),$$

$$\nu_{j_0,J_1^l}^{(l_0)}(n) = \sum_{I_1^{1+l_0+l} \in A^{1+l_0+l}(j_0^{l_0}, J_1^l)} \nu_{I_1^{1+l_0+l}}(n),\ \bar{\nu}_{J_1^{s+1}}(n) = 1/\sqrt{n^-}(\nu_{J_1^{s+1}}(n) - n^- \pi_{J_1^{s+1}}),$$

$$\bar{\nu}_{J_0^{B_*+1}}^l(n) = 1/\sqrt{n^-}(\nu_{J_0^{B_*+1}}^l(n) - n^- \pi_{J_0^{B_*+1}}^l);$$

$$\bar{q}_{J_0^{B_*+1}} = \sqrt{n^-}(\hat{q}_{J_0^{B_*+1}} - q_{J_0^{B_*+1}});\ \pi_{J_0^{B_*+1}}^l = \mathrm{P}\{x_t = j_0, X_{t+l+1}^{t+l+B_*+1} = J_1^{B_*+1}\};$$

E_{N^s} is the identity matrix of the order N^s; $P^* = \lim\limits_{n\to\infty} \bar{P}^n$ is the limit matrix; $Z = (E_{N^s} - \bar{P} + P^*)^{-1} = (z_{I_1^s, J_1^s})$, $I_1^s, J_1^s \in A^s$.

Lemma 1. The n-dimensional probability distribution ($n > s$) for the Markov chain of conditional order has the following form:

$$\mathrm{P}\{x_1 = j_1, \ldots, x_n = j_n\} = \pi_{J_1^s}^0 \prod_{t=s}^{n-1} \sum_{k=0}^{K} \delta_{<J_{t-B_*+1}^t>,k} q_{t-s+b_k, j_{t+1}}^{(m_k)}, j_1, \ldots, j_n \in A,$$

where $\pi_{J_1^s}^0 = \mathrm{P}\{x_1 = j_1, \ldots, x_s = j_s\}$, $J_1^s \in A^s$, is the starting probability distribution of the Markov chain (1).

Corollary. The likelihood function for the Markov chain of conditional order has the following form:

$$l_n(X_1^n, \{Q^{(i)}\}, B_*, \{b_k\}) = \ln \pi_{J_1^s}^0 + \sum_{u,v \in A, w \in A^{B_*}} \sum_{k=0}^{K} \delta_{<w>,k} \nu_{u,wv}^{(l_k)}(n) \ln q_{u,v}^{(m_k)},$$

where $l_k = s - b_k + B_*$.

Theorem 2. If true values of $B_*, \{b_k\}$, and $\{m_k = k\}$ are known, then the MLE for the one-step transition probabilities $q_{u,v}^{(m_k)}$, $u, v \in A$ are

$$\begin{cases} \hat{q}_{u,v}^{(m_k)} = \sum\limits_{w \in A^{B_*}} \delta_{<w>,k} \dfrac{\nu_{u,wv}^{(l_k)}(n)}{\nu_{u,w}^{(l_k)}(n)},\ if\ \nu_{u,w}^{(l_k)}(n) > 0, \\ \qquad\qquad 1/N,\ if\ \nu_{u,w}^{(l_k)}(n) = 0. \end{cases} \tag{2}$$

Remark. If some parameters $\{m_k\}$, $k = 0, 1, \ldots, K$, are equal, i.e. one transition matrix corresponds to different base memory fragments, then the MLE have the following form:

$$\begin{cases} \hat{q}_{u,v}^{(m_k)} = \dfrac{\sum\limits_{w \in M_{m_k}} \nu_{u,wv}^{(b_k)}(n)}{\sum\limits_{w \in M_{m_k}} \nu_{u,w}^{(b_k)}(n)}, \ if \ \sum\limits_{w \in M_{m_k}} \nu_{u,w}^{(b_k)}(n) > 0, \\ 1/N, \ if \ \sum\limits_{w \in M_{m_k}} \nu_{u,w}^{(b_k)}(n) = 0, \end{cases}$$

where $M_i = \{w \in A^{B_*} : m_{<w>} = i\}$, $i = 1, \ldots, M$, $\bigcup\limits_{i=1}^{M} M_i = A^{B_*}$.

Theorem 3. If the true values of B_*, $\{m_k\}$ are known, then the MLE of $\{b_k\}$ are

$$\hat{b}_k = arg \max_{1 \le b \le s - B_*} \sum_{i,j \in A} \nu_{i,wj}^{s-b-B_*}(n) \ln(\hat{q}_{i,wj}^{(m_k)}), \ k = 1, 2 \ldots, K. \tag{3}$$

Estimators of the order of the Markov chain s and the length of BMF B_* are constructed by minimization of the Bayesian informational functional [2]:

$$(\hat{s}, \hat{B}_*) = arg \min_{2 \le s \le \bar{S}, 1 \le B \le \bar{B}_*} BIC(s, B), \tag{4}$$

$$BIC(s, B) = -\Big(\sum_{u,v \in A, w \in A^B} \sum_{k=0}^{K} \delta_{<w>,k} \nu_{u,wv}^{(s-\hat{b}_k - B)}(n) \ln \hat{q}_{u,v}^{(m_k)} \Big) + 2N^B \log n,$$

where $\bar{S} \ge 2$, $1 \le \bar{B}_* \le \bar{S} - 1$ are maximal admissible values of the parameters s and B_* respectively; estimators $\hat{Q}^{(i)}, i = 1, \ldots, M$ and $\hat{b}_k, k = 0, \ldots, K$, can be found using (2) and (3) respectively.

Now consider asymptotic properties of the constructed estimators (2).

Theorem 4. If the Markov chain of conditional order is stationary, then (2) gives consistent estimators at $n \to \infty$:

$$\hat{q}_{u,v}^{(m)} \xrightarrow{\mathrm{P}} q_{u,v}^{(m)}, \ 1 \le m \le M.$$

Lemma 2. If the Markov chain of conditional order $\{x_t \in A, \ t \in \mathbf{N}\}$, is stationary, then at $n \to \infty$ the random variables $\{\bar{\nu}_{J_0^{B_*+1}}^{l}(n) : J_0^{B_*+1} \in A^{B_*+2}\}$ are jointly asymptotically normal with zero asymptotic means and asymptotic covariances:

$$\operatorname{cov}\{\bar{\nu}_{J_0^{B_*+1}}^{l}(n), \bar{\nu}_{I_0^{B_*+1}}^{l'}(n)\} = \delta_{J_0^{B_*+1}, I_0^{B_*+1}} \pi_{J_0^{B_*+1}}^{l} +$$
$$+ q_{J_0^{B_*+1}} q_{I_0^{B_*+1}} \left(h(J_0^{B_*+1}, I_0^{B_*}) + h(I_0^{B_*+1}, J_0^{B_*}) \right) - 3\pi_{J_0^{B_*+1}}^{l} \pi_{I_0^{B_*+1}}^{l'},$$

where $h(J_0^{B_*+1}, I_0^{B_*}) = \sum\limits_{E_1^{s+1} \in A_j} \sum\limits_{F_1^{s+1} \in A_i} \pi_{E_1^s} z_{E_2^{s+1}, F_1^s}$.

Theorem 5. If the Markov chain of conditional order $\{x_t \in A,\ t \in \mathbf{N}\}$, is stationary, then at $n \to \infty$ the random variables $\{\bar{q}_{J_0^{B_*+1}} : J_0^{B_*+1} \in A^{B_*+2}\}$ are jointly asymptotically normal with zero asymptotic means and asymptotic covariances:

$$\mathrm{cov}\{\bar{q}_{J_0^{B_*+1}}, \bar{q}_{I_0^{B_*+1}}\} = \delta_{J_0^{B_*}, I_0^{B_*}} q_{J_0^{B_*+1}} \frac{\delta_{j_{B_*+1}, i_{B_*+1}} - q_{J_0^{B_*} i_{B_*+1}}}{\pi_{J_0^{B_*}}}.$$

3 Hypothesis Testing

The problem of a deviation detection by the observed time series $\{x_t\}$ from "purely random" sequence appears usually in construction and reliability evaluation of the information security systems [4]. Construct now a statistical test for the hypotheses $H_0 = \{x_t \in A$ is a "purely random" sequence: $q_{i,j}^{(m)} = 1/N,\ \forall i, j \in A,\ m = 1, 2, \ldots, M\}$ and $H_1 = \{x_t \in A$ is the Markov chain of conditional order with one-step transition probabilities $q_{i,j}^{(m)} = q_{i,j}^{(m)}(n) = \frac{1}{N}(1 + \frac{\eta_{i,j}^{(m)}(n)}{\sqrt{n}}) > 0$, $\eta_{i,j}^{(m)}(n) \to \eta_{i,j}^{(m)}$ at $n \to \infty$, $\sum\limits_{j \in A} \eta_{i,j}^{(m)} = 0$, $\sum\limits_{m=1}^{M} \sum\limits_{i,j \in A} |\eta_{i,j}^{(m)}| > 0\}$.

Define the following statistic:

$$\xi_{j_0, J_1^l}^{(l_0)}(n) = \frac{\nu_{j_0, J_1^l}^{(l_0)}(n) - n/N^{l+1}}{\sqrt{n/N^{l+1}}},$$

$$\rho(n) = \sum_{w \in B_*, u, v \in A} \sum_{k=0}^{K} \delta_{<w>,k} (\xi_{u,wv}^{(l_k)})^2 - \frac{1}{N} \sum_{w \in B_*, u \in A} \left(\sum_{v \in A} \sum_{k=0}^{K} \delta_{<w>,k} \xi_{u,wv}^{(l_k)} \right)^2.$$

Theorem 6. If the hypothesis H_0 is true, then at $n \to \infty$ the probability distribution of the statistic $\rho(n)$ converges to the standard χ^2- distribution with $U = N^{B_*+1}(N-1)$ degrees of freedom.

Using this theorem we can construct the statistical test:

$$\text{decide in favour} \begin{cases} H_0 : \rho(n) \le \Delta, \\ H_1 : \rho(n) > \Delta, \end{cases}$$

where $\Delta = G_U^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the standard χ^2- distribution with U degrees of freedom, $\alpha \in (0, 1)$ is the given significance level.

4 Numerical Results

Evaluate performance of the proposed statistical procedures by computer experiments.

Example 1 (s, B_* are known). $A = \{0, 1\}, N = 2,\ s = 4,\ M = 2,\ B_* = 2,\ b_0 = b_1 = 2,\ b_2 = b_3 = 1,\ m_0 = m_2 = 1,\ m_1 = m_3 = 2,$

$$Q^{(1)} = \begin{pmatrix} 0.18 & 0.82 \\ 0.41 & 0.59 \end{pmatrix},\ Q^{(2)} = \begin{pmatrix} 0.77 & 0.23 \\ 0.09 & 0.91 \end{pmatrix}.$$

Numerical experiments were conducted by the following scheme. The Markov chain of conditional order was simulated using the developed computer program. Then estimates $\hat{Q}^1$ and $\hat{Q}^2$ of matrices $Q^{(1)}$ and $Q^{(2)}$ were calculated using the formula (2). After that the estimate of the variance $\hat{v}_n^u = \sum\limits_{k=1}^{2} \sum\limits_{i,j=0}^{1} (\hat{q}_{ij}^{(k)} - q_{ij}^{(k)})^2$ was calculated, where n is the length of the u-th realization of the Markov chain; $n = 500, 750, 1000, \ldots, 10000;\ u = 1, 2, \ldots, U$. Estimates $\{\hat{v}_n^u\}$ were calculated for each n and for $U = 1000$ independent replications. The total mean square error for the estimators (2) $\hat{v}_n = \frac{1}{U} \sum\limits_{u=1}^{U} \hat{v}_n^u$ was obtained; its dependence on n is plotted in Figure 1.

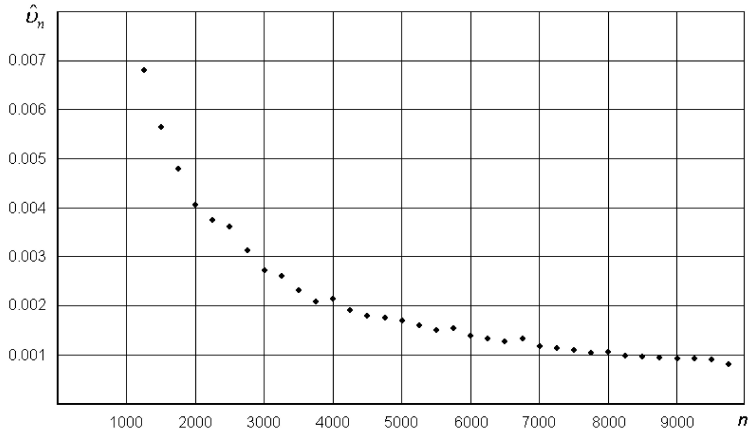


Figure 1: Mean square error

Similarly we evaluate performance for the estimators (3). In order to do this we calculate frequency of the true decision $\varepsilon = \frac{1}{U} \sum\limits_{u=1}^{U} \varepsilon_u$, $\varepsilon_u = \delta_{\hat{b},b}$, $b = (b_0, \ldots, b_K)$. The results are given in Table 1.

Table 1: Performance of the estimators (3)

n	500	1000	1500	2000	2500	3000	3500	4000	≥ 4500
ε	0.322	0.695	0.896	0.936	0.982	0.986	0.992	0.995	1.000

We can see from Table 1, that the probability of the true decision tends to one fast enough when the length of the observed time series grows.

Example 2 (s, B_* are unknown). The Markov chain of conditional order was simulated using the developed computer program. The length of the chain $n = 50000$. Other parameters were as in the Example 1, $\bar{S} = 8$, $\bar{B}_* = 4$.

The values of s, B and the corresponding values of $BIC(s, B)$ are given in Table 2. By (4) we find $\hat{s} = 4, \hat{B}_* = 2$.

Table 2: Performance of the estimators (3)

(s, B)	$BIC(s, B)$	(s, B)	$BIC(s, B)$
(2, 1)	31371	(6, 2)	22852
(3, 1)	30116	(6, 3)	22929
(3, 2)	28535	(6, 4)	23093
(4, 1)	27347	(7, 1)	27347
(4, 2)	22852	(7, 2)	22852
(4, 3)	22934	(7, 3)	22928
(5, 1)	27347	(7, 4)	23092
(5, 2)	22852	(8, 1)	27347
(5, 3)	22929	(8, 2)	22852
(5, 4)	23097	(8, 3)	22928
(6, 1)	27347	(8, 4)	23089

We can see from Table 2 that $BIC(s, B)$ reaches the minimum if $\hat{B}_* = 2$ and $\hat{s} = 4, 5, \ldots, 8$. The true order of the chain is equal to $\hat{s} = 4$; the following pairs (s, B): (4, 2), (5, 2), (6, 2), (7, 2), (8, 2) are equivalent. Then according to (3) we find $\hat{b} = (\hat{b}_0, \ldots, \hat{b}_K)$:

$$\hat{b}_k = arg \max_{1 \le b \le \hat{s} - \hat{B}_*} \sum_{i,j \in A} \nu_{i,wj}^{\hat{s}-b-\hat{B}_*}(n) \ln(\hat{q}_{i,wj}^{(m_k)}),\ k = 1, 2 \ldots, K,$$

$b = (2, 2, 1, 1)$.

Finally, we find the estimators for the transition matrices (each value of BMF has its own matrix):

$$\hat{Q}^{(1)} = \begin{pmatrix} 0.181 & 0.819 \\ 0.400 & 0.600 \end{pmatrix},\ \hat{Q}^{(2)} = \begin{pmatrix} 0.772 & 0.228 \\ 0.088 & 0.912 \end{pmatrix},$$

$$\hat{Q}^{(3)} = \begin{pmatrix} 0.183 & 0.817 \\ 0.402 & 0.598 \end{pmatrix},\ \hat{Q}^{(4)} = \begin{pmatrix} 0.774 & 0.226 \\ 0.089 & 0.911 \end{pmatrix}.$$

5 Conclusion

In this paper we present a new model of the high-order Markov chain with a small number of parameters called the Markov chain of conditional order. Probabilistic and statistical properties

of the model are analyzed. Ergodicity conditions are found, Maximum likelihood estimators of parameters are constructed and their properties are investigated. Statistical test for the deviation detection of the observed time series from the "purely random" sequence is constructed. Numerical experiments illustrate the theoretical results.

References

[1] Buhlmann P., Wyner A. (1999). Variable Length Markov Chains. *The Annals of Statistics*. Vol. **27**, N 2, pp. 480-513.

[2] Csiszar, I. Consistency of the BIC order estimator. (1999). *Electronic research announcements of the American mathematical society*. Vol. **5**, pp. 123-127.

[3] Kemeny J., Snell J. (1990). *Finite Markov Chains*. Princeton NJ: D. Van Nostrand Company.

[4] Kharin Yu. S., Bernik V. I., Matveev G. V., Agievich S. V. (2003). *Mathematical and computer foundations of cryptology*. Novoe znanie, Minsk.

[5] Kharin Yu. S., Piatlitski A. I. (2007). Markov Chain with Partial Connections MC(s, r) and statistical conclusions of its parameters. *Discrete mathematics*, Vol. **19**, N 2, pp. 109-130.

[6] Kharin Yu. S., Zhuk E. E. (2004). *Mathematical and applied statistics*. BSU, Minsk.

[7] Li Y., Dong Y., Zhang H., Zhao H., Shi H., Zhao X. (2010). Spectrum Usage Prediction Based on High-order Markov Model for Cognitive Radio Networks. *2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010)*, pp. 2784-2788.

[8] Raftery A. E. (1985). A Model for High-Order Markov Chains. *J. Royal Statistical Society*, Vol. **B-47**, N 3, pp. 528-539.

[9] Waterman M. S. (1999). *Mathematical Methods for DNA Sequences*. Chapman and Hall/CRC, Boca Raton.

Comparing Predictive Accuracy

Eva Ferreira[1], Winfried Stute[2]
[1] Univ. del Pais Vasco, Departamento de Economia Aplicada III, Bilbao, Spain
[2] Univ. of Giessen, Mathematical Institute, Giessen, Germany
e-mail: [1]eva.ferreira@ehu.es,[2]Winfried.Stute@math.uni-giessen.de

Abstract

In this work we provide new tests for the difference in predictive accuracy of two prognostic factors X_1 and X_2 on a common output Y. Given a set of independent replicates of (X_1, X_2, Y), we split this sample into a learning part for estimating the unknown regression functions, and a validation part for which the residuals need to be computed. We show that the null distributions of our test statistics may be approximated by a normal. In simulations, the power is promising already for small to moderate sample sizes.

Keywords: Predictive accuracy; residuals; nonparametric test, data split.

Introduction

Suppose that Y is an unknown variable of interest. For example, Y could measure the future status of a company. Rather than Y, what might be available already now, is a vector X_1 of covariates which may be helpful to predict the value of Y. In a regression context we decompose Y into a term depending only on X_1 and a noise variable ε_1 orthogonal to X_1, i.e.,

$$Y = m_1(X_1) + \varepsilon_1 \text{ such that } \mathbb{E}(\varepsilon_1|X_1) = 0 \text{ a.s.}$$

The function m_1 is the regression function of Y w.r.t. X_1. When Y is unknown, $m_1(X_1)$ is the best predictor based on X_1, i.e., among all functions $\varphi(X_1)$ of X_1, $m_1(X_1)$ is the one minimizing the expected squared prediction error. Unfortunately, in a practical situation, m_1 is unknown and needs to be estimated from a learning sample $(X_{11}, Y_1), \ldots, (X_{1T}, Y_T)$ of independent replicates of (X_1, Y). Denote with $\hat{m}_1$ any of such estimators, and let

$$e_1 := Y - \hat{m}_1(X_1)$$

be the associated residual. The availability of such data may be limited by the sampling costs for the covariate X_1. Therefore it makes sense to also take into account an alternative covariate X_2. Denote with

$$Y = m_2(X_2) + \varepsilon_2 \text{ such that } \mathbb{E}(\varepsilon_2|X_2) = 0 \text{ a.s.}$$

the corresponding decomposition. While $m_1(X_1)$ and $m_2(X_2)$ are optimal within their classes, it remains open which of $m_1(X_1)$ and $m_2(X_2)$ outperforms the other.

It is the purpose of this paper to provide some methodology for comparing the predictive accuracy of two prognostic factors X_1 and X_2 w.r.t. a common dependent variable Y. The

quality of the fit is measured through $g(\varepsilon_1)$ and $g(\varepsilon_2)$, where g is a weight function chosen by the statistician. We already mentioned the quadratic loss associated with $g(u) = u^2$. In Robust Statistics, a popular weight is $g(u) = |u|$ leading to the mean absolute deviations $|\varepsilon_1|$ and $|\varepsilon_2|$. Another possibility would be

$$g(u) = u^2 1_{\{|u|>\delta\}}.$$

When one applies this g to e_1 and e_2, one neglects deviations which fall below the threshold δ.

Our statistical analysis will be based on a learning (or estimation) sample $(X_{1t}, X_{2t}, Y_t), 1 \leq t \leq T$, of independent replicates of (X_1, X_2, Y). The covariates X_{1t} and X_{2t} can be quite different. They may coincide in some of their coordinates but not in others. Their dimensions d_1 and d_2 may also differ. Later we shall in detail discuss the case when X_1 is a subvector of X_2 so that $d_1 < d_2$. The learning sample will be used to estimate the unknown regression functions m_1 and m_2 through $\hat{m}_1$ and $\hat{m}_2$, say. In a parametric framework, m_1 and m_2 are of the type

$$m_1(x_1) = m_1(x_1, \beta) \text{ and } m_2(x_2) = m_2(x_2, \gamma).$$

In such a situation we have to estimate β and γ by the Least Squares Estimator or robust alternatives $\hat{\beta}$ and $\hat{\gamma}$. For $\hat{m}_1$ and $\hat{m}_2$ we then take the plug-in estimators

$$\hat{m}_1(x_1) = m_1(x_1, \hat{\beta}) \text{ and } \hat{m}_2(x_2) = m_2(x_2, \hat{\gamma}).$$

In a nonparametric framework one may take for $\hat{m}_1$ and $\hat{m}_2$ any nonparametric smoother. See Stone (1977) for general conditions on such smoothers to obtain universal consistency. Spiegelman and Sacks (1980) is a relevant reference for the consistency of the Nadaraya-Watson estimator.

Now, after having obtained $\hat{m}_1$ and $\hat{m}_2$, the associated residuals are computed for a validation sample $(X_{1t}, X_{2t}, Y_t), T+1 \leq t \leq T+n$, being independent of the first:

$$e_{it} = Y_t - \hat{m}_i(X_{it}), T+1 \leq t \leq T+n, i = 1, 2.$$

For a given weight function g, a comparison of the predictive accuracy will now be based on

$$\bar{d} = n^{-1} \sum_{t=T+1}^{T+n} [g(e_{1t}) - g(e_{2t})].$$

Typically, a large value of $\bar{d}$ indicates that X_1 has less predictive accuracy than X_2. A test for

$$H_0 : \mathbb{E}g(\varepsilon_1) = \mathbb{E}g(\varepsilon_2)$$

versus

$$H_1 : \mathbb{E}g(\varepsilon_1) > \mathbb{E}g(\varepsilon_2) \tag{1}$$

rejects H_0 in favor of H_1 when $\bar{d}$ exceeds a critical value. In general, the null distribution is very complicated. A quantity which is much easier to handle is one which is obtained after replacing the residuals by the true errors:

$$\bar{d}_1 = n^{-1} \sum_{t=T+1}^{T+n} [g(\varepsilon_{1t}) - g(\varepsilon_{2t})].$$

Under H_0, this is a sum of centered independent identically distributed summands to which the Central Limit Theorem applies. In our main results we show that

$$n^{1/2}\bar{d} = n^{1/2}\bar{d}_1 + o_{\mathbb{P}}(1) \quad \text{as } n \to \infty \tag{2}$$

under appropriate conditions so that by Slutsky's theorem the distribution of $n^{1/2}\bar{d}$ can in fact be approximated by a normal. Before we come to the main results, some further comments are in order.

Remark 1. *Suppose that all components of X_1 are also included in X_2. For $g(u) = u^2$ we then have*

$$\begin{aligned}
&\mathbb{E}\varepsilon_1^2 - \mathbb{E}\varepsilon_2^2 = \mathbb{E}[Y - m_1(X_1)]^2 - \mathbb{E}[Y - m_2(X_2)]^2 \\
&= \mathbb{E}m_1^2(X_1) - \mathbb{E}m_2^2(X_2) - 2\mathbb{E}[Y m_1(X_1)] + 2\mathbb{E}[Y m_2(X_2)] \\
&= \mathbb{E}m_2^2(X_2) - \mathbb{E}m_1^2(X_1) = \mathbb{E}[m_2(X_2) - m_1(X_1)]^2 = c \geq 0,
\end{aligned}$$

where the second but last equality follows from the facts that $m_1(X_1)$ and $m_2(X_2)$ are conditional expectations of Y w.r.t. X_1 and X_2. The last equality utilizes that X_1 is a subvector of X_2. In most situations c will be strictly positive so that (1) holds true and no extra test is necessary. In such a situation one may be interested to know whether the inclusion of more covariables would increase the predictive accuracy by an amount of at least c_0. In other words we want to test

$$H_0 : \mathbb{E}g(\varepsilon_1) = \mathbb{E}g(\varepsilon_2) + c_0$$

versus

$$H_1 : \mathbb{E}g(\varepsilon_1) > \mathbb{E}g(\varepsilon_2) + c_0.$$

Our approach also applies here. Just replace $\bar{d}$ by $\bar{d} - c_0$.

Of course there may be situations where the augmented X_1 will not improve the predictive accuracy at all. Consider, e.g., the two linear models

$$Y = X_1'\beta + \varepsilon_1 \quad \text{and} \quad Y = X_1'\gamma + U_2'\delta + \varepsilon_2. \tag{3}$$

Hence $X_2 = (X_1', U_2')'$. In this case two situations are possible. If $\delta = 0$, then $\varepsilon_1 = \varepsilon_2$ whence $c = 0$. In other words, the augmentation of X_1 has no effect on the predictive accuracy. In general, we have

$$c = \mathbb{E}[X_1'(\gamma - \beta) + U_2'\delta]^2.$$

If, e.g., X_1 and U_2 are centered and uncorrelated, then

$$c = \mathbb{E}[X_1'(\gamma - \beta)]^2 + \mathbb{E}[U_2'\delta]^2$$

so that the overall difference may be attributed to the variability contained in U_2 weighted by δ and a (possibly reduced) variability contained in X_1.

Remark 2. *Since the test statistic $\bar{d}$ depends on estimated residuals, a crucial role in our setup will be played by $\hat{m}_1$ and $\hat{m}_2$. For the replacement of the e's by the ε's the following error bounds on $\hat{m}_i - m_i$ will be needed:*

$$\mathbb{E}\left\{[\hat{m}_i(X_{i,T+1}) - m_i(X_{i,T+1})]^2\right\} = O(T^{-\beta_i}) \tag{4}$$

$i = 1, 2$. Generally the constant β_i depends on whether we are in a parametric or nonparametric framework. In the parametric case we typically can estimate unknown parameters at the rate $T^{-1/2}$ so that under smoothness of the model the bound (4) holds with $\beta_i = 1$. In the nonparametric framework the quality of the estimators deteriorates as the dimension of X_1 and X_2 gets large, a consequence of the so-called curse of dimensionality. E.g., for the Nadaraya-Watson estimator, Spiegelman and Sacks (1980) showed that (4) holds true with

$$\beta_i = \frac{2}{2 + d_i}.$$

Here again d_i is the dimension of $X_i, i = 1, 2$.

Remark 3. *Since (2) is concerned with standardized variables, the estimation error encountered in (4) has to become negligible compared with the sampling variances of the validation part. This may be achieved if the sample size of the learning sample, T, is large enough compared with n, the size of the validation sample.*

Remark 4. *Another issue is the choice of the weight function g. More or less this is up to the applicant of the statistical methodology. If one prefers to downweight large residuals, the absolute deviation function (or a robust variant) might be appropriate. Statistical inference is then based on the Mean Absolute Deviation (MAD), a popular means for goodness-of-fit in Robust Regression. If large deviations are to be upweighted large powers of e are in order. Note, however, that such a choice also requires higher moments for the errors ε. Therefore, in this paper, we shall focus only on g's which increase at most as fast as $g(u) = u^2$. In particular, g' and g'' (if they exist) are assumed to be Lipschitz and bounded, respectively. While Taylor expansion is an appropriate tool in the smooth case, g's with possible discontinuities need to be studied separately.*

Remark 5. *Prediction accuracy in regression has been often discussed in the context of model selection. See Efron (2004) for a discussion and review. For a given (nested) family of models one adds a penalty for the complexity of the model to the residual sum of squares. The resulting objective function is then minimized for a data set at hands. In our approach no penalty is considered and the regression need not be parametrically specified. Rather we study the unconditional distribution of the relevant quantity $\bar{d}$.*

Remark 6. *Concerning prediction, Granger and Newbold (1978) were pioneers in designing formal tests. Their procedure was based on the correlation of some combination of the residuals. Since then other authors extended their work into various directions. A simple but often applied test is due to Diebold and Mariano (1995). They compare two fully specified parameter-free models and compare the known errors. Typically, however, the assumed models are more complex to the effect, that the true errors are unknown and parameters need to be estimated. This changes the distributional character of the DM-test and the distribution presented in their work is not applicable.*

Hence it is not surprising that Clark and West (2004) found some bias in the DM-test when applied to residuals. In the context of classical goodness-of-fit tests based on the empirical distribution function, this is known at least since Durbin (1973). For the corresponding discussion in regression, see Stute (1997). To circumvent these problems one may, as in the present paper, split the whole sample of size $T + n$ into a learning sample of size T and a validation sample of size n.

1 Main Results

Let $(X_{1t}, X_{2t}, Y_t), 1 \leq t \leq T + n$, be a sample of independent random vectors with the same distribution as (X_1, X_2, Y), and let g be a given weight function. Recall $\bar{d}$ and $\bar{d}_1$.

Theorem 1. *Assume that g is differentiable such that g' is Lipschitz of order one, and let $\hat{m}_1$ and $\hat{m}_2$ be such that (4) is satisfied for $i = 1, 2$. Set $\beta = \min(\beta_1, \beta_2)$. Then, if $n = o(T^\beta)$, as n and T tend to infinity,*

$$n^{1/2}\bar{d} = n^{1/2}\bar{d}_1 + o_{\mathbb{P}}(1). \tag{5}$$

Furthermore, if g is twice continuously differentiable with

$$\mathbb{E}[g'(\varepsilon_i)|X_i] = 0 \text{ a.s. for } i = 1, 2 \tag{6}$$

then the assertion (5) holds true under the weaker condition $n = o(T^{2\beta})$.

Note that (6) is satisfied for $g(u) = u^2$.

While Theorem 2.1 covers the case of a differentiable g, in the next Theorem we consider the important special cases $g_1(u) = |u|$ and $g_2(u) = u^2 1_{\{|u|>\delta\}}$.

Theorem 2. *The expansion (5) holds true*

- *for g_1 whenever $n = o(T^\beta)$*
- *for g_2 whenever $n \ln n = o(T^{\beta/2})$ and*

the distribution function of $|\varepsilon|$ is differentiable at δ.

A careful check of the proof for g_2 shows that one may extend Theorem 2.1 to functions g with finitely many jumps but which are twice continuously differentiable in between. Details are omitted.

Our conditions on n and T show that $g(u) = u^2$ requires the smallest T. This is due to the fact that with this g the orthogonality of ε_i and $m_i(X_i)$ can be effectively used. On the other hand, discontinuities of g require a larger T and some smoothness of the distribution function of $|\varepsilon|$ in order to cope with the jump of g at δ.

Corollary 1. *Under the conditions of Theorem 2.1 or 2.2, we have*

$$\frac{n^{1/2}\bar{d}}{\hat{\sigma}_{\bar{d}}} \to \mathcal{N}(0,1) \text{ in distribution.}$$

Here $\hat{\sigma}_{\bar{d}}^2$ is a consistent estimator of the variance of $g(\varepsilon_1) - g(\varepsilon_2)$.

For example, we could take for $\hat{\sigma}_{\bar{d}}^2$ the sample variance of the $g(e_{1t}) - g(e_{2t}), T+1 \le t \le T+n$.

We have done several simulation results designed to support our theoretical findings with promising results.

References

[1] Billingsley, P. (1968), *Convergence of Probability Measures*, New York: Wiley.

[2] Clark, T., and West, K. (2004), "Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis," Working Paper 04-03, FRB of Kansas City, http://ssm.com/abstract=557101.

[3] Diebold, F., and Mariano, R. (1995), "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13, 253-263.

[4] Durbin, J. (1973), "Weak convergence of the sample distribution function when parameters are estimated," *Annals of Statistics*, 1, 279-290.

[5] Efron, B. (2004), "The estimation of prediction error: Covariance penalties and cross-validation," *Journal of the American Statistical Association*, 99, 619-632.

[6] Granger, C., and Newbold, P. (1977), *Forecasting economic time series*, Orlando, FL: Academic Press.

[7] Spiegelman, C., and Sacks, J. (1980), "Consistent window estimation in nonparametric regression," *Annals of Statistics*, 8, 240-246.

[8] Stone, C. (1977), "Consistent nonparametric regression," *Annals of Statistics*, 25, 613-641.

[9] Stute, W. (1997), "Nonparametric model checks for regression," *Annals of Statistics*, 25, 613-641.

Reduction of the Average Sample Number in Sequential Scheme of Testing Hypotheses

Sergey Postovalov, Madina Shakhmametova
Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: postovalov@ngs.ru, shakhmametova@gmail.com

Abstract

Problems of testing two simple hypotheses about the distribution of a random variable by the results of independent observations are considered. The main goal is finding critical values of SPRT, 2-SPRT and generalized optimal sequential Ayvazyan's test for testing hypotheses about normal and logistic distributions where probabilities of the first and second type errors have some specified values. It is shown that the use of the obtained critical values minimizes the average sample number as compared to the use of known theoretical approximate critical values.

Keywords: SPRT; 2-SPRT; generalized optimal sequential test; average sample number.

Introduction

The problem of testing hypotheses

$$H_0 : f(x) = f_0(x) \ vs. \ H_1 : f(x) = f_1(x)$$

of the random variable X based on the results of independent observations has various applications such as quality and reliability control, statistical regulation of technological processes, distinction of a signal against hindrances, etc. A common way of solving it is by the use of sequential testing. Methods of sequential analysis are characterized by a random moment of stopping observations. This moment depends on the values of data being observed. An advantage of sequential testing was shown by A. Wald [6] while solving the problem of testing two simple hypotheses. He proved that the average sample number of sequential testing methods is smaller than the sample number of any other method with the same probabilities of the errors of the first and second type where the sample number is fixed in advance. Wald formulated the sequential probability ratio test (SPRT) that became the most optimal one in the class of sequential testing methods. Later modifications of SPRT appeared, e.g., 2-SPRT formulated by Lorden [5], which is used now in different types of test problems [3], Li at al. [4], Bilias [2], and generalized optimal sequential Ayvazyan's test [1]. Similar to SPRT these modifications were developed for testing two simple hypotheses with the use of approximate critical values. We have a goal to define exact critical values for these tests, where probabilities of the first and second type errors are preset, by computer simulation. We consider testing the following hypotheses:

$$H_0 : f(x) = f_0(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(x-\theta)^2}{2}\}$$

$$H_1 : f(x) = f_1(x) = \frac{\pi}{\sqrt{3}} \exp\{-\frac{(\pi x)}{\sqrt{3}}\}/[1 + \exp\{-\frac{(\pi x)}{\sqrt{3}}\}]^2,$$

where θ is unknown parameter of normal distribution.

1 Sequential scheme of testing

Let us consider sequential probability ratio test schemes for two simple hypotheses. After a regular observation the statistics of sequential test are calculated according to formula (1) in the case of the SPRT and Ayvazyan's test and according to formula (2) for the 2-SPRT:

$$\lambda^{(n)} = \sum_{i=1}^{n} ln \frac{f_1(x_i)}{f_0(x_i)}. \tag{1}$$

$$\lambda_0^{(n)} = \sum_{i=1}^{n} ln \frac{f_2(x_i)}{f_0(x_i)}, \lambda_1^{(n)} = \sum_{i=1}^{n} ln \frac{f_1(x_i)}{f_2(x_i)}, \tag{2}$$

where $f_2 = \nu f_0 + (1-\nu) f_1$. An auxiliary parameter ν is defined by the condition of the Kulbak-Leybler's distances equality:

$$\int [f_2(x) - f_0(x)] ln \frac{f_2(x)}{f_0(x)} dx = \int [f_2(x) - f_1(x)] ln \frac{f_2(x)}{f_1(x)} dx$$

Regions of acceptance the hypotheses H_0 or H_1 are defined by formulas (3)-(5) for the SPRT, the 2-SPRT and the Ayvazyan's test correspondently:

$$\Lambda_n^{H_0} = \{\lambda : \lambda^{(n)} \le c_0\},\ \Lambda_n^{H_1} = \{\lambda : \lambda^{(n)} \ge c_1\} \tag{3}$$

$$\Lambda_n^{H_0} = \{\lambda_0 : \lambda_0^{(n)} \le c_0\},\ \Lambda_n^{H_1} = \{\lambda_1 : \lambda_1^{(n)} \ge c_1\}, \tag{4}$$

where $c_0 \approx \ln(\frac{\beta}{1-\alpha})$, $c_1 \approx \ln(\frac{1-\beta}{\alpha})$

$$\Lambda_n^{H_0} = \{\lambda : \lambda^{(n)} \le c_0(1 - \frac{n}{n^*})\},\ \Lambda_n^{H_1} = \{\lambda : \lambda^{(n)} \ge c_1(1 - \frac{n}{n^*})\}, \tag{5}$$

where $c_0 \approx -2\ln(\frac{1}{\beta})$, $c_1 \approx 2\ln(\frac{1}{\alpha})$ and n^* is a maximum number of observations within the Ayvazyan's test, which depends on the first and second type errors α, β:

$$n^* = f_0(x) = \frac{8}{\rho(H_0, H_1)} \ln \frac{1}{\min(\alpha, \beta)}$$

Figures (1)-(3) show the test process graphically.

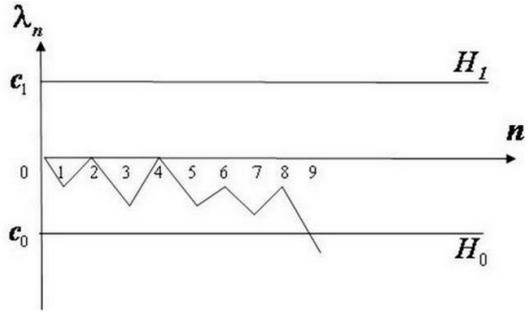


Figure 1: Procedure of acceptance the hypothesis H_0 by SPRT

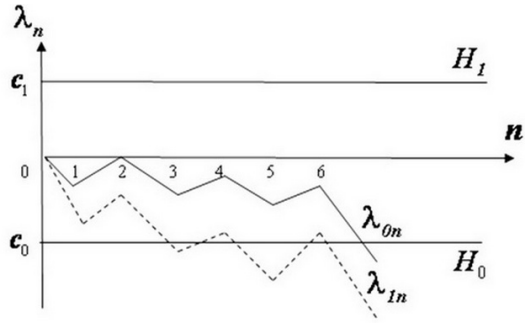


Figure 2: Procedure of acceptance the hypothesis H_0 by 2-SPRT

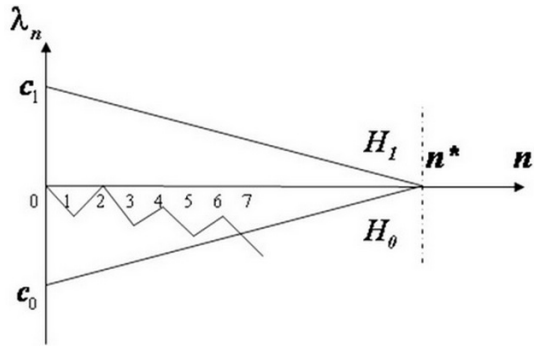


Figure 3: Procedure of acceptance the hypothesis H_0 by generalized optimal sequential Ayvazyan's test

1.1 Modification of the likelihood ratio statistics for randomly censored data

Let T be a continuous random variable with the distribution function F(t) determined on the interval $[0, \infty)$. The reliability function is defined as

$$R(t) = p\{T > t\} = \int_t^{\infty} f(u)du = 1 - F(t)$$

Let us introduce the notation:

$$q_i(x) = \begin{cases} f_i(x), & x \text{ - censored observation} \\ 1 - F_i(x), & x \text{ - uncensored observation,} \end{cases}$$

where $f_i(x)$ is a density function of the random variable ξ when the hypothesis H_i is true, i=0,1, and $F_i(x)$ is a distribution function. Then the sequential test statistic is defined as

$$\lambda^{(n)} = \sum_{i=1}^{n} ln\frac{q_1(x_i)}{q_0(x_i)},$$

where the summand contains either a ratio of the likelihood function for the uncensored points, or a ratio of reliability functions for censored points (depending on the x_i meaning, if it is the retirement or the failure).

1.2 Problem of the sequential tests exact critical boundaries calculation

If we need to calculate the exact boundaries C_0, C_1, we must calculate the probability of the first and second type errors:

$$1 - \alpha = P\{H_0|H_0\} = \sum_{i=1}^{\infty} P\{\lambda_i < C_0 \bigcap_{j=1}^{i-1} \lambda_j \in [C_0, C_1]|H_0\}$$

$$1 - \beta = P\{H_1|H_1\} = \sum_{i=1}^{\infty} P\{\lambda_i < C_1 \bigcap_{j=1}^{i-1} \lambda_j \in [C_0, C_1]|H_1\},$$

but this is a difficult mathematical task because λ_i depends on λ_j if $i > j$. Another possible way of solving this problem is by computer simulation.

2 Reduction the ASN of sequential tests by finding their exact critical boundaries

This section contains the description of Monte Carlo method computer simulation process and the obtained results, namely the exact critical boundaries and the average sample numbers of the sequential tests being considered.

2.1 The sequential tests exact critical boundaries calculation

The algorithm of finding critical values includes 3 steps.

1. Trying all values $c_0 \in [\underline{c_0}, 0], c_1 \in [0, \underline{c_1}]$ from the modeling area with a small step δ, we simulate dependences $\alpha(c_0, c_1), \beta(c_0, c_1)$.

2. We build lines of an equal level for selected values of probabilities of the first and second type errors $\alpha = 0.05, 0.1, 0.15$ and $\beta = 0.05, 0.1, 0.15$.

3. We obtain the required critical values (c_0, c_1) for the selected probabilities of the first and second type errors from the points where the lines $\alpha(c_0, c_1) = \alpha_i$ and $\beta(c_0, c_1) = \beta_i$ intersect.

2.2 Finding the exact critical boundaries for testing a simple hypothesis about normal distribution vs. simple hypothesis about logistic distribution

Before applying the algorithm described above, we find the modeling area by carrying out preliminary simulation on an obviously wider area with a big step and a small number of experiments. Then we carry out our algorithm containing simulation of 16600 experiments by the Monte Carlo method where probabilities of the first and second type errors are preset. As a result we obtain the functions $\alpha(c_0, c_1)$ and $\beta(c_0, c_1)$ lines of an equal level (see Figures 4-7) and the required critical values (see Tables 1-3) from the intersections of these lines.

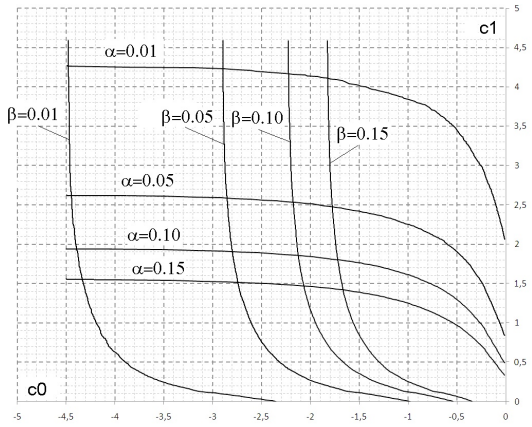


Figure 4: Lines of an equal level for testing two simple hypotheses about normal vs. logistic distribution by the SPRT

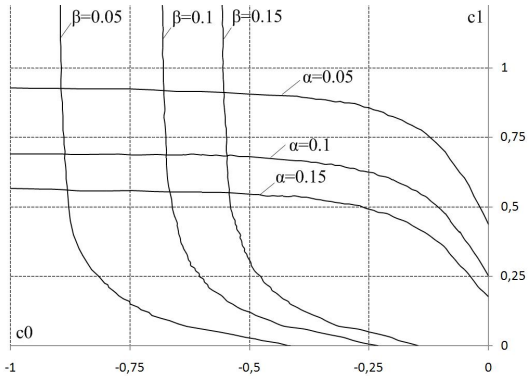


Figure 5: Lines of an equal level for testing two simple hypotheses about normal vs. logistic distribution by the 2-SPRT

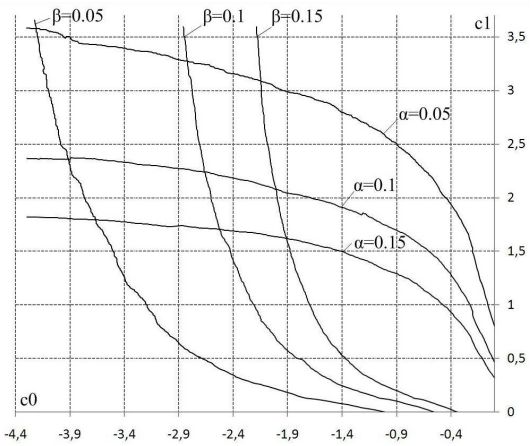


Figure 6: Lines of an equal level for testing a simple hypotheses about normal vs. logistic distribution by the Ayvazyan's test when $\min(\alpha, \beta) = 0.05$

Table 1: Exact critical values for testing a simple hypothesis about normal distribution vs. a simple hypothesis about logistic distribution by the SPRT

$\alpha \mid \beta$	0.15	0.1	0.05	0.01
0.15	-1.67, 1.42	-2.07, 1.47	-2.74, 1.52	-4.33, 1.56
0.1	-1.72, 1.81	-2.12, 1.86	-2.80, 1.91	-4.39, 1.95
0.05	-1.78, 2.48	-2.18, 2.54	-2.85, 2.59	-4.45, 2.63
0.01	-1.82, 4.11	-2.22, 4.17	-2.89, 4.23	-4.48, 4.26

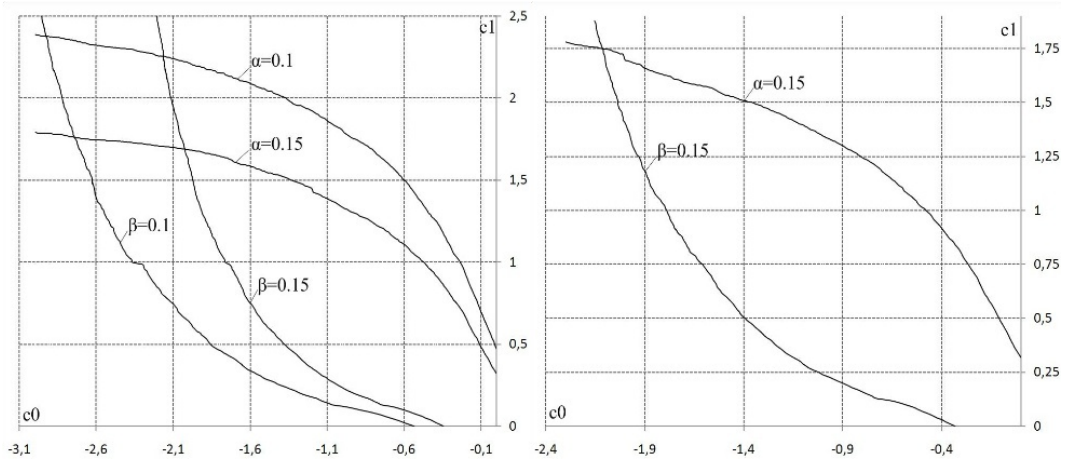


Figure 7: Lines of an equal level for testing a simple hypotheses about normal vs. logistic distribution by the Ayvazyan's test when $\min(\alpha, \beta) = 0.1$(left Figure), $\min(\alpha, \beta) = 0.15$ (right Figure)

Table 2: Exact critical values for testing a simple hypothesis about normal distribution vs. a simple hypothesis about logistic distribution by the 2-SPRT

$\alpha \mid \beta$	0.15	0.1	0.05
0.15	-0.54, 0.55	-0.67, 0.56	-0.89, 0.56
0.1	-0.55, 0.69	-0.68, 0.68	-0.89, 0.7
0.05	-0.56, 0.91	-0.68, 0.91	-0.89, 0.92

Table 3: Exact critical values for testing a simple hypothesis about normal distribution vs. a simple hypothesis about logistic distribution by the Ayvazyan's test

$\alpha \mid \beta$	0.15	0.1	0.05
0.15	-3.99, 3.57	-2.77, 3.29	-2.17, 3.12
0.1	-3.88, 2.36	-2.91, 2.37	-2.15, 2.21
0.05	-3.66, 1.78	-2.72, 1.79	-2.1, 1.74

2.3 Comparison of the sequential tests by an average sample number (ASN)

There have been made an ASN comparison of testing a simple hypothesis about normal distribution vs. a simple hypothesis about logistic distribution by different sequential schemes under the

assumption that the hypothesis H_0 is true. Tables 4 demonstrate the percent reduction in the ASN in the cases of exact and approximate critical boundaries. On the Figure 8 there is a chart of the gain in the ASN using the exact boundaries instead of the approximate with the first and second type errors set to 0.1.

Table 4: Reduction the ASN for SPRT, 2-SPRT and Ayvazyan's test

$\alpha \mid \beta$	0.15			0.1			0.05		
	SPRT	2-SPRT	Ayv.	SPRT	2-SPRT	Ayv.	SPRT	2-SPRT	Ayv.
0.15	11%	11%	47%	10%	12%	40%	10%	11%	27%
0.1	9%	8%	48%	9%	8%	42%	8%	9%	28%
0.05	7%	7%	58%	6%	6%	49%	6%	6%	34%

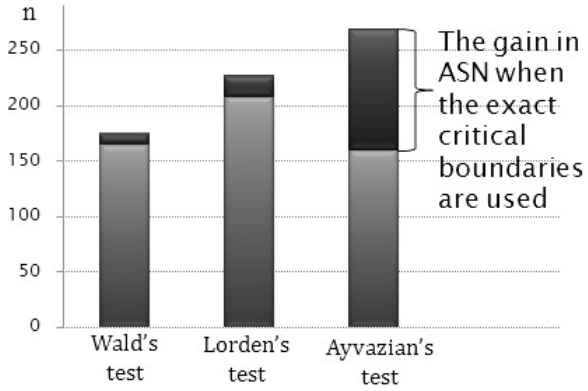


Figure 8: The gain in the ASN of sequential tests with the exact boundaries

Conclusions

We have found the exact critical boundaries of the SPRT, the 2-SPRT and the generalized optimal sequential Ayvazyan's test where probabilities of the first and second type errors have the specified values. The use of the exact critical boundaries decreases average sample number of the SPRT by 8-15%, average sample number of the 2-SPRT by 9-20% as compared to the use of approximate critical boundaries. In the case of the Ayvazyan's test up to twice ASN reduction is achieved.

References

[1] Ayvazyan S.A. (1965). Distinguishing Neighboring Hypotheses About Frequency Distributions in the Scheme of the Generalized Probability Ratio Test. *Theory of Probability and its Applications*. Vol. **10, 4**, pp. 646-658.

[2] Bilias Ya. (2000). Sequential testing of duration data: the case of the Pennsylvania 'reemployment bonus' experiment. *Journal of Applied Econometrics, Wiley&Sons, Ltd.*. Vol. **15, 6**, pp. 575-594.

[3] Lai T. L. (2001) Sequential Analysis: Some Classical Problems and New Challenges. *Statistica Sinica*. Vol. **11**, pp. 303-408.

[4] Li Ya., Pu X., Tsung F. (2009). Adaptive Charting Schemes Based on Double Sequential Probability Ratio Tests. *Quality and reliability engineering international*. Vol. **25**, pp. 21-39.

[5] Lorden G. (1976) 2-SPRTs and the modified Kiefer-Weiss problem of minimizing an expected sample size. *Annals of Statistics*. Vol. **4, 2**, pp. 281-292.

[6] Wald A. (1947) *Sequential analysis*. Wiley&Sons, New York.

A Chi-Squared Test for the Family of Inverse Gaussian Distributions for Censored Data

N. Saaidia
Department of Mathematics
Badji Mokhtar University, Annaba, Algeria
e-mail: solev@pdmi.ras.ru

Abstract

In this paper we propose a Chi-squared type test based on the Rao-Robson-Nikulin statistic under random censoring for the inverse Gaussian family.

Keywords: Pearson's Chi-squared test, Modified Chi-squared test, Inverse Gaussian distribution, Rondam censored data, Estimation, Rao-Robson-Nikulin statistic, Goodness-of-fit test, Maximum likelihood.

Introduction

In Reliability and survival analysis, we often encounter incomplete observations, and in this situation the usual methods are no longer valid. Several goodness-of-fit tests have been suggested by the statistician in the case of randomly right censored data. Habib and Thomas (1986) developed a Pearson-type Chi-squared statistic based on the differences of Kaplan-Meier estimate $\hat{S}_N(t)$ and parametric maximum likelihood estimators of survival functions $S(t, \hat{\theta}_N)$. Akritas (1988) proposed a Chi-squared statistic based on the idea of comparing the observed and expected number of failures in each class. Hjort (1990) developed a Chi-squared type statistic to test the validity of the parametric model for life history data based on the cumulative hazard process. Kim (1993) also proposed the Chi-squared goodness-of-fit test based on the product limit estimator. Solev and Nikulin (1999) constructed a type test Chi-square for double censored data. Finally Bagdonavičius and Nikulin (2011) gave several examples for distributions frequently used in reliability.

1 The Family of the Inverse Gaussian Distributions

Over a century the family of inverse Gaussian (**IG**) distributions had attracted the attention of many researchers in several fields [Seshadri (1993), Seshadri (1999), Chhikara and Folks (1989)], notably in reliability, survival analysis, actuarial science, histomorphometry, electrical networks, hydrology, management, demography, accelerated life testing, meteorology, mental health, physiology, economics and cardiology. In Reliability and survival analysis , the distributions that have a unimodal or $\cap-$shape hazard function are not too much, they include: log-normal, log-logistic, power generalized Weibull [Bagdonavičius et Nikulin (2002)], exponentiated Weibull [Mudholdkar and Srivastava (1993)] and **IG** distributions. The **IG** distribution offers certain advantages

over these distributions, because for these four distributions the hazard rate increases from 0 to its maximum value and then decreases to 0. For the **IG** distribution, the hazard rate increases from 0 to its maximum value and then decreases asymptotically to a constant which implies that the occurrence of a failure eventually becomes purely random and independent of past life. In contrast, the other four distributions vanishing hazard rate implies that eventually almost no possibility of failure remains, which is not reasonable for most real system [Gunes and all, 1996].

The **IG** family is mostly conveniently specified in terms of its density function,

$$f(x,\theta) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x \geq 0, \quad \theta = (\mu,\lambda)^T \in \mathbb{R}^1_+ \times \mathbb{R}^1_+ \subset \mathbb{R}^2,$$

All the positive and negative moments of the **IG** distribution exist. The corresponding survival function is

$$S(x,\theta) = \Phi\left(-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}-1\right)\right) - exp\left(\frac{2\lambda}{\mu}\right)\Phi\left(-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}+1\right)\right), \ \theta = (\mu,\lambda)^T,$$

where $\Phi(t)$ is the distribution function of the standard normal distribution.

The hazard rate function of **IG** distribution is

$$h(x,\theta) = \frac{\left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} exp\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\}}{\Phi\left(-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}-1\right)\right) - exp(\frac{2\lambda}{\mu})\Phi\left(-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}+1\right)\right)}, \ \theta = (\mu,\lambda)^T.$$

One can easily verify that this function is unimodal . This form is often applied in reliability and survival analysis.

We can also easily demonstrate that

$$\lim_{x\to\infty} h(x,\mu,\lambda) = \frac{\lambda}{2\mu^2}.$$

2 Goodness-of-fit test for the IG Family

Under the random censorship model, we assume that the failure times $T_1, T_2, \cdots, T_n$ are nonnegative and independent. The censoring variables $C_1, C_2, \cdots, C_n$ are also nonnegative, independent and assumed to be random sample and independent of $T_1, T_2, \cdots, T_n$. We observe only $X_i = min(T_i, C_i)$ and the indicator functions $\delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$.

Consider the problem of testing the composite hypothesis H_0 that the distribution of the sample T_i belongs to the **IG** family.

Let $S(t,\theta)$, $\theta = (\mu,\lambda)^T$ is the survival function (or reliability function) of **IG** distribution, $f(t,\theta)$ is the density function corresponding to $S(t,\theta)$, $G(t)$ the unknown survival function of consortship and $g(t)$ the density function corresponding to $G(t)$.

Denote by

$$\Lambda(t,\theta) = -lnS(t,\theta) = \int_0^t h(u,\theta)du,$$

the cumulative hazard function.

The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f^{\delta_i}(X_i,\theta)S^{1-\delta_i}(X_i,\theta)G^{\delta_i}(X_i)g^{1-\delta_i}(X_i).$$

Under *non-informative* censoring mechanism, that means that the survival function $G(t)$ and the density function $g(t)$ do not depend on the parameter θ, the expressions for the likelihood function becomes

$$L(\theta) = \prod_{i=1}^{n} f^{\delta_i}(X_i,\theta)S^{1-\delta_i}(X_i,\theta).$$

For the **IG** distribution, the log-likelihood function [Lemeshko et al. (2010)] is

$$\ell(\theta) = \ell(\mu,\lambda) = \sum_{i=1}^{n} \delta_i \left\{ \frac{1}{2}\ln\lambda - \frac{1}{2}\ln 2\pi - \frac{3}{2}\ln X_i - \lambda\frac{(X_i-\mu)^2}{2\mu^2 X_i} \right\} +$$

$$\sum_{i=1}^{n} (1-\delta_i)\ln\left\{ \Phi\left(A(X_i)\right) - exp(\frac{2\lambda}{\mu})\Phi\left(B(X_i)\right) \right\},$$

where

$$A(X_i) = A(X_i,\mu,\lambda) = -\sqrt{\frac{\lambda}{X_i}}\left(\frac{X_i}{\mu} - 1\right) \quad \text{and} \quad B(X_i) = B(X_i,\mu,\lambda) = -\sqrt{\frac{\lambda}{X_i}}\left(\frac{X_i}{\mu} + 1\right).$$

The score functions $U_l(\mu,\lambda)$, $l = 1,2$ [Lemeshko et al. (2010)] are

$$U_1(\mu,\lambda) = \frac{\partial\ell(\mu,\lambda)}{\partial\mu} = \frac{\lambda}{\mu^3}\sum_{i=1}^{n}\delta_i(X_i-\mu) +$$

$$\frac{1}{\mu^2}\sum_{i=1}^{n}(1-\delta_i)\frac{\sqrt{\lambda X_i}\varphi(A(X_i))+exp(\frac{2\lambda}{\mu})\left(2\lambda\Phi(B(X_i))-\sqrt{\lambda X_i}\varphi(B(X_i))\right)}{S(X_i,\mu,\lambda)},$$

$$U_2(\mu,\lambda)=\frac{\partial\ell(\mu,\lambda)}{\partial\lambda}=\sum_{i=1}^{n}\delta_i\left(\frac{1}{2\lambda}-\frac{(X_i-\mu)^2}{2\mu^2X_i}\right)+$$

$$\sum_{i=1}^{n}(1-\delta_i)\frac{\frac{1}{2\lambda}A(X_i)\varphi(A(X_i))-exp(\frac{2\lambda}{\mu})\left(\frac{2}{\mu}\Phi(B(X_i))+\frac{1}{2\lambda}B(X_i)\varphi(B(X_i))\right)}{S(X_i,\mu,\lambda)},$$

where $\varphi(t)$ is the density function of the standard normal distribution. To have the MLE $\hat{\theta}$ of θ one can solve the system formed by equalizing the score functions to zero.

It is easy to calculate the second partial derivatives $\ddot{\ell}(\theta)$ of the log-likelihood function.
The Fisher's information matrix is

$$I(\theta)=-E_\theta\ddot{\ell}(\theta),$$

where

$$\ddot{\ell}(\theta)=\sum_{i=1}^{n}\delta_i\frac{\partial^2}{\partial\theta^2}\ln h(X_i,\theta)-\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\Lambda(X_i,\theta).$$

Denote by τ the time for experiment, we suppose that all units fail or censored before τ. Divide the interval $[0,\tau]$ into k smaller intervals

$$I_j=(a_{j-1},a_j],\quad j=1,\cdots,k\quad a_0=0,\quad a_k=\tau,$$

and let denote by

$$U_j=\sum_{i:X_i\in I_j}\delta_i,$$

$$E_j=\sum_{i:X_i>a_{j-1}}\Big(\Lambda(a_j\wedge X_i,\hat{\theta})-\Lambda(a_{j-1},\hat{\theta})\Big)=\sum_{i:X_i>a_{j-1}}\ln\frac{S(a_{j-1},\hat{\theta})}{S(a_j\wedge X_i,\hat{\theta})}=$$

$$\sum_{i:X_i>a_{j-1}}\ln\frac{\Phi\left(A(a_{j-1},\hat{\mu},\hat{\lambda})\right)-e^{\frac{2\lambda}{\mu}}\Phi\left(B(a_{j-1},\hat{\mu},\hat{\lambda})\right)}{\Phi\left(A(a_j\wedge X_i,\hat{\mu},\hat{\lambda})\right)-e^{\frac{2\lambda}{\mu}}\Phi\left(B(a_j\wedge X_i,\hat{\mu},\hat{\lambda})\right)},$$

where $a \wedge b = \min(a, b)$.

Following Bagdonavičius and Nikulin (2011), to test the hypothesis H_0 we consider the Chi-squared type statistic

$$Y_n^2 = Z^T \hat{V}^- Z,$$

where

$$\hat{V}^- = \hat{A}^{-1} + \hat{A}^{-1}\hat{C}^T\hat{G}^-\hat{C}\hat{A}^{-1}, \quad \hat{G} = \hat{i} - \hat{C}\hat{A}^{-1}\hat{C}^T.$$

As mentioned in Bagdonavičius and Nikulin (2011), $\hat{V}^-$ is the general inverse of the matrix $\hat{V}$. But for inverse Gaussian distribution this matrix is not degenerate.

$$Z = (Z_1, Z_2, \cdots, Z_k)^T, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - E_j),$$

$$\hat{A} = \begin{pmatrix} \hat{A}_1 & 0 & \cdots & 0 \\ 0 & \hat{A}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{A}_k \end{pmatrix}, \quad \hat{A}_j = \frac{U_j}{n},$$

$$\hat{C} = \begin{pmatrix} \hat{C}_{11} & \hat{C}_{12} & \cdots & \hat{C}_{1k} \\ \hat{C}_{21} & \hat{C}_{22} & \cdots & \hat{C}_{2k} \end{pmatrix}, \qquad \hat{C}_{lj} = \frac{1}{n}\sum_{i:X_i \in I_j} \delta_i \frac{\partial}{\partial \theta_j} \ln h(X_i, \hat{\theta}),$$

for the **IG** distribution, the elements of the matrix $\hat{C}$ are

$$\hat{C}_{1j} = \frac{\hat{\lambda}}{n\hat{\mu}^3}\sum_{i:X_i \in I_j} \delta_i (X_i - \hat{\mu}) - \frac{\sqrt{\hat{\lambda}}}{n\hat{\mu}^2}\sum_{i:X_i \in I_j} \delta_i \frac{\sqrt{X_i}\varphi(A(X_i, \hat{\mu}, \hat{\lambda}))}{S(X_i, \hat{\mu}, \hat{\lambda})} -$$

$$\frac{e^{\frac{2\hat{\lambda}}{\hat{\mu}}}}{n\hat{\mu}^2}\sum_{i:X_i \in I_j} \delta_i \frac{2\hat{\lambda}\Phi(B(X_i, \hat{\mu}, \hat{\lambda})) - \sqrt{\hat{\lambda}X_i}\varphi(B(X_i, \hat{\mu}, \hat{\lambda}))}{S(X_i, \hat{\mu}, \hat{\lambda})},$$

$$\hat{C}_{2j} = \frac{1}{2n\hat{\lambda}}\sum_{i:X_i \in I_j} \delta_i - \frac{1}{2n\hat{\mu}^2}\sum_{i:X_i \in I_j} \delta_i \frac{(X_i - \hat{\mu})^2}{X_i} -$$

$$\frac{1}{2n\hat{\lambda}}\sum_{i:X_i \in I_j} \delta_i \frac{A(X_i, \hat{\mu}, \hat{\lambda})\varphi(A(X_i, \hat{\mu}, \hat{\lambda}))}{S(X_i, \hat{\mu}, \hat{\lambda})} +$$

$$\frac{e^{\frac{2\hat{\lambda}}{\hat{\mu}}}}{n}\sum_{i:X_i\in I_j}\delta_i\frac{\frac{2}{\hat{\mu}}\Phi(B(X_i,\hat{\mu},\hat{\lambda}))+\frac{1}{2\hat{\lambda}}B(X_i,\hat{\mu},\hat{\lambda})\varphi(B(X_i,\hat{\mu},\hat{\lambda}))}{S(X_i,\hat{\mu},\hat{\lambda})}.$$

$$\hat{i}=\begin{pmatrix}\hat{i}_{11} & \hat{i}_{12}\\ \hat{i}_{21} & \hat{i}_{22}\end{pmatrix},\qquad \hat{i}_{ll'}=\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial\ln h(X_i,\hat{\theta})}{\partial\theta_l}\frac{\partial\ln h(X_i,\hat{\theta})}{\partial\theta_{l'}}.$$

The partial derivatives of the function $\ln h$ with respect to μ and λ are

$$\frac{\partial\ln h(\mu,\lambda)}{\partial\mu}=\frac{\lambda(t-\mu)}{\mu^3}-\frac{1}{S(t,\theta)}\left\{\frac{\sqrt{\lambda t}}{\mu^2}\varphi(A(t,\mu,\lambda))+\right.$$
$$\left.\left(\frac{2\lambda}{\mu^2}\Phi(B(t,\mu,\lambda))-\frac{\sqrt{\lambda t}}{\mu^2}\varphi(B(t,\mu,\lambda))\right)exp(\frac{2\lambda}{\mu})\right\},$$

$$\frac{\partial\ln h(\mu,\lambda)}{\partial\lambda}=\frac{1}{2\lambda}-\frac{(t-\mu)^2}{2\mu^2 t}-\frac{1}{S(t,\theta)}\left\{\frac{1}{2\lambda}A(t,\mu,\lambda)\varphi(A(t,\mu,\lambda))-\right.$$
$$\left.\left(\frac{2}{\mu}\Phi(B(t,\mu,\lambda))+\frac{1}{2\lambda}B(t,\mu,\lambda)\varphi(B(t,\mu,\lambda))\right)exp(\frac{2\lambda}{\mu})\right\}.$$

The statistic Y_n^2 can be written in the form

$$Y_n^2=\sum_{j=1}^{k}\frac{(U_j-e_j)^2}{U_j}+Q,$$

where

$$Q=W^T\hat{G}^{-1}W,\quad W=\hat{C}\hat{A}^{-1}Z.$$

Under H_0 the statistic Y_n^2 possesses the Chi-squared distribution χ_k^2 with k degrees of freedom in the limit.

Choice of $\hat{a}_j$: Set

7, 34, 42, 63, 64, 74*, 83, 84, 91, 108, 112, 129, 133, 133, 139, 140, 140, 146, 149, 154, 157, 160, 160, 165, 173, 176, 185*, 218, 225, 241, 248, 273, 277, 279*, 297, 319*, 405, 417, 420, 440, 523*, 523, 583, 594, 1101, 1116*, 1146, 1226*, 1349*, 1412*, 1417.

Tab. 1. Arm A data for the Head-and-Neck cancer.

$$b_i = (n-i)\Lambda(X_{(i)}, \hat{\theta}) + \sum_{l=1}^{i} \Lambda(X_{(l)}, \hat{\theta}).$$

where $X_{(i)}$ is the ith element in the ordered statistics $(X_{(1)}, \cdots, X_{(n)})$.
If i is the smallest natural number verifying $E_j \in [b_{i-1}, b_i],\ j = 1, \cdots, k-1$ then:

$$(n-i+1)\Lambda(a, \hat{\theta}) + \sum_{l=1}^{i-1} \Lambda(X_{(l)}, \hat{\theta}) = E_j,$$

then

$$\hat{a}_j = \Lambda^{-1}\left([E_j - \sum_{l=1}^{i-1} \Lambda(X_{(l)}, \hat{\theta})]/(n-i+1), \hat{\theta}\right), \quad \hat{a}_k = \max(X_{(n)}, \tau),$$

where Λ^{-1} is the inverse of the cumulative hazard function Λ.

Reanalysis of the Arm A data for the Head-and-Neck cancer study

The survival times (in days) for patients of Arm A of the Head-and-Neck cancer trial (table 1) were first considered by Efron (1988) for 51 head-and-neck cancer patients. (*: censoring).

By transforming the data into months (1 month=30.438 days), Efron (1988) estimated the hazard function for these data by showing that it is unimodal by using the standard logistic regression techniques. Mudholkar et al. (1996) showed that the exponentiated Weibull distribution provide an acceptable fit to these data. Nikulin and Haghighi (2004) showed by using the Akritas test that the generalized Weibull distribution gives also a good fit for these data.

For the **IG** distribution, the MLE's $\hat{\theta}$ is

$$\hat{\theta} = (\hat{\mu}, \hat{\lambda})^T = (456.8747, 3.8889)^T, \quad \text{log-vrais} = -5.52968.$$

For a number of classes $k = 5$, the points $\hat{a}_j$ are

$$\hat{a}_1 = 1.689404,\ \ \hat{a}_2 = 2.922234,\ \ \hat{a}_3 = 4.755015,\ \ \hat{a}_4 = 9.441967.$$

For a significance level $\alpha = 0.05$, the critical value $\chi^2_k = 11.07050$. The observed value of RRN statistic is $Y^2_n = 9.823176$ hence we accept the hypothesis H_0. It means the **IG** distribution gives also a good fit for these data.

References

[1] Akritas, Michael G. 1988. Pearson-Type Goodness-of-Fit Tests: The Univariate Case. Journal of the American Statistical Association, 83 (401):222-230.

[2] Bagdonavičius, V., Kruopis, J., Nikulin, M. 2011. Non-parametric Tests for Censored Data, John Wiley & Sons Inc.

[3] Bagdonavičius V., Nikulin, M. 2002. Accelereted Life Models: Modeling and Statistical Analysis, Chapman and Hall.

[4] Bagdonavičius, V., Nikulin, M. 2011. Chi-squared goodness-of-fit test for right censored data. Int. J. Appl. Math. Stat, V. 24, NSI-11A.

[5] Efron, B. 1988. Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. Journal of the American Statistical Association 83, 414-425.

[6] Greenwood, P. S., Nikulin, M. 1996. A guide to Chi-squared Testing, John Wiley and Sons, New York.

[7] Gunes, H., Dietz, D.C., Auclai, P.F., Moor, A.H. 1997. Modified goodness-of-fit tests for the inverse Gaussian distribution Computational statistics and Data analysis, 24, p.63-77.

[8] Habib, M.G., Thomas, D.R. 1986. Chi-squared Goodness-of-Fit Tests For Randomly Censored Data, Annals of Statistics, V.14, N2, p 759-765.

[9] Haghighi, F. 2004. Fiabilité et Analyse de Survie: Le Modèle semi-paramétrique de Dégradation. Une généralistion de la famille Weibull et ses application, Thèse de Doctorat, Université Victor Segalen Bordeaux 2.

[10] Hjort, N.L. 1990. Goodness of Fit Test in Models for Life History Data Based on Cumulative Hazard Rates, The annals of statistics, V.18, N3, p1221-1258.

[11] Kim, J.H. 1993. Chi-squared Goodness-of-Fit Tests For Randomly Censored Data, Annals of Statistics, the annals of statistics, V.21, N3, p 1621-1639.

[12] Lemeshko, B.Y., Lemeshko, S.B., Akushkina, K.A., Nikulin, M., Saaidia, N. 2010. Inverse Gaussian Model and Its Applications in Reliability and Survival Analysis, In: Mathematical and Statistical Models and Methods in Reliability (Eds. Rykov, V.V., Balakrishnan, N., Nikulin, M.S.), Birkhäuser Boston, p 293-315.

[13] Mudholkar, G.S, Srivastava D., Kollia, G. D. 1996. A generalization of the Weibull distribution with application, J. Amer. Statist. Assoc. 91, 15751583 .

[14] Nikulin, M.S., Haghighi, F. 2004. A Chi-Squared Test for the Generalized Power Weibull Family for the Head-and-Neck Cancer Censored Data, journal of Mathematical Sciences Volume 133, Number 3, 1333-1341.

[15] Nikulin, M.S. 1973a. Chi-square Test For Continuous Distributions with Shift and Scale Parameters, teor. Veroyatn. Primen., 18, No. 3, 559-568.

[16] Nikulin, M. S. (1973b) On a chi-square test for continuous distributions, Theory of Probability and its Applications, 18, 638639.

[17] Nikulin, M.S., Saaidia, N. 2009. Inverse Gaussian family and its applications in reliability, Study by simlulation, St. Petersburg, June 28–July 4. VVM comm. Ltd., St. Petersburg **V. 2**, 657-661.

[18] Nikulin, M.S., Solev, V.N. 1999. Chi-squares Goodness-of-fit Test for Doubly Censored Data with Applications in Survival Analysis and Reliability, In: Statistical and Probabilistic Models in Reliability, D.C.Ionescu, N.Limnios (eds), Birkhauser, Boston, 101-112.

[19] Saaidia, N. Seddik-Ameur, N. 2010. Chi-Squared Type Test for the Inverse Gaussian Distribution, les annales de l'ISUP, Volume LIV - Fascicule 3, pp. 67-84.

[20] Saaidia, N. (2011) La famille des lois Gaussiennes inverses et ses applications en Fiabilité et Analyse de Survie. Doctoral thesis, Badji Mokhtar University Annaba, Algeria / Victor Segalen University Bordeaux, France.

[21] Van Der Vaart, A.W. 1998. Asymptotic Statistics. Cambridge Series in Statistics and probabilistic Mathematics, Cambridge: Cambridge University Press.

Application of Classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling Tests for Censored Samples

Ekaterina Chimitova[1], Hannelore Liero[2] and Mariya Vedernikova[1]
[1] *Novosibirsk State Technical University, Novosibirsk, Russia*
[2] *Institute of Mathematics, University of Potsdam, Potsdam, Germany*
e-mail: chim@mail.ru, liero@uni-potsdam.de, vedernikova.m.a@gmail.com

Abstract

Problems of testing statistical goodness-of-fit hypotheses for censored data are considered in the paper. The application of the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for a complete sample obtained from original censored sample by using randomization is proposed. By means of computer simulation methods we have investigated the test statistic distributions and the power of considered tests for close competing hypotheses when testing simple and composite hypotheses.

Keywords: censored data, goodness-of-fit testing, Kolmogorov test, Cramer-von Mises-Smirnov test, Anderson-Darling test, randomization.

Introduction

In the survival and reliability analysis lifetimes are typically right censored. A lot of papers describing various approaches for testing goodness-of-fit for censored samples have been published recently. In [1], [3] and [10] modifications of the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests are given. But these modifications cannot be applied in the case of randomly censored samples. The classical tests above can be modified by using the Kaplan-Meier estimate instead of the empirical distribution function in the formulas of statistics (see, for example, [4], [5], [9], [11]).

The limiting distributions of test statistics based on the Kaplan-Meier estimate are unknown, because they considerably depend on the type of censoring and censoring degree [8]. In the case of randomly censored samples the distributions of considering test statistics also depend on the distribution of censoring times, which in turn is usually unknown in practice. It is necessary to simulate statistic distribution for testing goodness-of-fit by using modified nonparametric tests. The process of modeling statistic distribution requires time and computational costs and it is not possible without special software. This is a significant disadvantage of the modified nonparametric goodness-of-fit tests for censored samples.

A good possibility was shown in [7] to use the Smirnov transformation with randomization for the correct application of classical goodness-of-fit tests for grouped and censored data in case of testing a simple hypothesis. In this paper we develop this idea for testing simple and composite hypotheses by randomly censored samples.

1 Nonparametric Goodness-of-Fit Tests

In this paper we consider testing simple goodness-of-fit hypotheses $H_0 : F = F_0$ and composite hypotheses which can be presented as $H_0 : F \in \{F_0(\cdot;\theta), \theta \in \Theta\}$.

One approach for testing these hypotheses by complete samples is the application of the classical nonparametric tests: Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests. The Kolmogorov test statistic

$$D_n = \sup_{-\infty<t<\infty} |F_n(t) - F_0(t;\theta)|,$$

where $F_n(t)$ is the empirical distribution function and n is the sample size. In testing hypotheses, the statistic is usually used with Bolshev's correction [2] of the form

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

where $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n} - F_0(t_i;\theta)\right\}$, $D_n^- = \max\limits_{1\le i\le n}\left\{F_0(t_i;\theta) - \frac{i-1}{n}\right\}$.

The Cramer-von Mises-Smirnov test statistic

$$S_\omega = \int_{-\infty}^{\infty} (F_n(t) - F_0(t;\theta))^2 \, dF_0(t;\theta) \tag{2}$$

can be calculated by the following formula

$$S_\omega = \frac{1}{12n} + \sum_{i=1}^{n} \left\{ F_0(t_i;\theta) - \frac{2i-1}{2n} \right\}^2.$$

The Anderson-Darling test statistic

$$S_\Omega = \int_{-\infty}^{\infty} (F_n(t) - F_0(t;\theta))^2 \frac{dF_0(t;\theta)}{F_0(t;\theta)(1 - F_0(t;\theta))} \tag{3}$$

can be calculated as follows

$$S_\Omega = -n - 2\sum_{i=1}^{n} \left\{ \frac{2i-1}{2n} \ln F_0(t_i;\theta) + \left(1 - \frac{2i-1}{2n}\right) \ln(1 - F_0(t_i;\theta)) \right\}.$$

Let us denote the distribution of a test statistic under hypothesis H_0 as $G(S\,|H_0)$. In the case of testing simple hypotheses the distributions $G(S\,|H_0)$ of considering statistics do not depend on the hypothetical distribution, in the limit statistic S_K belongs to the Kolmogorov distribution, S_Ω and S_ω belong to the $a1$ and the $a2$ distributions, respectively [2]. For composite hypotheses nonparametric test statistic distributions $G(S\,|H_0)$ are affected by a number of factors: the

form of tested lifetime distribution $F_0(t;\theta)$, the type and the number of estimated parameters, the method of parameter estimation and other factors. Approximations of limiting statistic distributions for testing various composite hypotheses when using maximum likelihood estimates of unknown parameters have been proposed in [6]. In this paper we denote these approximations as $\tilde{G}(S)$.

2 Censored Sample Transformation to Complete One by Using Randomization

In reliability or survival studies the observed data are usually presented as $(X_1, \delta_1), \ldots, (X_n, \delta_n)$, $X_i = \min(T_i, C_i)$ is the observation value, T_i is an observed lifetime and C_i is a censoring time. The identifier $\delta_i = 1$ if $X_i = T_i$ and $\delta_i = 0$ if $X_i = C_i$. Let lifetime T and censoring time C be independent random variables from the distribution functions $F(t)$ and $F^C(t)$, respectively. All lifetimes and censoring times are assumed mutually independent. Then censoring is called *independent random censoring*.

We propose transforming the original censored sample to a complete sample by using randomization. Randomization makes it possible to use the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for censored samples.

First, let us consider the simple hypothesis. That is, F_0 is completely known. For each censored random variable, that is for X_i with $\delta_i = 0$ a value $\hat{T}_i$ is generated by $\hat{T}_i = F_0^{-1}(\xi_i)$, where $\xi_i \sim U[F_0(C_i), 1]$. The corresponding values of censored observations in the original sample are replaced by generated values. Hereby, we obtained a transformed complete sample $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$, in which $\hat{X}_i = X_i$ if i-th observation is complete ($\delta_i = 1$) and $\hat{X}_i = \hat{T}_i$ if one is censored ($\delta_i = 0$).

2.1 Testing Simple Hypotheses

In the case of testing simple hypotheses by nonparametric goodness-of-fit tests based on transformed samples the known parameter values θ are used in randomization. The values $\hat{X}_i$ in transformed sample are distributed as follows

$$
\begin{aligned}
P(\hat{X}_i \le t) &= P(\hat{X}_i \le t | T_i \le C_i) P(T_i \le C_i) + P(\hat{X}_i \le t | T_i > C_i) P(T_i > C_i) = \\
&= P(T_i \le t | T_i \le C_i) P(T_i \le C_i) + P(\hat{T}_i \le t | T_i > C_i) P(T_i > C_i) = \\
&= P(T_i \le t, T_i \le C_i) + P(\hat{T}_i \le t, T_i > C_i) = \\
&= \int P(x \le t, x \le C_i) dF(x) + \int P(\hat{T}_i \le t, T_i > c | C_i = c) dF^C(c) = \\
&= \int_0^t (1 - F^C(x_-)) dF(x) + \int P(\xi_i \le F(t), T_i > c | C_i = c) dF^C(c) = \\
&= \int_0^t (1 - F^C(x_-)) dF(x) + \int P(\xi_i \le F(t) | T_i > c, C_i = c) P(T_i > c | C_i = c) dF^C(c) = \\
&= \int_0^t (1 - F^C(x_-)) dF(x) + \int_0^t \frac{F(t) - F(c)}{1 - F(c)} (1 - F(c)) dF^C(c) = \\
&= \int_0^t (1 - F^C(x_-)) dF(x) + \int_0^t (F(t) - F(c)) dF^C(c) = F(t).
\end{aligned}
$$

Therefore, under the simple null hypothesis the distribution the completed sample is F_0. So, the algorithm of testing simple hypotheses can be written as follows:

1. Specify the significance level α;
2. Transform the original censored sample $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ to complete sample $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$ by using randomization;
3. Calculate the test statistic S^* ((1), (2) or (3)) by the obtained sample and the theoretical distribution F_0;
4. Compute the p-value: $p = 1 - G(S^* | H_0)$;
5. The hypothesis H_0 is rejected if obtained p-value is less then α.

2.2 Testing Composite Hypotheses

In the case of testing composite hypotheses maximum likelihood estimates $\hat{\theta}$ obtained by the original censored sample are used in the randomization. We define

$$\hat{T}_i = F^{-1}(\xi_i, \hat{\theta}) \qquad \text{with} \qquad \xi_i \sim U\left[F(C_i; \hat{\theta}), 1\right].$$

Hereupon, the distribution of simulated values $\hat{X}_i$ can differ from the "true" distribution of X_i. After replacing censored observations $(X_i = C_i, \delta_i = 0)$ by generated observations with values $\hat{T}_i$ it is necessary to estimate unknown parameters θ again. The algorithm of testing composite hypotheses by using randomization can be written as follows

1. Specify the significance level α;
2. Calculate the maximum likelihood estimates $\hat{\theta}$ by the original sample $(X_1, \delta_1), \ldots, (X_n, \delta_n)$;
3. Transform the sample $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ to complete sample $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$ by using randomization;
4. Calculate the maximum likelihood estimates $\tilde{\theta}$ by the transformed sample $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$;
5. Calculate the test statistic S^* ((1), (2) or (3)) by the obtained sample and the theoretical distribution $F_0(\cdot; \tilde{\theta})$;
6. Compute the p-value: $p = 1 - G(S^*|H_0)$;
7. The hypothesis H_0 is rejected if obtained p-value is less then α.

Transformation may lead to deviations of statistic distributions from limiting distributions. It is necessary to investigate test statistic distributions $G(S\,|H_0\,)$ obtained by using transformed samples.

3 Simulation Results

We simulate empirical statistic distributions of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests when testing simple and composite goodness-of-fit hypotheses for various sample sizes and distributions of censoring times $F^C\,(t)$.

3.1 Testing Simple Hypotheses

By computer simulation methods we have shown that test statistic distributions based on the transformed samples when testing simple hypotheses do not depend on the distribution $F^C(t)$ and coincide with the corresponding limiting distribution (for example, the Kolmogorov distribution for Kolmogorov's test statistic).

Figure 1 illustrates the empirical distribution of the Kolmogorov test statistic obtained by completed samples of the size $n = 300$, when testing the simple hypothesis H_0: Weibull distribution $F_0\,(x;\theta) = 1 - \exp\left(-\left(\frac{x-\theta_1}{\theta_2}\right)^{\theta_3}\right)$ with parameter values $\theta_1 = 0,\ \theta_2 = 2,\ \theta_3 = 2$, and the Kolmogorov distribution. Completed samples were obtained from censored samples in which censoring times are distributed from the Weibull distribution with parameter values $\theta_1 = 0,\ \theta_2 = 1.87,\ \theta_3 = 3.74$. These values of parameters were taken so that the censoring degree would be approximately equal to 50%. The number of simulated samples is $N = 10000$.

As you can see from Figure 1 the empirical distribution of the Kolmogorov statistic by completed samples practically coincide with the Kolmogorov distribution. Similar results were

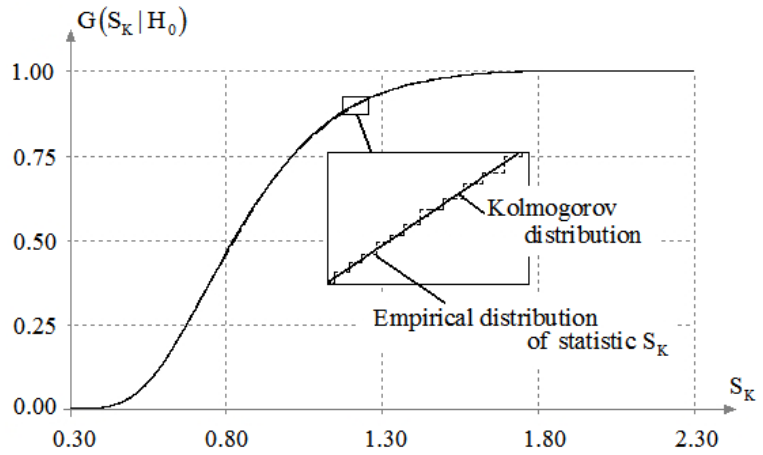


Figure 1: Kolmogorov's statistic distribution

obtained for the Cramer-von Mises-Smirnov and Anderson-Darling tests. This fact confirms the result given in Section 2.1 that under H_0 the distribution of completed samples is F_0.

Let us compare the power of goodness-of-fit tests by using randomization for different distributions of censoring times $F^C(t)$ at the example of the following pair of close simple competing hypotheses.

H_0: Weibull distribution $F_0(x;\theta) = 1 - \exp\left(-\left(\frac{x-\theta_1}{\theta_2}\right)^{\theta_3}\right)$ with parameter values $\theta_1 = 0,\ \theta_2 = 2,\ \theta_3 = 2$;

H_1: gamma-distribution $F_0(x;\theta) = \frac{1}{\Gamma(\theta_3)}\Gamma\left(\frac{x-\theta_1}{\theta_2},\ \theta_3\right)$ with parameter values $\theta_1 = 0,\ \theta_2 = 0.5577,\ \theta_3 = 3.1215$.

In Table 1 there are the estimates of the power for considered goodness-of-fit tests applied to samples, which were obtained by randomization from censored samples with $F^C(t)$ from

- Beta I distribution (the corresponding rows are denoted as "B"),
- Weibull distribution (the corresponding rows are denoted as "W").

The significance level $\alpha = 0.1$, the sample size $n = 200$.

As it is seen from the table if the censoring degree is not large ($\leq 40\%$) there are no significant losses in the power of nonparametric goodness-of-fit tests by completed samples comparing with the power of tests by originally uncensored samples. When testing simple hypotheses the distribution of censoring times does not influence the power of tests in study.

3.2 Testing Composite Hypotheses

Figure 2 illustrates the dependence of Kolmogorov's statistic distributions on the censoring distribution $F^C(t)$ in case of testing composite hypotheses. There are two empirical distributions

Table 1: Test Power for Simple Hypotheses

Censoring Degree		0%	10%	20%	30%	40%	50%	60%	70%	80%
Kolmogorov	B	0.35	0.33	0.30	0.29	0.28	0.25	0.21	0.16	0.15
Test	W		0.33	0.33	0.33	0.31	0.27	0.21	0.17	0.12
Cramer-von	B	0.35	0.34	0.32	0.28	0.28	0.28	0.25	0.21	0.19
Mises-Smirnov Test	W		0.34	0.33	0.33	0.32	0.28	0.22	0.17	0.12
Anderson-Darling	B	0.35	0.34	0.33	0.29	0.30	0.29	0.23	0.21	0.19
Test	W		0.32	0.32	0.32	0.32	0.27	0.22	0.18	0.12

of the statistic obtained by $N = 10000$ completed samples of the size $n = 300$. In the first case original censored samples had the Weibull distribution of censoring times and in the second one $F^C(t)$ is the Beta I distribution. The censoring degree is about 50%.

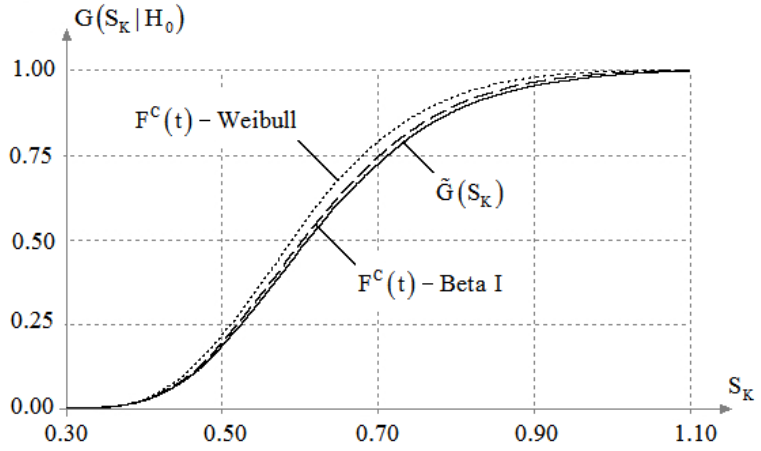


Figure 2: Kolmogorov's statistic distribution for various censoring distributions

As it is seen from the figure, the empirical test statistic distribution in the case of Weibull censoring distribution deviates from the approximation of limiting distribution $\tilde{G}(S_K)$ (gamma-distribution with parameters $\theta_1 = 0.2598$, $\theta_2 = 0.0563$, $\theta_3 = 6.6012$, see [6]) more than the statistic distribution of the same test in the case of Beta I censoring distribution. The difference between these two cases is the following. For the distribution of lifetimes $F(t)$ from the Weibull distribution with parameters $\theta_1 = 0$, $\theta_2 = 2$, $\theta_3 = 2$: if the censoring distribution is the Beta I distribution, then censored observation values are "uniformly distributed" in a variational series. If $F^C(t)$ is the Weibull distribution then the censored observation values are focused at the end of variational series, and in this case statistic distributions of classical tests applied for completed samples can be far from the $\tilde{G}(S_K)$. Thus, the closeness of statistic distributions using

randomization to the corresponding approximation strongly depends on the location of censored observations in the variational series of the original sample.

In Figure 3 there are the empirical distributions of Kolmogorov's statistic obtained by samples which completed from censored samples with various censoring degrees. The size of original samples is $n = 300$. The number of simulated samples is $N = 10000$.

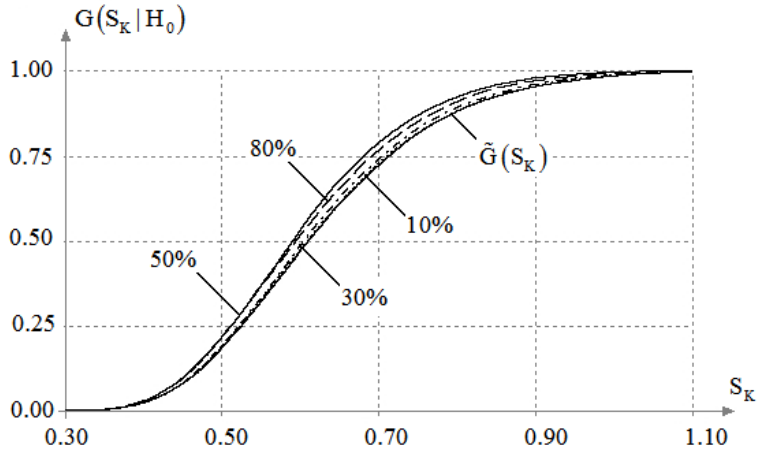


Figure 3: Kolmogorov's statistic distribution for various censoring degrees

It may be noticed that statistic distributions depend on censoring degrees. When the censoring degree is less than 30%, statistic distributions practically coincide with approximation of the limiting distribution $\tilde{G}(S_K)$ obtained by testing composite hypotheses for complete samples [6]. When the censoring degree increases, statistic distributions are on the left side from approximation. A similar dependence of statistic distributions is observed for Cramer-von Mises-Smirnov and Anderson-Darling tests by using the randomization.

Let us compare the power of goodness-of-fit tests by using randomization for different distributions of censoring times $F^C(t)$ at the example of the following pair of close competing hypotheses - H_0: Weibull distribution and H_1: gamma-distribution.

In Table 2 there are the estimates of the power for considered goodness-of-fit tests applied to transformed samples from censored samples with $F^C(t)$ from Beta I distribution ("B") and Weibull distribution ("W"). The significance level $\alpha = 0.1$, the sample size $n = 200$.

As it is seen from Table 2 when testing the composite hypothesis the power of considered tests is influenced by the distribution of censoring times $F^C(t)$: in the case when censored observations are "uniformly distributed" in a variational series the tests power is larger than when censored observations are located at the end of the empirical distributions of an original censored sample.

Comparing results presented in Tables 1 and 2 it may be noticed that the losses in the power (in comparison with the test power for originally uncensored samples) of nonparametric goodness-of-fit tests by completed samples for composite hypothesis are larger than in the case

Table 2: Test Power for Composite Hypotheses

Censoring Degree		0%	10%	20%	30%	40%	50%	60%	70%	80%
Kolmogorov	B	0.42	0.37	0.33	0.28	0.24	0.20	0.16	0.13	0.10
Test	W		0.33	0.27	0.21	0.17	0.14	0.10	0.09	0.08
Cramer-von	B	0.53	0.45	0.39	0.32	0.28	0.22	0.17	0.13	0.10
Mises-Smirnov Test	W		0.39	0.31	0.24	0.18	0.14	0.10	0.09	0.08
Anderson-Darling	B	0.59	0.49	0.44	0.35	0.29	0.23	0.18	0.13	0.10
Test	W		0.41	0.32	0.24	0.18	0.13	0.10	0.09	0.08

of testing simple hypothesis.

Conclusions

- When testing simple goodness-of-fit hypotheses the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling test statistic distributions, obtained by transformed samples, agree with the limiting Kolmogorov, $a1$ and $a2$ distributions, respectively.

- When testing composite goodness-of-fit hypotheses the statistic distributions practically do not differ from corresponding approximations of the limiting distributions, if the censoring degree is small ($< 30\%$).

- If the censoring degree is small there are no significant losses in the power of nonparametric goodness-of-fit tests by using randomization comparing with the power tests by complete samples.

So, the method of testing goodness-of-fit for censored samples by using randomization, unlike most others ways for testing goodness-of-fit for censored samples, does not require large computational costs. For the reason stated above, if the censoring degree is not large ($< 30\%$) the application of classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests by transformed samples is the easy and effective solution of testing goodness-of-fit for censored samples.

Acknowledgements

These research have been supported by the Federal targeted program of the Ministry of Education and Science of the Russian Federation “Scientific and scientific-educational personnel of innovative Russia”.

References

[1] D.M. Barr and T. Davidson. A Kolmogorov-Smirnov test for censored samples. *Technometrics*, 15, 4, 1973.

[2] L.N. Bolshev, N.V. Smirnov. *Tables of Mathematical Statistics*, Moscow: Science. (in Russian), 1983.

[3] R.B. D'Agostino and M.A. Stephens. *Goodness of Fit Techniques*, New York: Marcel Dekker, 1986.

[4] N.L. Hjort. On Inference in Parametric Survival Data. *International Statistical Review*, 60, 3, 355–387, 1992.

[5] J.A. Koziol and S.B. Green. A Cramer-von Mises Statistic for Randomly Censored Data. *Biometrika*, 63, 3, 465–474, 1976.

[6] B.Yu. Lemeshko, S.B. Lemeshko and S.N. Postovalov. Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses. *Communications in Statistics - Theory and Methods*, 39, 3, 460–471, 2010.

[7] B.Yu. Lemeshko, E.V. Chimitova and S.S. Kolesnikov. Nonparametric goodness-of-fit tests for discrete, grouped or censored data. *ASMDA'2007 International Conference. Book of Abstracts.*

[8] B.Yu. Lemeshko, E.V. Chimitova and T.A. Pleshkova. Testing simple and composite goodness-of-fit hypotheses by censored samples. *Nauchniy vestnik NGTU*, 4(41), 13–28, 2010.

[9] V. Nair. Plots and tests for goodness of fit with randomly censored data. *Biometrika*, 68, 99–103, 1981.

[10] A.N. Pettitt and M.A. Stephens. Modified Cramer von Mises statistics for censored data. *Biometrika*, 63, 2, 1976.

[11] D. Reineke and J. Crown. Estimation of Hazard, Density and Survival Functions for Randomly Censored Data. *Journal of Applied Statistics*, 31, 10, 1211–1225, 2004.

Part IV
Nonparametric Methods

Using FRiS-Function for Nine Medical Tasks Solving

Nikolay Zagoruiko[1,2], Irina Borisova[1,2], Vladimir Dyubanov[1],
Olga Kutnenko[1,2]
[1]*Institute of Mathematics SD RAS*
[2]*Design Technological Institute of Digital Techniques SD RAS*
Novosibirsk, Russia
e-mail: zag@math.nsc.ru

Abstract

In tasks of modern biology the quantity of features often on orders exceeds quantity of objects. For the decision of such tasks the method Data Mining based on use the new measure of similarity between objects in the form of Function of Rival Similarity (FRiS) is offered. On this basis the methods of a quantitative estimation of compactness of patterns, of a construction decision rules and of feature selection are developed. High efficiency of methods is illustrated by results of the decision of nine tasks of recognition of diseases on microarray dataset.

Keywords: feature selection, FRiS-function, compactness.

Introduction

In a modern stream of Data Mining tasks often there are tasks, in which the quantity of attributes N on orders exceeds quantity of objects M. It is especially characteristic for biological and biomedical research, using the information about the expression levels of genes. If N is equal tens thousand, and the quantity of relevant attributes needed to be selected n - to several tens, its exact decision connected with search of all combinations from N on n, is impossible. For this reason there are tens of algorithms which for polynomial time provide the decisions to some extent close to optimum. The big variety of these algorithms generates a task of **a choice of the best algorithm of the feature selection**.

The first stage of elimination of weak methods can be made by the frequency of their use in scientific publications. This approach is described in work [6]. Its authors have taken the trouble selection of most often used methods and have lead huge work on their comparative test. The list of applicants includes 10 following methods: significance analysis of microarrays, analysis of variance (ANOVA), empirical Bayes t-statistic, template matching, maxT, between group analysis, area under the receiver operating characteristic (ROC) curve, the Welch t-statistic, fold change and rank products.

These methods were applied to 9 different two class microarray datasets. These datasets include the well-known publicly available colon [1], lymphoma [10] and leukemia datasets [3, 5]. For recognition a four decision rules were used: support vector machines (SVM) [12], between group analysis (BGA) [4], naive Bayes classification (NBC) [9], and k-nearest neighbours (kNN) [7].

Each feature selection method was applied to training datasets and the most highly ranked genes were selected to generate gene lists of length between 2 and 100 genes. The ability of these gene lists to form successful classifiers was evaluated. To limit sampling bias during cross-validation (CV) a 50% sample data were randomly split into two equal groups 10 times. The first group was used as a training dataset for feature selection and classifier training.In training and test cross-validation, all four classification methods were applied. The prediction success of each model was assessed using the blind test dataset, which were not used for feature selection of gene lists or training of classifiers. Quality of the decision of each task was estimated by two characteristics: the percentage accuracy scores and the RCI scores [11].

We used the results of article [6] for estimation the quality of our algorithm of the feature selection FRiS-GRAD in a combination with decision rule FRiS-Stolp.

In sections 2 and 3 we shall describe FRiS and its use for construction of a decision rule and for estimation of the informativeness of attributes. In section 4 the feature selection algorithm is presented. In section 5 the decision of 9 tasks by these methods are described.

1 Function of Rival Similarity

Usually the distance and the similarity in a metric space are considered as absolute categories. Actually, if used standard units are known (meters, micrometers, etc) then the distance is measured in absolute scale. Nevertheless, the similarity measure used in pattern recognition cannot be measured in such way. When recognizing membership of the object z to one of two patterns A or B it is important to know the distance to the pattern A as well as the distance to the competitive pattern B and to compare these distances with each other. It is means that similarity is not an absolute but **relative category** in terms of pattern recognition. answer the question "how much is an object z similar to an object a?", we need to know the answer to another question: "in comparison with what?"

All statistical recognition algorithms consider the competition between classes. In terms of the method of "k nearest neighbors" (kNN), a new object z is recognized as an object of the pattern A if the distance $r(z, A)$ to this pattern is less than the distance $r(z, B)$ to the nearest competitive pattern B. Similarity in this algorithm is estimated in scale of order.

We propose to use the following form of the Function of Rival Similarity (FRiS) of an object z with an object a in competition with an object b [13]:

$$F(z, a|b) = [R(z, b) - R(z, a)]/[R(z, a) + R(z, b)]$$

The measure of rival similarity has values in the range of -1 and 1. If $R(z, a) = R(z, b)$, then $F(z, b|a) = 0$.

2 Compactness and informativeness

Practically all recognition algorithms are based on the application of the compactness hypothesis [2]. Patterns are called simple or compact if the boundaries between them are "not very unconventional". We would like to obtain quantitative measure of compactness, and its value should be directly concerned with an expected reliability of recognition.

Each i-th object, $i = 1, 2, \ldots, M_A$ brings contribution F_i to compactness of the pattern A. This contribution depends on similarity to it the other objects of the pattern A, and differences from it the objects of the competing pattern B. The value F_i is defined as follows.

Step 1. For an arbitrary object j, $j \neq i$, of the pattern A we find the distance $R(j, i) = r_1$ to the object i and the distance $R(j, b) = r_2$ to the nearest object b of the pattern B (see Figure 1). Then we find the value of the similarity according to these distances: $F(j, i|b) = (r2-r1)/(r1+r2)$. The value $F(j, i|b)$ should be added to the counter C_i^1.

Step 2. Step 1 is repeated for all objects of the pattern A. Resulting value of the counter C_i^1 represents the sum of similarities of the all objects of A with object i. If we divide this sum by M_A, and we obtain the estimate F_i^1 of "defensive" capability of object i: $F_i^1 = C_i^1/M_A$.

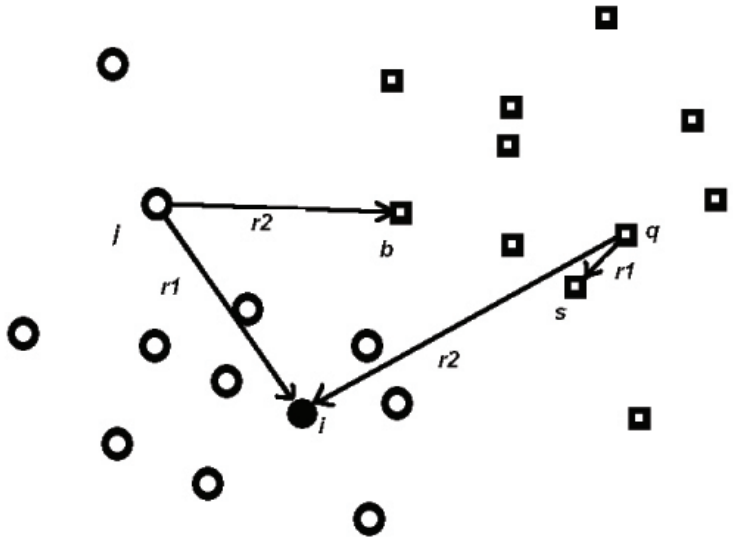


Figure 1: The definition of contribution F_i.

Step 3. Now we need to check the tolerance F_i^2 of the object i relative to the objects of the pattern B. For this purpose we estimate the similarity of all objects q, $q = 1, \ldots, M_B$, of the pattern B with the object i in competition with the object s, the nearest neighbor of the object q, and add this value to the counter C_i^2.

Step 4. Step 3 is repeated for all objects of the pattern B. Then we obtain the estimate F_i^2 of tolerance of the object i relative to the objects of the patter B: $F_i^2 = C_i^2/M_B$.

Step 5. The contribution F_i of the object i to the compactness of the pattern A is $F_i = (F_i^1 + F_i^2)/2$.

Step 6. Repeating Steps 1-5 we obtain the estimates of the contribution for all objects of the pattern A. Summing F_i and dividing their sum by M_A, we obtain the mean value G_A, which characterizes the compactness of the pattern A:

$$G_A = (1/M_A) \sum_{i=1}^{M_A} F_i.$$

Step 7. The same way we obtain values G_B of the pattern B.

This way we can estimate the compactness G_j, $j = 1, \ldots, K$ for all K patterns in the given feature space and, therefore, informativeness of this space can be obtained by using geometrical averaging of estimates of G_j.

$$G = \sqrt[K]{\prod_{j=1}^{K} G_j}.$$

The value of compactness increases with the increasing of the density of objects within patterns and the distance between patterns. The measure suggested by Fisher for estimating informativeness of features has the same property. It is rather natural to use the compactness as criterion of informativeness of the feature space. Experiments with this criterion demonstrated its great advantage over the widely used minimum error criterion when some test dataset is recognized in a mode Cross-Validation.

3 Algorithm FRiS-GRAD

To select most informative attribute subsystem we use main ideas of basic greedy approaches (forward and backward searches) [8]. Forward selection, which is used in the algorithm Addition (Ad), tries to increase attribute subsystem quality as much as possible for each inclusion of attributes, and backward elimination, which is used in the algorithm Deletion (Del), tries to achieve this for each deletion of attributes. In the algorithm AdDel the next combination of these two approaches is used: at first, n_1 informative attributes are selected by method Ad. Then n_2 worst of them ($n_2 < n_1$) are eliminated by method Del. Such consecution of actions (Ad and Del) repeats until the quality of selected attributes is maximum. On first steps while the number of attributes increases the quality increases too. But at some moment when all informative attributes have already taken in the selected subsystem, the quality becomes decreasing. Inflection on the curve of the quality allows specifying the optimum number of attributes. This is very important property of the algorithm AdDel.

In the algorithm GRAD [14] we add and eliminate "granules" (sets consisting of several attributes) instead of single attributes. To find the best subsystem of attributes the algorithm FRiS-GRAD uses procedure of directed search, offered in the algorithm AdDel. On each step

some variant of attribute subsystem is formed and then the algorithm FRIS-Compactness is run to estimate the informativeness of this variant. As a result the algorithm FRiS-GRAD [13] selects the most informative attribute subspace, which dimension is determined automatically. For this subspace the algorithm FRiS-Stolp constructs the rule for classifying of new objects in this subspace.

4 The solution of the nine medical tasks

From many variants of the decision of the nine tasks listing in introduction, for comparison a variant was chosen in which the training set was done on a randomly chosen 50% of the objects. In the control, the other 50% were shown. We have found that ten repetitions of the decision of each task are not enough for a steady estimate of the results. We did a 30-fold recurrence of the random division of the sample between training and control. The quantity of selected attributes n was not set in advance, as our method defines the optimal value for n automatically. For different tasks, subsets of informative attributes in range from 10 up to 25 have been chosen. For this reason for comparison of our method with 10 methods the results, received in [6] with use of 20 best attributes were used.

Each of the nine tasks in [6] was solved by the ten methods for feature selection in a combination with four types of decision rules. We chose the best (record results) from them, and our results were compared with these results. The results of our experiments are presented in Table 1.

Table 1: Results of experiments

Tasks	ALL	Leuk	Prost	DLBCL	Colon	ALL4	Myel	ALL3	ALL2
N	12625	7129	12625	7129	2000	12625	12625	12625	12625
m_1/m_2	95/33	47/25	50/53	58/19	22/40	26/67	36/137	65/35	24/91
Record	**90.00**	95.85	90.19	**93.00**	88.6	82.06	79.4	59.58	78.23
	0.823	0.678	0.565	**0.518**	0.402	**0.278**	**0.062**	0.031	**0.02**
FRiS	**90.00**	**90.00**	**93.13**	89.79	**89.52**	**83.87**	**81.50**	**73.82**	**80.75**
	0.823	**0.704**	**0.647**	0.342	**0.466**	0.225	0.047	**0.112**	0.01
Rating	1	1	1	4	1	1	1	1	1

Here N is the quantity of attributes in the initial dataset, $m1$ and m_2 are the numbers of objects in the first and second classes, respectively. In the line 'Records' the best of all results from [6] are presented, and in the line 'FRiS' the results received by a FRiS-method are presented. The first subrow reflects the percentage accuracy scores, the second, the RCI scores [11]. The place of FRiS-GRAD among the competitors according to their accuracy is specified in the line 'Rating'. Average value of accuracy of the 'Record' algorithms over nine tasks is 85.21, for the

FRiS-GRAD algorithm is $88,04$. Average value of RCI score of the 'Record' algorithms is 0.3752, for the FRiS-GRAD algorithm is $0,3751$.

For the given tasks, for each method it is possible to choose the best result in combination with all four decision rules. By these results it is possible to define the order of the methods. By combining the ranks of a method in each of the nine tasks, one obtains the general rating of that method in this competition: the lower the sum of the ranks, the higher the rating. The results of this calculation are shown in Table 2.

Table 2: Ratings of methods

Feature selection method	**Rating**
Fold change	47
Between group analysis	43
Analysis of variance (ANOVA)	43
Significance analysis of microarrays	42
Rank products	42
Welch t-statistic	39
Template matching	38
Area under the ROC curve	37
AmaxT	37
Empirical Bayes t-statistic	32
FRiS-GRAD	12

For each of the four decision rules used in combination with ten methods of feature selection we find out the best results (the accuracy in percentages) on all nine tasks. Taking in consideration the results received by method FRiS-GRAD as well, on each task we define a place (rank) occupied by each of the five methods. For the first place we add the penalty 1, for last, i.e., the fifth place, the penalty 5. As a result it is possible to obtain ratings of five decision rules (see Table 3):

Table 3: Ratings of methods

Decision rules	**Rating**
Between group analysis (BGA)	35
K-nearest neighbours (kNN)	32
Naive bayes classification (NBC)	25
Support vector machines (SVM)	19
FRiS-GRAD	12

From these results it is seen that the method of feature selection and the decision rules construction based on FRiS-function has high competitive qualities.

The computing time of the algorithm is equal to $O((N + n^3/6)M^3)$, where N is the initial number of attributes, $n << N$ is the number of attributes of which granules are composed, and M is the number of objects of the training sample. The computing time depends only slightly on the initial number of attributes N and increases rapidly with an increasing number of training objects M. In the considered tasks, M was insignificant and for constructing of a decision rules and 30 cycles of CV the computer Pentium has spent on a task "Colon" 40 seconds and on a task "Myeloma" 10 minutes.

5 Conclusions

The advantage of FRiS-approach in comparison to ten methods and four decision rules can be explained by the several reasons.

1. All 10 methods consider the attributes to be independent. To select such informative combinations of attributes, it is necessary to analyse not separate attributes, but their combinations, which demands the decision of difficult combinatory problem. Unlike FRiS-GRAD, the methods presented in [6] are not oriented to solve such kind problem.

2. As for almost all methods of recognition, these ten methods for estimation of the informativeness of attributes use the quantity of the errors received in the process of Cross-Validation (CV). In Section 3 it was noted, that the criterion of informativeness, based on CV, is essentially inferior to a criterion such as that of FRiS-Compactness.

The main reason for the specified advantages consists in the transition from an absolute measure of similarity to the relative measure considering a competitive situation in immediate proximity to objects which with the similarity is to be estimated. Transition from absolutism to relativism allows of improving essentially all algorithms of data mining.

Acknowledgements

The given work was supported by the Russian Fund of Basic Researches (grant № 11-01-00156).

References

[1] Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA*. Vol. **96(12)**, pp. 6745-6750.

[2] Braverman .. (1962). Experiences n rining the Machine to Rgnitin of Visual Patterns. *Automatics and lmhnis* . Vol. **23(3)**, pp. 349-365.

[3] Chiaretti S., Li. X., Gentleman R., Vitale A., Vignetti M., Mandelli F., Ritz J., Foa R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*. Vol. **[9]3(7)** , pp. 2771-2778.

[4] Culhane A.C., Perriere G., Considine E.C., Cotter T.G., Higgins D.G. (2002). Between-group analysis of microarray data. *Bioinformatics* . Vol. **18(12)**, pp. 1600-1608.

[5] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. Vol. **286(5439)** , pp. 531-537.

[6] Jeffery I., Higgins D., Culhane A. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* . Vol. **7**, p.359.

[7] Michotte Y., Kaufman L. (1988). The K-nearest neighbor method. *Data Handling in Science and Technology*. Vol. **2**, pp. 395-397.

[8] Pudil P., Novicova J., and Kittler J.. (1994). Floating search methods in feature selection. *Pattern Recognit. Lett*. Vol. **15(11)**, pp. 1119-1125.

[9] Robertson S.E., Sparck-Jones K. (1976). Relevance weighting of search terms. *J Am Soc Inf Sci*. Vol. **27** , pp. 129-146.

[10] Shipp M.A., Ross'K.N., Tamayo P., Weng A.P., Kutok J.L., Aguiar R.C., Gaasenbeek M., Angelo M., Reich M., Pinkus G.S., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med*. Vol. **8(1)** , pp. 68-74.

[11] Tsai C.A., Chen J.J. (2004). Significance analysis of ROC indices for comparing diagnostic markers: applications to gene microarray data. *J Biopharm Stat* . Vol. **14(4)** , pp. 985-993.

[12] Vapnik V.N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

[13] Zagoruiko N.G., Borisova I.A., Dyubanov V.V., Kutnenko O.A. (2008). Methods of Recognition Based on the Function of Rival Simi1arity. *Int. J. Pattern Recognition and Image Anlsis* . Vol. **18 (1)**, pp. 1-6.

[14] Zagoruiko N.G., Kutnenko O.A., Ptitsin A.A. (2005). Algorithm GRAD for Selection of Informative Genetic Features. *Proceedings of the International Conference on Computational Molecular Biology* . pp. 81-89.

Non-Parametric Stochastic Approximation in Adaptive Systems Theory

A.V. Medvedev
Siberian state aerospace university named after M.F. Reshetnev
Krasnoyarsk, Russia
e-mail: saor_medvedev@sibsau.ru

Abstract

The paper discusses the problem of parametric and non-parametric stochastic approximations according to the experimental information. In this respect the identification problem in a "wide" sense and working sample generation from the initial training one are considered. Some modifications of the known non-parametric estimations of the regression curve according to the observations are introduced; their use in adaptive identification and control problems in conditions of non-parametric uncertainty is analyzed. The results of numerical study are presented.

Introduction

The identification problem is one of the most important in the control theory and other fields united by a capacious term – cybernetics. The main attention is paid to the identification problem in a "wide" sense along with a well-developed identification theory in a "narrow" sense. Earlier [1, 2] discrete-continuous processes and identification ways of stochastic systems were described. They are closely connected with the existed a priory information.

Moreover we are interested first of all in identification in conditions of non-parametric uncertainty as well as in a case when a priori information about the investigated process corresponds simultaneously to the either non-parametric [3] or parametric processes.

1 Levels of a priory information

The availability of this or that volume of a priori information about the investigated process determines a mathematical statement of the identification and control problems and in its turn it assumes the approach to the problem solution. Let's give the types of a priori information and corresponding control systems:

- *Systems with parametric uncertainty.* The parametric level of a prior information assumes a parametric structure of the model and some characteristics of random noise. Zero mathematical expectation and bounded dispersion are typical for them. To estimate parameters different iterated probability procedures are usually used. In these conditions the problem of identification in narrow sense [3] is solved as in all previous cases;

- *Systems with non-parametric uncertainty.* Non-parametric level of a prior information doesn't assume the model existence but it needs some information of the qualitative character about the process, e.g. uniqueness or a lack of uniqueness of its characteristics, linearity for dynamic processes or the character of its non-linearity. To solve the identification problem at this level of a priori information (identification in a wide sense) methods of non-parametric statistics are used;
- *Systems with parametric or non-parametric uncertainty.* The identification problems of multiply connected systems in conditions when quantity of the initial information doesn't correspond to any type mentioned above. For example, one can derive parametric correlations for the particular features of the multiply connected process on the basis of physicochemical regularities, energy, mass conservation law, balance correlation etc., but not for others. So, we have the situation when the identification and control problem is stated in conditions of either parametric or non-parametric a priori information.

2 Statement of the problem

Having passed a standard scheme illustrating the identification problem we shall give a general one. In fig. 1 the following designations are taken: $x(t)$ – vector output variable of the process,

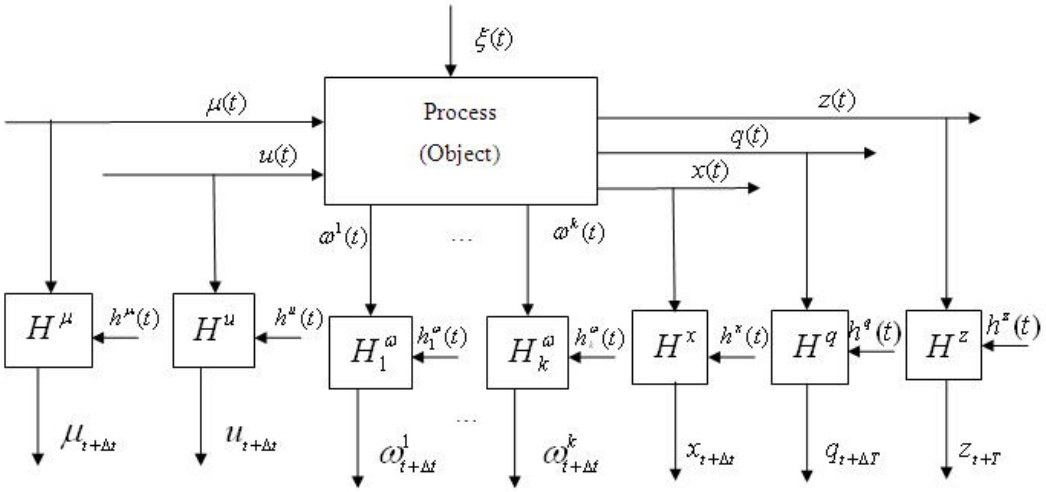


Figure 1: General scheme of the multivariate stochastic process

$u(t)$ – vector control influence, $\mu(t)$ – vector input variable of the process, controlled, but not controlling, $\xi(t)$ – vector stochastic influence, t – continuous time, H^{μ}, H^{u}, H^{E}, H^{ω} – relation channels, corresponding to different variables included control means, devices to measure the observed variables, μ_t, u_t, x_t – denote a measure $\mu(t)$,$u(t)$,$x(t)$ at discrete time. $\omega^i(t) : i = 1, 2, ..., k$ – variables of the process controlled as well according to the object length.

Let's pay attention to the significant difference of output variables $z(t)$, $q(t)$ and $x(t)$ presented in fig. 1. The output variable $x(t)$ is controlled by time intervals Δt, $q(t)$ are controlled by essentially larger time intervals ΔT, z by T ($T >> \Delta T >> \Delta t$). From a practical point of view variable control $z(t)$ often is of the most importance for the investigated process. For example output variables $x(t)$ are controlled with various inductive, capacitate and other data units, $q(t)$ on the basis of laboratory analysis, and $z(t)$ – as a result of prolonged chemical analysis, physical and mechanical testing and so on. It stipulates a great difference the control discreteness of output variables $x(t)$ nd $z(t)$. The peculiarity is the measured output object variables will be known only in certain time intervals. It explains delay in measurements of the output object variables $x(t)$, $q(t)$ and $z(t)$, Δt, ΔT and T – discreteness the changes occur.

In this case output variables depend on input variables in the following way:

$$E(t) = (u(t-\tau), \mu(t-\tau), \xi(t), t), \tag{1}$$

where τ – delay at different process channels, but we do not give any indexes due to the simplicity.

The detailed enough analysis of the similar process was carried out in [1, 2]. The concrete identification problems will be given below pay attention to the differences in each examined case. It is evident from fig. 1 that output object variables $x(t)$, $q(t)$, $z(t)$ of the object depend on input $u(t)$, $\mu(t)$, $\xi(t)$. The received $\omega(t)$ present additional information about the investigated process passing. This information is purposeful to be used at the model building. So identification problem is to build models that can be done generally as follows:

$$\hat{x}(t) = \hat{A}(u(t-\tau), \mu(t-\tau), \omega(t-\tau)), \tag{2}$$

$$\hat{q}(t) = \hat{A}(u(t-\tau), \mu(t-\tau), \omega(t-\tau), \hat{x}(t)), \tag{3}$$

$$\hat{z}(t) = \hat{A}(u(t-\tau), \mu(t-\tau), \omega(t-\tau), \hat{x}(t), \hat{q}(t)). \tag{4}$$

The variety of identification problems could be stipulated by different volume of a priori information, types of processes, and delay in the object and connection channel.

3 *H*-models of non-inertia objects

Let's make the following assumption: an object presented in fig. 1, a non-inertia one, $\omega(t) = (\omega^{'}(t), ..., \omega^{k}(t))$, $q(t)$, $z(t)$ are absent with their measure channels. In this case the object is described by dependency

$$E(t) = (u(t), \mu(t), \xi(t), t). \tag{5}$$

The example demonstrated a model of such an object with delay is as follows:

$$\hat{x}(t) = \hat{A}(u(t-\tau), \mu(t-\tau), \alpha), \tag{6}$$

where $\hat{A}$ a chosen function class, α – a parameter vector, τ – delay. Delay in different connection channels will be surely different, but we shall not mark it in formulas not to make them too complex with indexes. Let's note that all the delays in different connection channels are known.

Let's consider one more process that occurs in practice [7]. The essence of the process is that it has a tube form due to the practically always known stochastic dependency. Let $u \in R^1$, $\mu \in R^1$, $x \in R^1$ (see fig. 2). The intervals of changes $(u, \mu, x) \in R^3$ are always known from practical point of view. Without losing generality let's point out in R^3 a unit cube. The really proceeding process belongs to the subfield $\Omega^H(u, \mu, x) \subset \Omega(u, \mu, x)$, which is always unknown. So, $u \in [0; 1]$, $\mu \in [0; 1]$, $x \in [0; 1]$ and a triad is $(u, \mu, x) \in \Omega^H(u, \mu, x)$. It is clear that not every

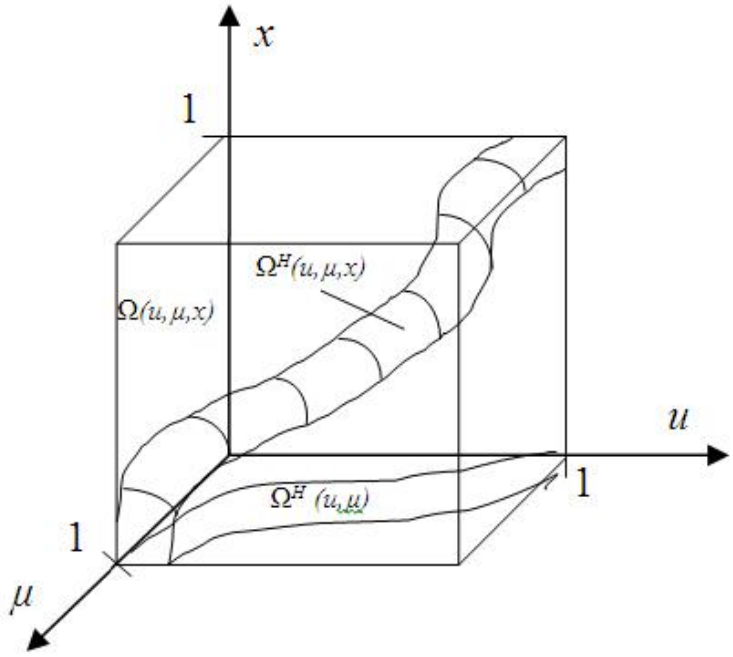


Figure 2: Process proceeding in a tube

of a triad (u, μ, x), get while the experiment or estimated in a real process will belong to the unit cube $\Omega(u, \mu, x)$. It is necessary to note that in the identification theory domains $\Omega(u, \mu, x)$, $\Omega(u, \mu)$, $\Omega(u)$, $\Omega(\mu)$, $\Omega(x)$ are always known, and a domain $\Omega^H(u, \mu, x)$ is always unknown. In the case of stochastic independency of input variables of the process, $\Omega^H(u, \mu, x)$ coincide with $\Omega(u, \mu, x)$, i.e. $\Omega^H(u, \mu, x) = \Omega(u, \mu, x)$. If an object is dynamic, the variables of the phase space are stochastically dependent for sure. It of course leaves its mark on the peculiarity of such processes modeling. We will take them further.

A parametric model of the static process is presented in fig. 1 and it could be done as follows [5]:

$$\hat{x}(u, \mu) = F(u, \mu, \alpha), \tag{7}$$

where $F(\cdot)$ – a function, α – a parameter vector, e.g.

$$\hat{x}(u,\mu) = \sum_{i=1}^{N} \alpha_i \varphi_i(u,\mu), \tag{8}$$

where $\varphi_i, i = \overline{1,N}$ – a system of linearly independent functions $(u,\mu) \in \Omega(u,\mu)$, $(u,\mu,x) \in \Omega(u,\mu,x)$. In the case of stochastic dependency a vector component $u \in R^k$, $\mu \in R^n$ the investigated process has a "tube" structure [7]. Then a parametric model must be taken as follows:

$$\hat{x}(u,\mu) = F(u,\mu,\alpha) I(u,\mu), \tag{9}$$

where $I(u,\mu)$ an indicator, such that

$$I(u,\mu) = \begin{cases} 1,\ if\ (u,\mu) \in \Omega^H(u,\mu); \\ 0,\ if\ (u,\mu) \notin \Omega^H(u,\mu). \end{cases}$$

It's evident that if $\Omega^H(u,\mu) = \Omega(u,\mu)$, then a model (9) coincides with a generally excepted one (7) or with (8). As the estimation of the indicator function $I(u,\mu)$ statistics [7] could be taken.

$$I(u,\mu) = \begin{cases} 1,\ if\ \sum_{i=1}^{s} \prod_{j=1}^{n} \Phi\left(\frac{u_j - u_j^i}{c_s}\right) \prod_{j=1}^{m} \Phi\left(\frac{\mu_i - \mu_j^i}{c_s}\right) > 0; \\ 0,\ if\ \sum_{i=1}^{s} \prod_{j=1}^{n} \Phi\left(\frac{u_j - u_j^i}{c_s}\right) \prod_{j=1}^{m} \Phi\left(\frac{\mu_i - \mu_j^i}{c_s}\right) \leq 0. \end{cases} \tag{10}$$

where bell-shaped functions $\Phi(\cdot)$ and a smooth parameter c_s satisfy convergence conditions [6] $u \in R^n$, $\mu \in R^m$, $x \in \Omega(x) \subset R^l$.

4 *KH*-models of free-inertia objects

Here we will take the same assumptions as before.

$$x_i(u(t),\mu(t)) = F_i(u(t-\tau),\mu(t-\tau)), i = \overline{1,l_1} \tag{11}$$

or more generally,

$$f_i(u^{\langle i\rangle}(t-\tau),\mu^{\langle i\rangle}(t-\tau),x^{\langle i\rangle}(t)) = 0, i = \overline{1,l_2}, \tag{12}$$

Here $F(\cdot)$ and $f(\cdot)$ – unknown functions corresponding to this or that class, an index $\langle i\rangle$ shows that vectors u, μ, x are compound in this case. It means that $u^{\langle i\rangle}$, $\mu^{\langle i\rangle}$, $x^{\langle i\rangle}$, $i = \overline{1,l}$ combined from different component sets of the corresponding vectors $u = (u_1,...,u_n)$, $\mu = (\mu_1,...,\mu_m)$, $x = (x_1,...,x_e)$, i.e. $u^{\langle 1\rangle}$, for example differs also from $u^{\langle 2\rangle}$. In particular we can have the following: $u^{\langle 1\rangle} = (u_1,u_3,u_4,u_7,$ and $u^{\langle 2\rangle} = (u_2,u_3,u_5,u_8,u_9)$. We could say the same about vectors $\mu \in \Omega(x) \subset R^e$ and further analogues designations.

Let the investigated processes have a "tube" structure, then a model (11) will be as follows:

$$\hat{x}_i(u(t),\mu(t)) = \hat{F}_i(u(t-\tau),\mu(t-\tau)) I_s(u(t-\tau),\mu(t-\tau)), \tag{13}$$

where an indicator $I_s(\cdot)$

$$I_s(u(t-\tau),\mu(t-\tau)) = I_s(u(t-\tau),\mu(t-\tau),\vec{u}_s,\vec{\mu}_s) \tag{14}$$

is so that

$$I_s\left(u\left(t-\tau\right),\mu\left(t-\tau\right)\right) = \begin{cases} 1,\ if\ \left(u\left(t-\tau\right),\mu\left(t-\tau\right)\right) \in \Omega_s\left(u,\mu\right); \\ 0,\ if\ \left(u\left(t-\tau\right),\mu\left(t-\tau\right)\right) \notin \Omega_s\left(u,\mu\right). \end{cases} \tag{15}$$

The model of the process (12) will be the following

$$\hat{f}_i\left(u^{\langle i\rangle}\left(t-\tau\right),\mu^{\langle i\rangle}\left(t-\tau\right),x^{\langle i\rangle}\left(t\right)\right) I_s^i\left(u^{\langle i\rangle}\left(t-\tau\right),\mu^{\langle i\rangle}\left(t-\tau\right)\right) = 0, i = \overline{1,l}, \tag{16}$$

where an indicator $I_s(\cdot)$ is

$$I_s\left(u^{\langle i\rangle}\left(t-\tau\right),\mu^{\langle i\rangle}\left(t-\tau\right),x^{\langle i\rangle}\left(t-\tau\right)\right) = \begin{cases} 1,\ if\ \left(u,\mu,x\right) \in \Omega_s\left(u,\mu,x\right); \\ 0,\ if\ \left(u,\mu,x\right) \in \Omega_s\left(u,\mu,x\right). \end{cases} \tag{17}$$

Models of the class (13) – (17) we shall name *KH*-models of free-inertia objects with a delay as they differ fundamentally from the generally excepted models [4, 5] not only because they describe "tube" structure processes, but by a priory information existence about different channels of multivariate object of both parametric and non-parametric type. The model of the similar object could be presented in more details by the following way:

$$\begin{cases} \hat{f}_i\left(u^{\langle i\rangle}\left(t-\tau\right),\mu^{\langle i\rangle}\left(t-\tau\right),x^{\langle i\rangle}\left(t\right),\alpha\right) I_s^i = 0,\ i = \overline{1,k}; \\ \hat{S}_i\left(u^{\langle i\rangle}\left(t-\tau\right),\mu^{\langle i\rangle}\left(t-\tau\right),x^{\langle i\rangle}\left(t\right),\vec{u}_s^{\langle i\rangle},\vec{\mu}_s^{\langle i\rangle},\vec{x}_s^{\langle i\rangle},\alpha\right) I_s^i = 0,\ i = \overline{k+1,l_1}, \end{cases} \tag{18}$$

where α – a parameter vector, $\vec{u}_s^{\langle i\rangle},\vec{\mu}_s^{\langle i\rangle},\vec{x}_s^{\langle i\rangle}$ – time vectors, I_s^i, $i = \overline{1,k}$ – an indicator (15), $\hat{S}_i$, $i = \overline{k+1,l_1}$ – corresponding non-parametric statistics [6].

The equation system (18) is a model of multivariate multilinked free-inertia object which belongs to K-models class which differs fundamentally from known ones [4, 5]. Their difference is that according to some channels of the multivariate process, its parametric structure is known to within parameters α. The second group of equations (18) $\hat{S}_i$, $i = \overline{k+1,l_1}$ corresponds to the level of non-parametric uncertainty. In this case we can have a priori information only of the qualitative character about the investigated object and use stochastic approximation of non-parametric type.

The estimation of parameters α in k-equations (18) can be reduced to the identification problem in a "narrow" sense. The estimation of the second group of equations $\hat{S}_i$, $i = \overline{k+1,l}$ in (18) can be realized by non-parametric methods [6]. If the system of equations (18) is disintegrated, i.e. in the case with the described process (11) we can easily forecast $x(t+\tau)$ by known values $u(t)$and $\mu(t)$. In the opposite case we must solve the system (otherwise, K-model) relative to the vector $x \in \Omega(x) \subset R^e$.

In the case we have one root of the system (18) in $\Omega\left(x\right)$ a method stated in [6] can be used. Generally speaking, this problem needs special study.

5 *K*-models of dynamic objects

We'll take further the problem of the dynamic process model building presented in fig. 1. Let's note that ΔT and T exceed considerably a time constant of the object about all the rest channels. Without losing generality we could consider that the control of variables $u(t)$, $\mu(t)$, $\omega(t)$, $x(t)$ is realized by the time interval $\Delta t << \Delta T << T$. Hence the process about channels $q(t)$ and $z(t)$ belongs to the class of free-inertia with delay, and about channels $\omega(t)$ and $x(t)$ can be referred to the dynamic class as their control is realized through the interval Δt that is considerably less than a time constant of the object about corresponding channels. In this case, a general enough K-model can be taken as follows.

$$\begin{cases} \hat{f}_i\left(u^{\langle i\rangle}(t-\tau), \mu^{\langle i\rangle}(t-\tau), \omega^{\langle i\rangle}(t-\tau), x^{\langle i\rangle}(t), \frac{dx^{\langle i\rangle}(t)}{dt}, \frac{d^2x^{\langle i\rangle}(t)}{dt^2}, ..., \alpha\right) = 0, \\ i = \overline{1,k}; \\ \hat{f}_i\left(u^{\langle i\rangle}(t-\tau), \mu^{\langle i\rangle}(t-\tau), \omega^{\langle i\rangle}(t-\tau), x^{\langle i\rangle}(t), q^{\langle i\rangle}(t), z^{\langle i\rangle}(t), \beta\right) I_s^i = 0, \\ i = \overline{k+1, l_1}; \\ \hat{S}_i\left(u^{\langle i\rangle}(t-\tau), \mu^{\langle i\rangle}(t-\tau), \omega^{\langle i\rangle}(t-\tau), x^{\langle i\rangle}(t), q^{\langle i\rangle}(t), z^{\langle i\rangle}(t), W_s^{\langle i\rangle}\right) = 0, \\ i = \overline{l_1+1, l_2},\ l_2 > l_1 > l, \end{cases} \tag{19}$$

where the first system of equations (19) is found on the basis of known fundamental laws corresponding to the investigated process to within parameters α. The second system of equations of the object is found on the basis of existed a priori information to within vector parameters β. The third group of equations (19) is unknown to within parameters, but class functions characterizing interconnection "input-output" and intermediate variables is determined on the basis of a priori information. The appearing designation $W_s^{\langle i\rangle}$ is a unity of all i-th variable observations of s – volume, i.e.

$$W_s^{\langle i\rangle} = \left(\overrightarrow{u_s^{\langle i\rangle}}, \overrightarrow{\mu_s^{\langle i\rangle}}, \overrightarrow{\omega_s^{\langle i\rangle}}, \overrightarrow{x_s^{\langle i\rangle}}, \overrightarrow{q_s^{\langle i\rangle}}, \overrightarrow{z_s^{\langle i\rangle}}\right),\ i = \overline{l_1+1, l_2}.$$

The estimation of vector component values of output variables $x(t)$, $q(t)$, $z(t)$ can be found in the result of equation system (19) solving at fixed variables $u(t)$, $\mu(t)$, $\omega(t)$. K-models differ very much from generally excepted ones first of all because they take into account in interconnections all existed variables and connections among them in the situation when their discreteness of control differs very much. The levels of a priori information about different cannels of the investigated process are different. So K-models present an integral synthesis describing the investigated process or a system of interconnected objects in their variety.

6 Control of variables, measurements

Here we shall underline the problem's importance of "input-output" variables measurement of the investigated object, process. Earlier [2, 3] it was mentioned that differing control means even for one and the same processes lead to different identification problems formulation. The main we

must point out in this problem is that a dynamic object sometimes has to be regarded as statistic with delay due to the prolong control procedure (measurement, control) of some variables. It exceed significantly a time constant of the object.

Undoubtedly at modeling and control by discrete-continuous processes it is expedient to use signals or analogue ones but it needs a thorough analysis not only of the concrete object but control methods and techniques of all the available variables and a priori information that can correspond simultaneously to different levels of a priori information about different measurement cannels of multivariate object system of variables. These or those variables, parameters, type of the measurement and control, a priori information and also some "freedom" when any assumption inevitable at mathematical statement of the problem that can finally lead to negative circumstances. All these questions are often missed while modeling problem study from theoretical point of view [4, 5]. It is impossible to solve applied problems of the concrete processes models building otherwise "truth doesn't suffer if somebody ignores it" (I.F. Shiller). It is appropriate to pay special attention of the investigator to the identification problem of the real process formulation at the very initial level: "it is much more difficult to find problem than to find a solution. For the first one it is necessary to have imagination, and for the second one - knowledge" (D.D. Bernal).

Let's point out more important feature that accompanies a lot of variables measurement. It is a non-representative sample for control. The problem is that the measurement (analysis) results of these or those variables are given for the whole set of product (articles). In this connection a dozens of grams of the product are taken, but a result is conferred to multi-ton set of goods. We should pay attention the analysis itself is carried out with high accuracy, chemical, physicochemical, physico-mechanical and so on. One more problem is of great importance: where and how must we take samples? In some branches it is regulated by GOST, in others some recommendations are adopted. In a word, this problem is very serious and needs a thorough analysis in each concrete case. Inaccuracies at this level lead to "rough" models of the process, and hence to unsatisfactory control systems. We shall not discuss here the problem of destroying control. It is an additional, independent problem that needs special study.

7 Mathematical problem statement of modeling and control

There is a quite evident fact about different a priori information of the investigated process existence [2]. As its consequence is different mathematical problem statement from the point of view about mathematical strictness. The disparity of our assumptions about the investigated object to the object itself is one of the most important "stumbling-blocks". After traditional phrase "Let a process..." some assumptions, hypothesis with remote relation to the reality follow. It is difficult to imagine a process, object, features whose features were unchangeable or they are changing according to the known law in time. We mean the processes described in [1], methods and technologies of the object's variables measurements that were of a great interest for the theory of automatic control. Their main features are in the lack of a priori information, random factors

influence, unknown features, lack and imperfection of variable control means, non-representative sample for measurements and so on. We have to change the ignorance to our regret speaking "Let...". It is clear that if our assumptions are quite nearby to the reality, then, in the result we can count on the success solving this or that problem. In another case a failure is inevitably. In reality on the basis of numerous processes and objects are fundamental laws of physical, chemical, electric, mechanical events; they could be described with a high level of accuracy. Correspondently special models and control systems of high quality could be built for them.

If assumptions are too "rough", there exist two ways to solve problem. The first way is to complete our "ignorance" about a process when a problem statement is done accurately from mathematical point of view. The second way is to develop a mathematical approach to adequate level of a priori information we have really.

Thereupon we'd like to remind you some know facts from R. Kalman's paper [8]. Let's quote some extracts from this paper: "... classical (Kholmogorov's) probabilistic approach can't operate in real problems with uncertain information. To model uncertainty by probabilistic method it is necessary to have excessively much information unextracted from available data of numerous practical problems." and more:

L.S. Pontryagin: "Mathematicians don't believe in probability".

A.N. Kholmogorov: "...something is wrong with statistics".

A bit different axiomatics of probabilistic theory is sated in [9].

In future we will model and control real including organizational processes described in [1] because reality and practice need it. In particular we could refer a lot of economical processes to organizational ones. As far back as in the middle of the previous century G. von Neumann and O. Morgenstern wrote about applying mathematics in economics [10]: "First of all let's be aware that there is no universal system in economic theory and it hardly be formulated in the nearest future. The reason is that economics is a very complex science...". And further "Often arguments against applying mathematics in economics include only references to subjective elements and psychological factors and so on...". "It is important to realize that economists can't expect easier fate than scientists of other fields have".

A century has passed but mathematics hasn't appeared for economics as well as modeling and operation process control though we've some steps in this direction: smooth sets theory, decision making theory, system analysis and theory of systems etc.

8 Avalanche processes

During the recent decade the interest to the processes step-wise changing at smooth influence from the outside has increased substantially. Hence a new division of mathematics with an intrigue name catastrophe theory has appeared. Though the source of this theory goes back to the XVII century, but the first information about the catastrophe theory appeared in 1970 in professor R. Tom works. The basic role in foundation and development of the catastrophe theory had the preceding investigations of A. Poincare and A.A. Andronov.

Nowadays it is a powerful mathematical technique with a wide sphere of use in natural science and technique. In a remarkable book of V.I. Arnold [12] together with fundamentals of catastrophe theory for a reader who is not a mathematician some facts about the years of perestroika started at the end of 1980s in our country are given.

Let's go back to the scheme presented in fig. 1. A numerical experiment when a process is under the influence of smoothly changing variables $u(t) = (u_1(t), u_2(t))$, $\mu(t) = (\mu_1(t), \mu_2(t))$, $\xi(t) = (\xi_1(t), \xi_2(t))$ was carried out. The results of the experiment are given in figures above.

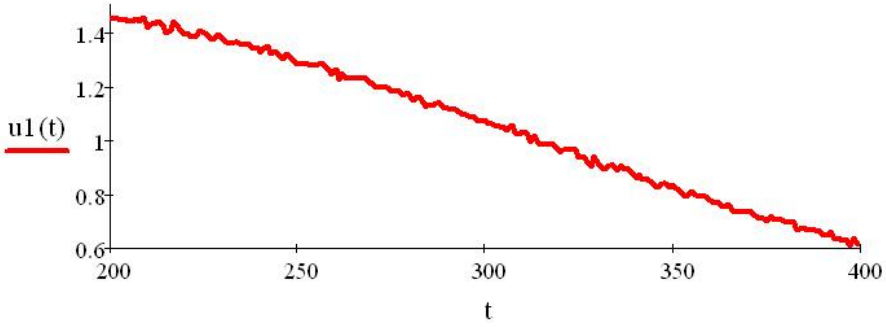


Figure 3

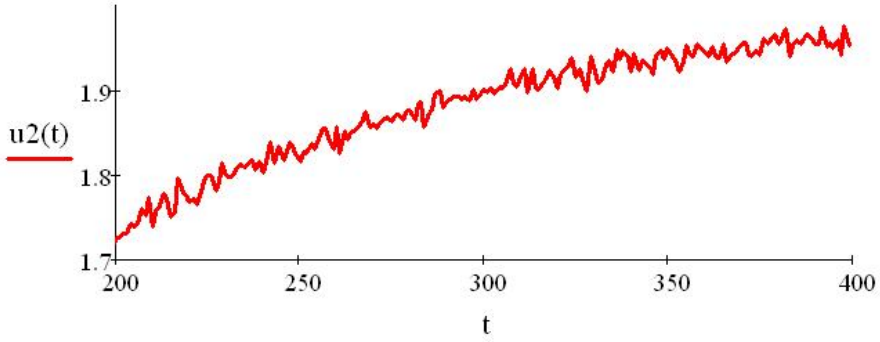


Figure 4

Their analysis shows that with a smooth changing of $u(t)$, $\mu(t)$ and $\xi(t)$ the response of the system (input variable value) $x(t)$ is smoothly changing as well. But in some time the formation of avalanche process of increasing $x(t)$ that is transferring into a catastrophe. In other words the loss of stability occurs and – "explosion". Such types of processes have been a subject of investigation for many years in physics, technique.

These processes are of a great interest in organizational systems. They are systems in society – social, economic and other spheres.

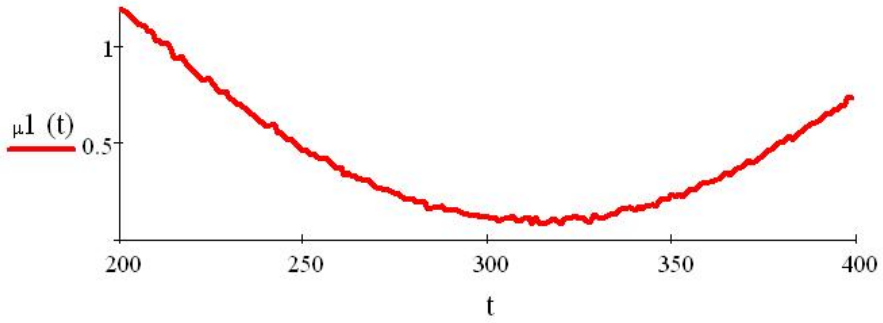


Figure 5

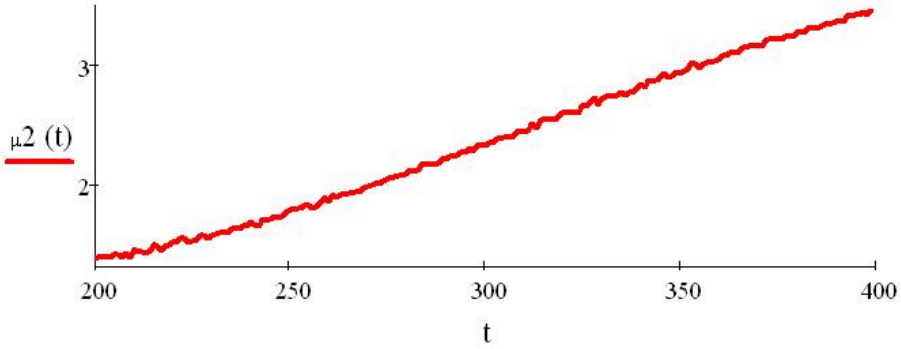


Figure 6

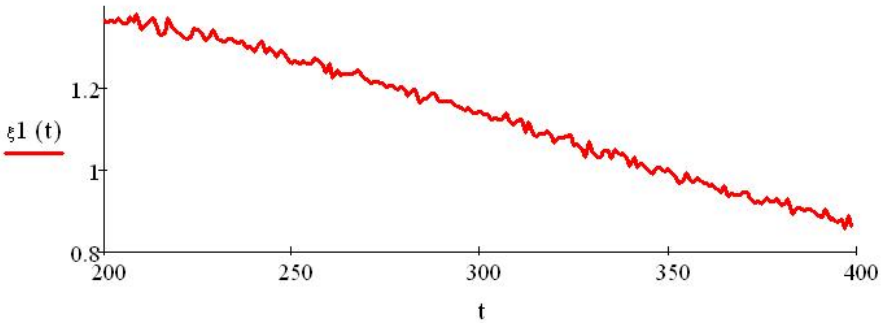


Figure 7

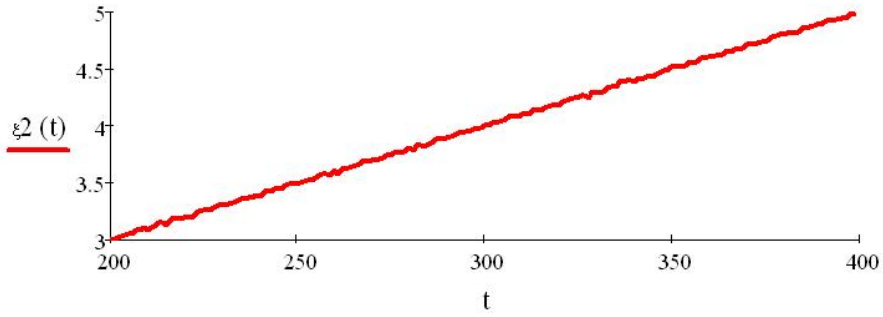


Figure 8

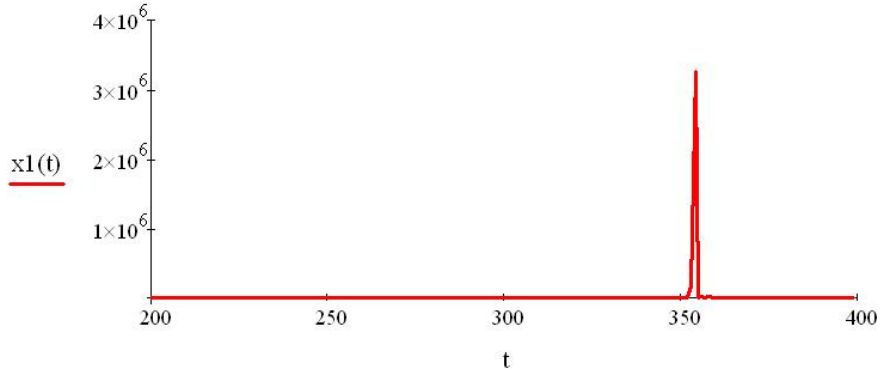


Figure 9

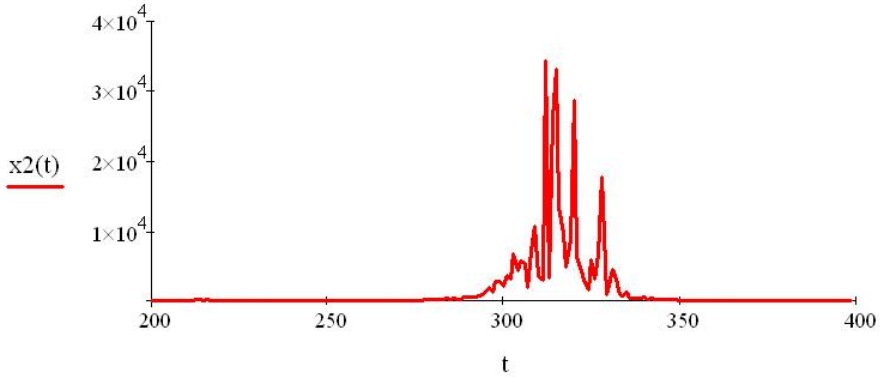


Figure 10

These processes study might lead to forecasting an incipient "inside an avalanche" and as a consequence, we have counteraction to the latest control methods, feedback and taking into account reality. K-models, KH-models and their system formations are of the most perspective.

9 Control system by the discrete-uninterrupted process

The control system by the discrete-uninterrupted object is presented generally in the underlying figure [11].

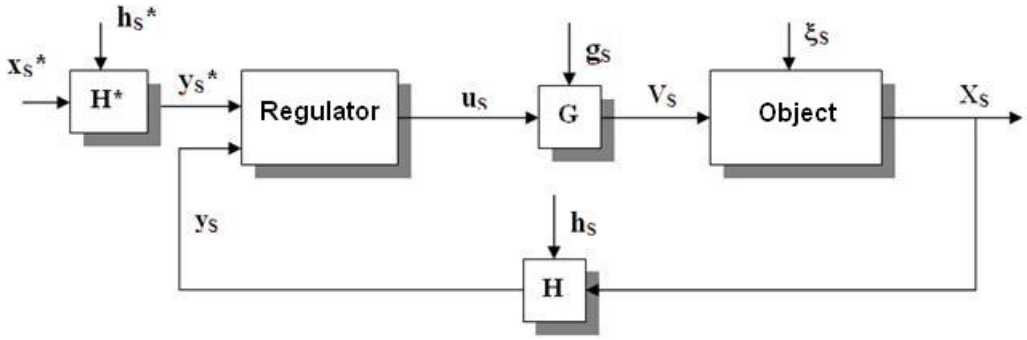


Figure 11

The following designations are accepted: x_s^* – a control input, mixing with the noise h_s^* through the channel H^* enters the regulator as y_s^*; the output of the object x_s coming through the channel H and mixing with the noise h_s enters y_s the regulator as u_s; the control effect u_s coming through the channel G, and mixing with the noise disturbance g_s enters to the object which is under the effect of ξ_s as v_s.

In the theory of dual control [11] and in the theory of adaptive systems [5] one supposes mathematical description of the object with the exactness to within the parameters vector. In most cases it is not enough to have only a priori information to choose the control of the investigated process well-grounded. That's why we should conduct a number of experiments on the object (often long and expensive) to solve the identification problem qualitatively from practical point of view. The more detailed analysis of a priori information levels is given in [4].

10 Non-parametric dual control

In conditions of non-parametric uncertainty [6] the equation of the process with the exactness to within the parameters vector is unknown but we know object's features of the qualitative character, e.g. characteristics uniqueness or a lack of uniqueness for non-inertia processes; linearity or a type of non-linearity for dynamic processes. If a type of the equation describing a process

is unknown then the known parametric methods of the control theory can't be applied for the identification and control problems solution.

Let the process is described by the equation:

$$x(t) = A < u(t) > \tag{20}$$

where A – an unknown operator, describing the process. If there exists an operator inverse to , i.e. $^{-1}$, $AA^{-1} = I$ – a single operator, then

$$u(t) = u^*(t) = A^{-1} < x^*(t) > \tag{21}$$

Having defined a path $x^*(t)$, let's find an ideal value $u^*(t)$ from (21). So (21) could be referred to the category of ideal regulators. Further we'll call it u-regulator to distinguish it from already known regulators. But the problem is in impossibility to build it in most cases; moreover the operator A is unknown. The attempt to solve this problem b at least partially by introduction control systems of correcting chains, compensating links etc. was undertaken. In some technical systems it led to the success.

In 50s years of the previous century academician V.S. Kulebakin offered and developed substantially a method of K(D)-images. It led to the theory of automatically regulated and controlled systems invariance. But in that case it is necessary to have high exactness of the investigated processes description by differential equations. If this kind of equation estimating the investigated process is unknown, classical methods of control theory can't be applied.

We take a particular case. Let an object is described by a linear differential equation of the unknown order, e.g. n, n – unknown. In this case at zero initial conditions $x(t)$ is:

$$x(t) = \int_0^t h(t-\tau)u(\tau)d\tau \tag{22}$$

where $h(t-\tau)$ – a weight function of the system, which is a derivative of the transient function $k(t)$, i.e. $h(t) = k\prime(t)$. It is known that the inverse operator (22) is an operator:

$$u(t) = \int_0^t v(t-\tau)x(\tau)d\tau \tag{23}$$

where $v(t)$ – a weight function of the object in the direction "input-output" and $v(t) = w\prime(t)$, where $w(t)$ – a transient function of the system in the same direction. In this case A is presented by the operator (22), and $^{-1}$ – by the expression (23). Hence, the problem is in finding a weight function $h(t)$. One of the possible ways for this problem solution is to solve the Wiener-Hopf equation. The second way is in taking a transient characteristic on a real object with a further estimation of its weight function according to the measurement results $\{x_i = k_i, t_i, i = \overline{1,s}\}$.

The non-parametric model is as follows:

$$x_s(t) = \int_0^t h_s(t-\tau, \vec{k}_s, \vec{t}_s)u(\tau)d\tau \tag{24}$$

where $\vec{k}_s$, $\vec{t}_s$ – time vectors $\vec{k}_s = (k_1, \ldots, k_s)$, $\vec{t}_s = (t_1, \ldots, t_s)$, and $h_s(\cdot)$ is:

$$h_s(t) = \frac{1}{sA_s} \sum_{i=1}^{s} k_i H' \left(\frac{t - t_i}{A_s} \right) \tag{25}$$

$H(\cdot)$ – bell-shaped (kernel) functions, c_s - a spreading parameter, satisfying conditions of convergence.

It is not possible to "take off" a weight function $v(t)$ in the direction "input-output" as well as a transient function $w(t)$ on the object. It was offered to take off a transient function "backwards" on the model. Apparently it was done for the first time in [6]. So, from the correlation:

$$x_s(t) = 1(t) = \int_0^t h_s(t - \tau, \vec{k}_s, \vec{t}_s) u(\tau) d\tau \tag{26}$$

where $1(t)$ – Heaviside function; it is possible to receive samples $\{u_j, t_j, j = \overline{1, s}\}$. Then the non-parametric control algorithm by a linear dynamic system is as follows:

$$u_s^*(t) = \int_0^t \left(\frac{1}{sc_s} \sum_{j=1}^{s} w_j H' \left(\frac{t - \tau - t_i}{c_s} \right) \right) x^*(\tau) d\tau \tag{27}$$

It's evident a sample number while taking transient characteristics on a real object and a model couldn't be coincided.

As operators A and A^{-1} according to the real data will be estimated not exactly, the control system could be represented as follows: In figure 12: CU – control unit, A_s^{-1} – non-parametric

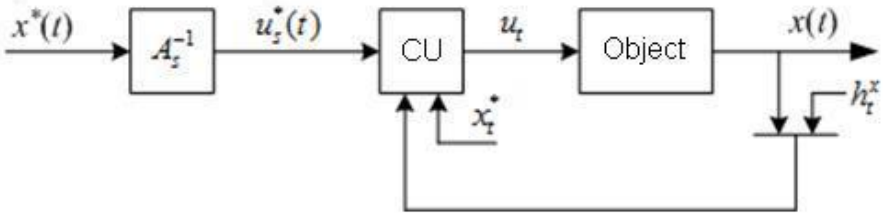


Figure 12

estimation of the reverse operator of the object, u_s^* – output (estimation A^{-1}, the noise disturbance h_t^x operates in the channel of the feedback. Non-parametric algorithm of dual control is as follows:

$$u_{s+1,t} = u_{s,t}^* + \Delta u_{s+1,t} \tag{28}$$

Here $u_{s,t}^*$ is determined according to the formula (27), and $\Delta u_{s+1,\mathrm{t}} = \varepsilon(x_t^* - x_t, s)$ – search steps. So, in $u_{s,t}^*$ "the information" about the object and in $\Delta u_{s+1,t}$ – "studied" search steps are concentrated. That is the essence of the algorithm (28) dualism.

Let's clarify it as an example the non-inertia object $x = f(u)$. For its estimation we'll take non-parametric estimation of the regression function according to the observations $\{x_i, u_i, i = \overline{1,s}\}$ [4].

$$x_s(u) = \sum_{i=1}^{s} x_i \Phi\left(\frac{u-u_i}{c_s}\right) \Bigg/ \sum_{i=1}^{s} \Phi\left(\frac{u-u_i}{c_s}\right), \tag{29}$$

where a bell-shaped functions $\Phi(\cdot)$ and a spreading parameter c_s satisfy some convergence conditions [4]. In this case $u = f^{-1}(x)$ will be the analogue of the expression (27), and u_s^* from (28) will be equal to:

$$u_{s,t}^* = \sum_{i=1}^{s} u_i \Phi\left(\frac{x_{s+1}^* - x_i}{c_s}\right) \Bigg/ \sum_{i=1}^{s} \Phi\left(\frac{x_{s+1}^* - x_i}{c_s}\right) \tag{30}$$

Let's analyze the dualism feature of the algorithm (28). At the initial control stage the second summand $\Delta u_{s+1,t}$ of the formula (28) is of the most importance. It is the case of active information storage in the system of dual control. It begins with the first observation of input and output variables of the object. During the training process (the information storage) the first summand, i.e. $u_{s+1,t}$ begins to play an increasing role the control influence formation $u_{s,t}^*$.

11 Numerical experiments

We shall represent below certain results of calculations illustrating particular results while investigating dynamic and non-inertia objects.

The first experiment illustrates "taking" of the weight functions of the linear dynamic system. The following figure shows a weight function received by solving differential equation of the type when a delta-function is accepted. Here we will represent weight functions for different analogues of delta-functions.

Weight functions received with various d-shaped step-functions of the following type as numerical investigations have shown are practically coincided:

$$\delta_1(t) = \begin{cases} \Delta t^{-1}, & \Delta t \in [0; 0.1] \\ 0, & \Delta t \notin [0; 0.1] \end{cases} \quad \delta_2(t) = \begin{cases} \Delta t^{-1}, & \Delta t \in [0; 0.01] \\ 0, & \Delta t \notin [0; 0.01] \end{cases}$$

As you can see, a weight function received analytically as well as weight functions received under effect on the object of considerably different d-shaped functions are practically coincided. It leads to the conclusion that in practice at the experiment about weight functions observation it is possible to give considerably different d-shaped controlled effects to he object's input. The connection (correlation) of two elements – "a dynamic object (process) – a differential equation" are of a great importance here.

The results of the linear dynamic object control are shown in figures 13 in the conditions that x_t^* – is a random variable, generated uniformly by distributed random numerals.

The experiment was conducted according to the following plan: at first the transient characteristics were taken; the operator A was estimated according to the formula (26) by these

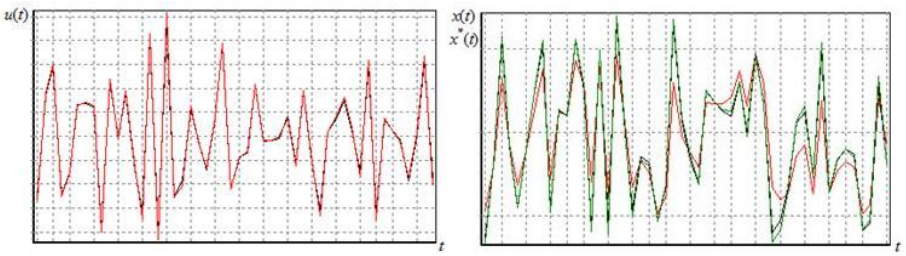


Figure 13

characteristics. The inverse operator A^{-1} was evaluated according to the formula (27). The figure illustrates that good quality of control though in such an "exotic" case. Any known regulator can't manage to solve this problem.

12 About the non-parametric systems theory

The terms "non-parametric identification", "non-parametric methods of data processing" are met in monographs about identification, but non-parametric identification algorithms are not sited. Usually non-parametric identification of linear dynamic processes is related to searching weight or transient functions of the system in the result of integral Fredholm equations of the 1st kind solving, in particular Wiener-Hopf equations.

We were speaking above about the models and u-regulators that are free of choice with the with the exactness to within the parameters vectors of the models of the investigated process or a parametric structure of the control units, as well as we were speaking about parametric structure of other process characteristics, e.g. correlative functions, spectral density and etc. So we are speaking about identification and control in conditions of non-parametric uncertainty [3]. It is conceived that the least level of the a priori information of the investigated object when the solution of a large number of cybernetics problems adequate to real processes is possible. Also let's note that the first investigations about non-parametric control by non-inertia objects belongs to the beginning of 70s of the previous century. One may consider the theory of non-parametric systems covers different cybernetics problems oriented to the non-parametric level of the a priori information.

Conclusion

The stated above information covers some identification and control problems at the level of parametric and non-parametric a priori information. In contrast to the well-developed parametric theory, the non-parametric one is oriented at the less level of the a priori information about the investigated objects and processes. The special attention is paid for the systems of dual control

of the Bayes type, an adaptive dual control and a non-parametric dual control. Some non-parametric models and algorithms of dual control and particular results of numerical calculations of the illustrative character are given.

References

[1] Medvedev A.V. (2010). Theory of non-parametric systems. Processes. *Vestnik SubSAU*. Vol. **3**.

[2] Medvedev A.V. (2010). Theory of non-parametric systems. Modeling. *Vestnik SubSAU*. Vol. **4**.

[3] Medvedev A.V. (2010). To theory of non-parametric systems. *Vestnik SubSAU*. Vol. **5**.

[4] Eickhoff P. (1975). *Fundamentals of control systems identification*. Mir, Moscow.

[5] Zupkin Ja.Z. (1986). *Adaptation and learning in automatic systems*. Nauka, Moscow.

[6] Medvedev A.V. (1983). *Non-parametric systems of adaptation*. Nauka, Novosibirsk.

[7] Medvedev A.V. (1995). Data analysis in identification problem. *Computer data analysis of modeling*. Vol. **2**. BGU, Minsk.

[8] Kalman R.E. (1985). Systems identification with noise. *Success of mathematics sciences*. Vol. **10**, part **4**.

[9] Wittle P. (1982). *Probability*. Nauka, Moscow.

[10] John fon Nueman (1970). *Game theory and economic behavior*. Nauka, Moscow.

[11] Feldbaum A.A. (1963). *Fundamentals of optimal automatic systems theory*. Fismatgis, Moscow.

[12] Arnold V.I. (1990). *Catastrophe theory*. Nauka, Moscow.

About Regression Characteristics Nonparametric Estimation in the Identification Problem

D. Bezmen, L. Golub, A. Medvedev
Siberian State Aerospace University named after Academician MF Reshetnev, Krasnoyarsk, Russia
e-mail: saor_medvedev@sibsau.ru

Abstract

In many applications, the regression function restoration from observational data with random errors is directly related to the identification problem. In this regard, the case where by the "input-output" variables supervision of an object are significant errors such as "crude error" and the reducing the dimension of the input variables vector problem is most attractive. This report focuses on the these issues analysis.

Introduction

During the stochastic objects identification we are often faced with the need to restore the various characteristics of the regression variables from observations with random errors. When given a parametric of regression functions form, the problem reduces to estimating the parameters. This class of problems relates to the identification problem in a "narrow" sense of [1, 2]. If a priori information about the process is insufficient to determine a parametric process model, the identification problem is formulated in the "broad" sense [1] and reduces to the nonparametric estimation of the characteristics regression [3]. Below we consider the problem of extracting the essential variables in the problem identification and the case where by the object "input-output" supervision may occur errors such as "misses".

1 The classical scheme of the identification problem

Give a general process scheme adopted in the modeling and identification theory.

The following designations are accepted: - unknown object operator, $E(t)$ - output variable vector, $u(t)$ - vector control action, $\mu(t)$ - input unguided vector, but controlled variable, $\xi(t)$ - vector random effects, (t) – continuous time, u_t, μ_t, x_t - means the measurement $u(t)$, $\mu(t)$, $x(t)$ in discrete time t. Control variables (x, u, μ) by a time interval Δt, ie $u_i,\ \mu_i,\ x_i\ \ i = \overline{1, s}$ - sample measuring process variables, s - sample size, h^u, h^μ, h^x- with an icon at the top – the process variables random noise measurement.

Object shown in Figure 1, described by an unknown operator A, ie,

$$x(t) = A(u(t-\tau), \mu(t-\tau), \xi(t), t) \tag{1}$$

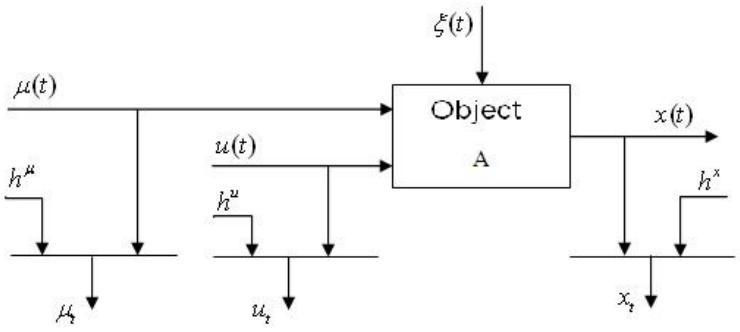


Figure 1: A classical identification problem scheme

where τ - retardation, characterized by various means, let accept the designation uniform for all channels that characterize the process under study by the simplicity reasons. The identification problem is to estimate (1) with available a priori information about the process and measurements with random errors $u(t-\tau)$, $\mu(t-\tau)$, $x(t)$.

The theory of identification in the "narrow" sense is currently dominated by the variety of discrete-continuous processes simulation . Its content consists in the fact that the first stage is somehow determined by the operators A parametric class, for example,

$$\tilde{x}_\alpha(t-\tau) = A^\alpha(u(t-\tau), \mu(t-\tau), t, \alpha), \tag{2}$$

and the second stage, the parameters α estimation based on the available sample $\left\{x_t, u_t, \mu_t, t = \overline{1,s}\right\}$. In this case the identification problem solving success depends on how "successful" the operator (2) is designated.

A "broad" sense identification, implies the parametric class phase selection absence of (1), unless, of course, this is not sufficient a prioriinformation. Operator to define a class of operators (1) on the basis of qualitative character information is often much easier, for example, the linearity of the process or the type of nonlinearity, etc. The identification problem consists in estimating this statement based on a sample $\left\{x_t, u_t, \mu_t, \quad t = \overline{1,s}\right\}$ in the form of

$$\tilde{x}_S(t-\tau) = A_S(u(t-\tau), \mu(t-\tau), t, \vec{x}_S, \vec{u}_S, \vec{\mu}_S), \tag{3}$$

where $\vec{x}_S$, $\vec{u}_S$, $\vec{\mu}_S$ - time vectors, $\vec{x}_S = (x_1, x_2, .., x_s)$, $\vec{u}_S = (u_1, u_2, .., u_s)$, $\vec{\mu}_S = (\mu_1, \mu_2, .., \mu_s)$. The presence of one or another of a priori information about the studied process, in essence, defines the identification problem mathematical formulation, and this, in turn, determines the approach to its solution. Let us dwell on the characteristics a prior information two types:

- *a parametric indeterminacy level.* A priori information parametric level requires a parametric model structure and some random noise characteristics, usually because they are zero mean and bounded variance. For example, to multiply the individual characteristics of the process on the basis of physical and chemical laws, energy, mass conservation, relations balance and

others can be derived parametric relations, but not for others. For estimating the parameters most commonly used variety of iterative probabilistic procedures [1, 2]. Under these conditions, we solve the identification problem in the "narrow" sense;

- *non-parametric indeterminacy level.* A priori information nonparametric level does not imply the model existence , but requires some information about the qualitative nature of the process [4], for example, the uniqueness or ambiguity of its characteristics, the linearity for the dynamicprocesses or the nature of its nonlinearity. To solve the identification problems at this level, a priori information (identification of a "broad" sense) nonparametric statistics methods [3].

2 The essential variables search in the reconstructing a regression function from the observations problem

Let the process under study by one of the channels is described by the regression characteristic form $x = f(u_1, ..., u_n)$, where the probability density $p(x) > 0 \ \forall x \in \Omega(x)$ and $p(y)$ the unknown. Then, for the restoration $f(u_1, ..., u_n) = M\{x|u\}$ of a statistically independent sample s of observations $(\overline{u_1}, x_1), (\overline{u_2}, x_2), ..., (\overline{u_s}, x_s)$, where $\overline{u_i} = (u_1^i, ..., u_n^i)$. To approximate the unknown function in conditions indeterminacy are often used nonparametric regression, nonparametric estimator has the form [3]:

$$x_s(u_1, ..., u_n) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{n} \Phi(c_{sj}^{-1}(u_j - u_j^i))}{\sum_{i=1}^{s} \prod_{j=1}^{n} \Phi(c_{sj}^{-1}(u_j - u_j^i))}, \tag{4}$$

where $\Phi(z_j)$ - a bell-shaped function(the "kernel") that has some convergence properties [4], $z_j = \frac{u_j - u_j^i}{c_{sj}}$, $j = \overline{1, s}$, c_s – fuzziness factor satisfying: $c_s > 0$, s = 1,2,. . . , $\lim\limits_{s\to\infty} c_s = 0$, $\lim\limits_{s\to\infty} sc_{\mathrm{s}}^{\mathrm{n}} = \infty$. In the sequel for simplicity we omit the designation of retardation. Essential in the evaluation of (4) is that with each component of the vector $u = (u_1, u_2, .., u_n)$ component of the vector $c_s = (c_{s1}, ..., c_{sn})$ is placed.

Parameter c_{sj} is adjusted based on the minimum mean-square criterion in a moving the exam mode:

$$W(A_{sj}) = s^{-1} \sum_{i=1}^{s} (x(u_i) - x_s(u_i, c_{sj}))^2 \underset{c_{sj}}{\longrightarrow} \min. \tag{5}$$

Before solving the problem (5), hold the centering and standardization of components of the vector $u = (u_1, u_2, .., u_n)$ on the observations basis $\vec{u}_s = (u_1, u_2, .., u_s)$. As a solving result the problem (5) will be found the components $\widehat{c}_s$, where $j = 1, 2, ..., n$. Having the vector $\widehat{c}_s$ components numerical values, we form a inequalities chain. Suppose that in the private case, it looks like this:

$$\widehat{c}_{s2} < \widehat{c}_{s3} < \widehat{c}_{s1} < ... < \widehat{c}_{sn} < \widehat{c}_{s4} < ... < \widehat{c}_{s5}, \tag{6}$$

where the numbers are marked with the components $\widehat{c}_s$. In this case, the following

Rule: the least impact on the value of $x(u_1, ..., u_n)$ providing the vector u, component which corresponds to the most important components of the parameter vector blur $\widehat{c}_s = (\widehat{c}_{s1}, ..., \widehat{c}_{sn})$ optimal estimate.

In particular, by analyzing the inequalities chain (6), we can conclude that the least impact on the value $x(u_1, ..., u_n)$ of providing component u_5.

Example. Let the solution of the optimization problem (5) ($n = 5$)were found parameter estimates blur: $\widehat{c}_{s1} = 0.05$, $\widehat{c}_{s2} = 0.01$, $\widehat{c}_{s3} = 0.9$, $\widehat{c}_{s4} = 0.4$, $\widehat{c}_{s5} = 0.3$. Form the inequalities chain:

$$\widehat{c}_{s2} < \widehat{c}_{s1} < \widehat{c}_{s5} < \widehat{c}_{s4} < \widehat{c}_{s3}.$$

Variable is excluded from the regression estimates, as the corresponding parameter $\widehat{c}_{s3}$ score has the greatest value. Provided that the absence u_3 does not increase the approximation error (this issue is solved researcher).

Numerical studies

1. Consider the case x depends on one variable vector u, but researchers do not know and uses another vector u component, thus seeking dependence as a function of two variables $x = f(u_1, u_2)$. As a result, the observations sample is presented in the form: $\left\{x_i, u_1^i, u_2^i, i = \overline{1, s}\right\}$, $u_1 \in [0; 3]$, $u_2 \in [0; 3]$. Sample size $s = 500$. Noise is 5% of the corresponding variables measurement result. The true yield of the object described by the equation $x(u_1) = u_1^2$. Effect u_2 on the yield of the object is absent. Nevertheless, the dependence $x = f(u_1, u_2)$ is sought in the form of ratings:

$$x_s(u_1, u_2) = \frac{\sum_{i=1}^{s} x_i \Phi(c_{s1}^{-1}(u_1 - u_1^i))\Phi(c_{s2}^{-1}(u_2 - u_2^i))}{\sum_{i=1}^{s} \Phi(c_{s1}^{-1}(u_1 - u_1^i))\Phi(c_{s2}^{-1}(u_2 - u_2^i))}. \tag{7}$$

where $\Phi(z_j) = \begin{cases} 1 - |z_j|, & |z_j| \le 1, \\ 0, |z_j| \ge 1, \end{cases}$

Estimate the fuzziness $\widehat{c}_{s1}$ and $\widehat{c}_{s2}$ parameters are determined by solving problem (5) minimize the mean-square criterion in a moving the exam mode:

$$W(A_{s1}, A_{s2}) = s^{-1} \sum_{i=1}^{s} (x(u_i) - x_s(u_i, c_{s1}, A_{s2}))^2 \underset{c_{s1}, A_{s2}}{\longrightarrow} \min. \tag{8}$$

Using the above rule, we conclude that the components u, namely u_2, no effect on the value $x = f(u_1, u_2)$, that x is not dependent on u_2. We illustrate this fact in the following figure. In accordance with the above rule variable u_2 is excluded from the regression estimates. Fig. 2 shows the object true output and its evaluation, built in depending on one variable u_1.

Fig. 2 and the following figures show the evaluation object function and the true output, depending on the iterations number. Standard error is calculated as follow $\delta^2(\widehat{A}_{s_1}, \widehat{c}_{s_2}) = \frac{1}{s}\sum_{i=1}^{s}(x_i - x_s(u_1^i, u_2^i))^2$. In the present case $\delta^2 = \delta^2(\widehat{A}_{s_1}, \widehat{c}_{s_2}) = 0,00232$.

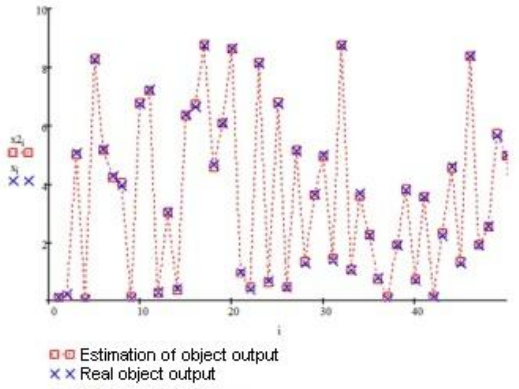


Figure 2

Construct the estimate (7) $x(u)$ on two variables: the essential - u_1 and irrelevant - u_2. At the same time $\widehat{c}_{s1} = 0.06$, $\widehat{c}_{s2} = 0.6$. Fig. 3 shows the object true output and its evaluation, built based on two variables; the essential - u_1 and irrelevant - u_2.

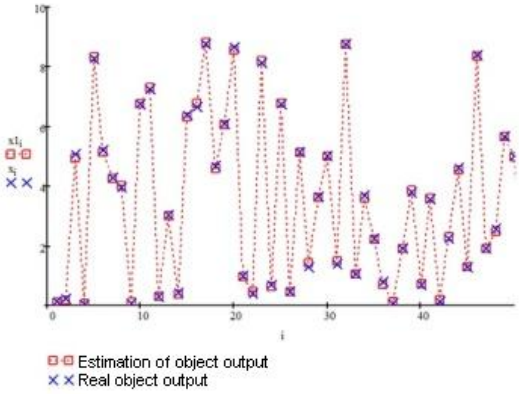


Figure 3

Standard error in this case his increased now $\delta^2 = 0,0086$. Components u_2 effect on an assessment $x(u_1, u_2)$ in this case is negative.

2. Consider the following example. Suppose that$x$ depends on two components of the vector u, however, one component influence, namely u_2, the object slightly output. As a result, the observations sample is presented in the form: $\left\{x_i, u_1^i, u_2^i, i = \overline{1,s}\right\}$, $u_1 \in [0;3]$, $u_2 \in [0;3]$. Sample size $s = 500$. Noise is 5% of the measurement result of the corresponding variables. The true yield of the object described by the equation $x(u_1, u_2) = u_1^2 + 0.1u_2$. Dependence $x = f(u_1, u_2)$ is sought in the form of (7).

Estimate the parameters of fuzziness $\widehat{c}_{s1}$ and $\widehat{c}_{s2}$ (5) are determined by solving the minimizing the mean square criterion problem (5) in the moving the exam mode. As a result, its solutions have been found estimates $\widehat{c}_{s1} = 0.06$, $\widehat{c}_{s2} = 0.6$.

Using the above rule, we conclude that the components u, namely u_2, has a significant impact on the values $x = f(u_1, u_2)$, that is x weakly dependent on u_2. In accordance with the above rule variable u_2 is excluded from the regression estimates. We illustrate this fact in the following figure. Fig. 4 shows the true output of the object and its score, constructed according to one variable u_1.

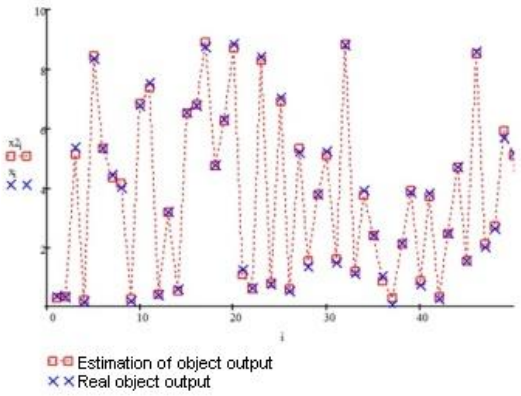


Figure 4

Standard error was $\delta^2 = 0,01$.

Construct the estimate (7) x on two variables: the essential - u_1 and irrelevant - u_2. Fig. 5 shows the object true output and its score, constructed based on two variables u_1 u_2.

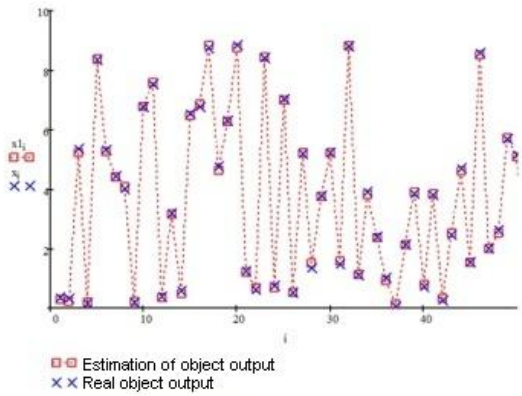


Figure 5

Standard error was 0.01.

3. Consider the case x depends on four vector u components, however, the influence of one component, namely u_4, the object output is slightly. As a result, the sample of observations is presented in the form: $\left\{x_i, u_1^i, u_2^i, u_3^i, u_4^i, i = \overline{1,s}\right\}$, $u_1 \in [0;3]$, $u_2 \in [0;3]$,$u_3 \in [0,3]$,$u_4 \in [0;3]$. The volume of samples $s = 500$. Noise is 5% of the corresponding variables measurement result. The true yield of the object described by the equation $x\left(u_1, u_2, u_3, u_4\right) = 0.4u_1^2 + 0.3u_2^3 + 0.2u_3^4 + 0.1u_4$. Dependence $x = f\left(u_1, u_2, u_3, u_4\right)$ is sought in the form of ratings:

$$x_s(u_1, u_2, u_3, u_4) = \frac{\sum_{i=1}^{s} x_i \Phi(c_{s1}^{-1}(u_1 - u_1^i))\Phi(c_{s2}^{-1}(u_2 - u_2^i))\Phi(c_{s3}^{-1}(u_3 - u_3^i))\Phi(c_{s4}^{-1}(u_4 - u_4^i))}{\sum_{i=1}^{s} \Phi(c_{s1}^{-1}(u_1 - u_1^i))\Phi(c_{s2}^{-1}(u_2 - u_2^i))\Phi(c_{s3}^{-1}(u_3 - u_3^i))\Phi(c_{s4}^{-1}(u_4 - u_4^i))}. \quad (9)$$

Estimate the parameters of blurring $\widehat{c}_{s1}$, $\widehat{c}_{s2}$, $\widehat{c}_{s3}$, $\widehat{c}_{s4}$ are determined by solving the minimizing the mean square criterion problem (5) in the moving the exam mode. As a result, its solutions have been found following the parameter estimates blur: $\widehat{c}_{s1} = 1$, $\widehat{c}_{s2} = 0.4$, $\widehat{c}_{s3} = 0.2$, $\widehat{c}_{s4} = 4.1$.

Using the above rule, we conclude that the components u, namely u_4, has a significant impact on the values $x = f\left(u_1, u_2, u_3, u_4\right)$, that is x weakly dependent on u_4. In accordance with the above rule variable u_4 is excluded from the regression estimates. We illustrate this fact in the following figure. Fig. 6 shows the of the object true output and its score, constructed according to three variables u_1, u_2, u_3.

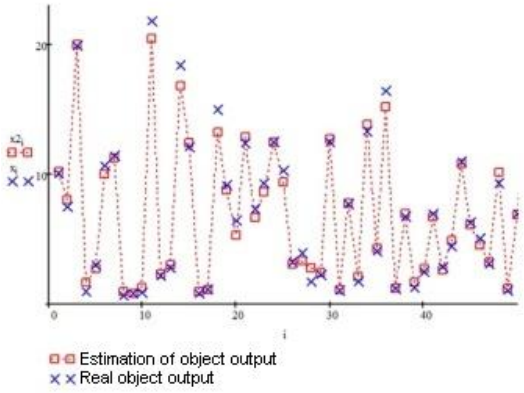


Figure 6

Standard error is $\delta^2 = 0,353$.

Construct estimation (9) x for the four variables u_1, u_2, u_3, u_4. Fig. 7 shows the true output of the object and its score, constructed according to the variables u_1, u_2, u_3, u_4.

Standard error was 0,361.

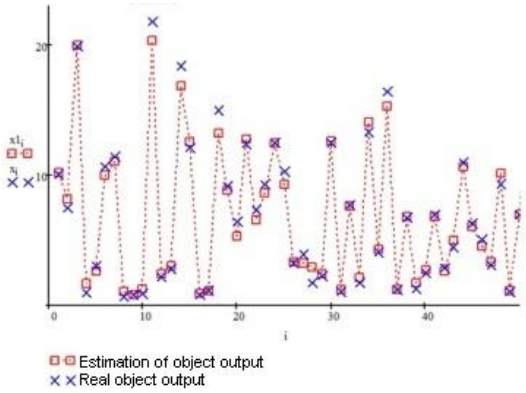


Figure 7

3 The regression functions robust estimation from observations with emission

When restoring a stochastic dependence on observations with errors random variables $x\,(u)$ and u errors, related unknown regression dependent $\tilde{x} = M\,\{x|u\}$, on the result of observations $(\overline{u_1}, x_1), (\overline{u_2}, x_2), ..., (\overline{u_s}, x_s)$, which $\overline{u_i} = (u_1^i, ..., u_k^i)$, faced with a situation where the dependence in the nonparametric estimates class is a priori unknown. In this case, the assessment can be used nonparametric estimation for Nadaraya-Watson [3] of the form (4).

In practice, one often encounters a situation where the sample dimensions are present emissions. Then use the nonparametric estimate (4) may lead to unsatisfactory results.

A solution to this problem, we propose the following method, which consists of two phases:

1. With the help of (4) recovering the regression at any point u_j, where $j = \overline{1,k}$, as a result we obtain $x1_j$. Smearing parameter c_s defined as in (5).

2. The second stage $x\,(u)$ is estimated in accordance with the statistics:

$$x2_s\,(u_1, ..., u_s) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{k} \Phi\left(c_s^{-1}\left(u_j - u_j^i\right)\right) \Phi\left(c_s^{-1}\left(x1_j - x_j^i\right)\right)}{\sum_{i=1}^{s} \prod_{j=1}^{k} \Phi\left(c_s^{-1}\left(u_j - u_j^i\right)\right) \Phi\left(c_s^{-1}\left(x1_j - x_j^i\right)\right)}. \tag{10}$$

The above assessment is final, ie, a regression curve according nonparametric estimation to data containing outliers.

Numerical studies

Need to recover an unknown dependence $x\,(u)$ on the available sample of observations $(\overline{u_1}, x_1), (\overline{u_2}, x_2), ..., (\overline{u_s}, x_s)$, where $\overline{u_i} = (u_1^i, ..., u_k^i)$.

We introduce the coefficient a for the two emissions are calculated by the formula $a = \frac{b}{s}$, where b - number of observations between two consecutive ejections. This is necessary only in the experimental conditions, when working with real data in this factor is not necessary. Thus,

on this factor can be judged on the relative emissions position. The more a, the farther apart are outliers. $0 < a < 1$.

In the numerical experiment was made: s=100, $u \in [0; 26]$, $x \in [-0,986; 1,061]$,$a = 0.6$. Fig. 8 shows a sample $(\overline{u_1}, x_1), (\overline{u_2}, x_2), ..., (\overline{u_s}, x_s)$.

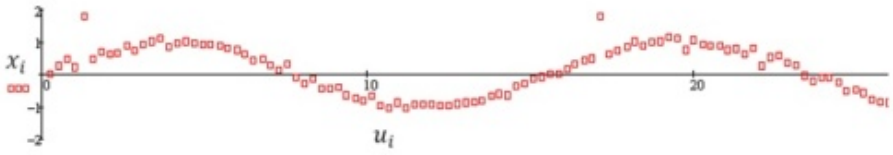


Figure 8

Restore the regression function estimate from observations of different ways:$x_s(u)$ - a non-parametric estimation for Nadaraya-Watson (1), the iterative formula for calculating the robust estimation $x3_s(u)$ is [7]:

$$x3_s^{l+1}(u) = \sum_{i=1}^{n} \left(\frac{\left|x_i - x3_s^l(u)\right|^{-1} \Phi\left(c_s^{-1}(u - u_i)\right)}{\sum_{j=1}^{n} \left|x_i - x3_s^l(u)\right|^{-1} \Phi\left(c_s^{-1}(u - u_i)\right)} \right) x_i, l = 0, 1, 2, ... \tag{11}$$

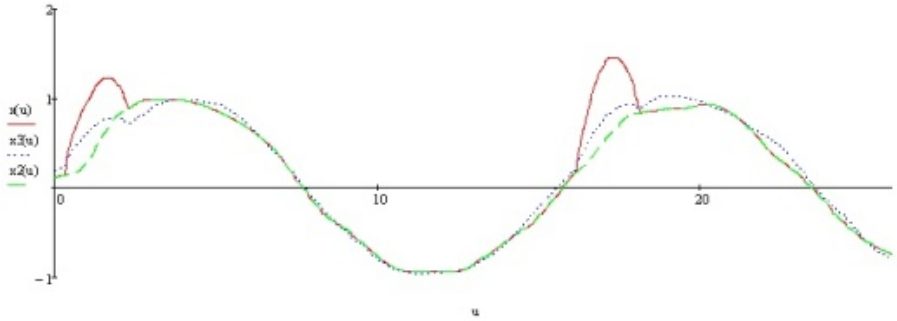


Figure 9

Assessment ends iterative correction when:$\left|x3_s^{l+1}(u) - x3_s^l(u)\right| \leq \varepsilon$, accepted: $\varepsilon = 0,01$, and $x2_s(u)$ the proposed robust estimation (10). The results are shown in Fig. 9.

Standard error is calculated as follows: $\delta^2(x_s(u)) = \frac{1}{s}\sum z_{i=1}^{s}(x_i - x_s(u))^2$. In this case, $\delta^2(x_s(u)) = 1,6$, $\delta^2(x2_s(u)) = 0,2$, $\delta^2(x3_s(u)) = 0,8$. $a = 0$ (Fig. 10)

$$\delta^2(x_s(u)) = 5,7, \delta^2(x2_s(u)) = 0,98, \delta^2(x3_s(u)) = 2,6.$$

Consider the case with the six releases. $a_{1,2} = 0.$, $a_{2,3} = 0.29$, $a_{3,4} = 0$, $a = 0.6$, $a_{4,5} = 0.39$, $a_{5,6} = 0$. Fig. 12 shows a sample (u_i, x_i).

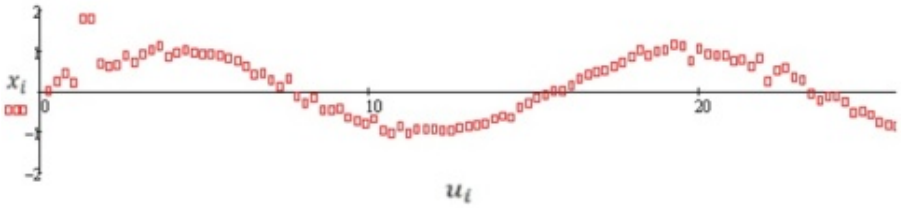


Figure 10

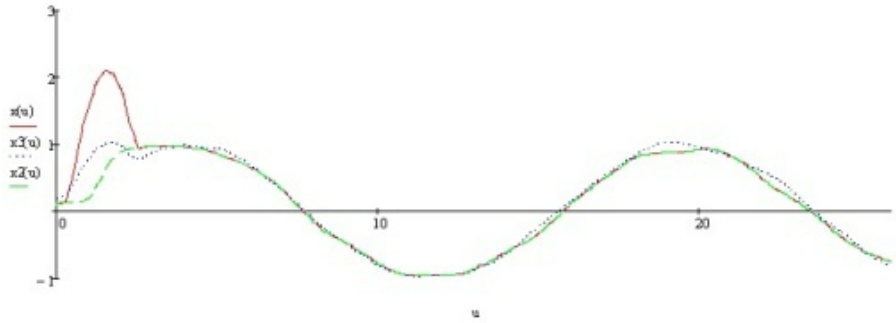


Figure 11

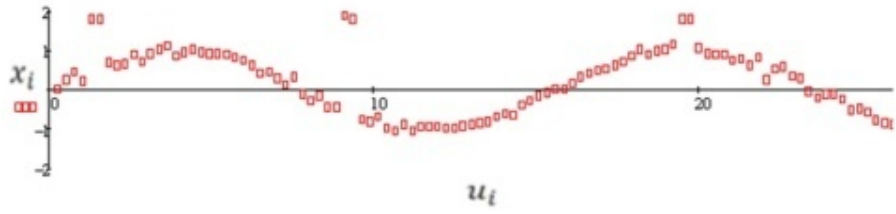


Figure 12

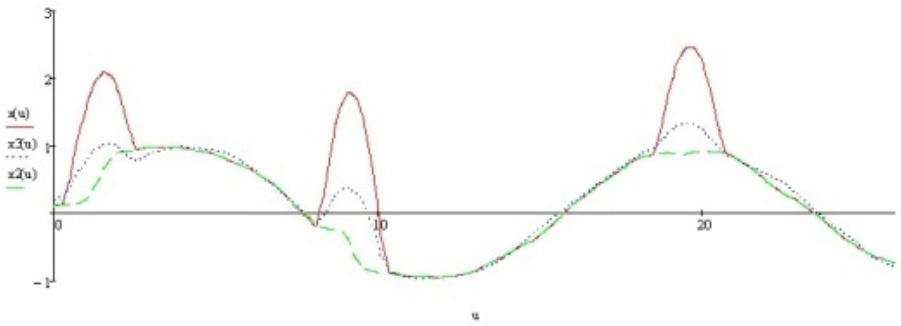


Figure 13

Standard error for $x2_s(u)$ – 0.25, for $x_s(u)$ – 9.947, for $x3_s(u)$ – 3.2
We consider the multidimensional version: $x_j \in [0; 30]$. Inputs - 5.

Table 1

Sample size	Coefficient of accumulation	Interference	Error for (1)	Error for (2)
100	0,63	5%	1.31	0.087
500	0,55	5%	0.585	0.022
100	0,02	5%	0.722	0.373
500	0,03	5%	0.723	0.023
100	0,63	10%	1.724	0.415
500	0,55	10%	0.731	0.252
100	0,02	10%	1.315	0.908
500	0,03	10%	0.95	0.264

References

[1] Eickhoff P. (1975). *Fundamentals of control systems identificaion*. Mir, Moscow.

[2] Zipkin Ja.Z. (1984). *Fundamentals of information theory identificaion*. Nauka, Moscow.

[3] Nadaraja E.A. (1983). *Nonparametric estimation of probability density and regression curve*. Tbilisy's University, Tbilisy.

[4] Medvedev A.V. (1983). *Adaptation nonparametric systems*. Nauka, Novosibirsk.

[5] Huber P. (1984). *Robustness in statistics*. Mir, Moscow.

[6] Hurrin Ju.S. (2008). *Optimality and robustness in statistical prognostication*. BGU, Minsk.

[7] Ruban A.I. (2004). *Methods of data analisys*. IPC KSTU, Krasnoyarsk.

Non-Parametric H-Models of Thermal Processes

R.S. Boyko, Ya.I. Demchenko
Siberian state aerospace university named after academician M.F. Reshetnev
Krasnoyarsk, Russia
e-mail: yademchenko86@mail.ru

Abstract

Some discrete-continuous processes identification and modeling tasks are considered. The new non-parametric estimations of probabilities distribution density function and regression curve according to observations are set, and also the convergence theorems. The new modifications of H-models are applied during the modeling of thermal process of oil decomposition.
Key words: modeling, identification, non-parametric estimations of regression function, non-parametric estimations of probabilities distribution density functions, convergence theorems, linear models, H-models.

Introduction

The problem of modeling and identification is one of cybernetics central problems. The main role during mathematical assertion of identification task belongs to information a priori [1] of researched process. Further we shall consider in a short form the levels of parametric and non-parametric uncertainty. In this work we shall draw a special attention at a situation, when the components of input variables vector are statistically dependent. The researched process runs in a space subregion of "input-output" variables, which have "tubular" structure [2]. The "tubular" structure was found during modeling of thermal processes of one-class During modeling of discrete-continuous processes according to real data we often face with some features, which appear in an observations sample selection of "input-output" variables. That is because of the following features:

1. variables meanings concentration in some subregions of measuring space;
2. emtinesses - absence of variables meanings in some sections of measuring interval;
3. rarefaction - presence of several meanings of measuring variables in some subregions.

The conditions 1,2,3 we shall call the A-conditions. As a matter of fact these features drove to necessity of construction of the new modifications of non-parametric estimations of probabilities distribution density and regression curve.

The content of identification in "narrow" meaning [3] (level of parametric uncertainty) supposes model parametric structure presence and some characteristics of random noises. It is simple for them to have a zero-mathematical expectation and bounded variance. In order to estimate the parameters the various iterative probabilistic procedures are used most often.

If the level of a priori information is non-parametric it is supposed that the model's parametric structure is absent, but in this way we need to know some qualitative information about the process, for example, certainty, or uncertainty of it's characteristics, linearity for dynamic processes or character of it's nonlinearity. For solution the identification task at this level of a priori information (identification in "broad" sense) the methods of non-parametric statistics may be used [4].

1 Parametric identification. H-models

Identification in "narrow" meaning supposes that one is to choose the class of models up to parameters accuracy with posterior parameters estimation with new object's "input-output" measured data coming.

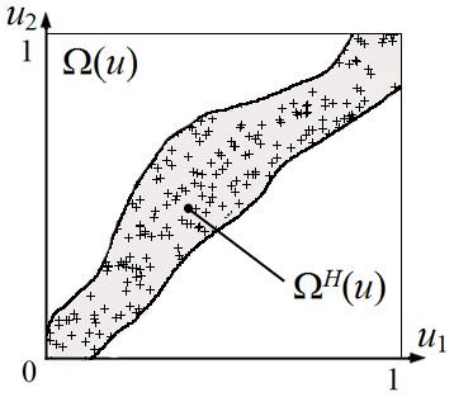


Figure 1

Parametric identification methods are developed good enough. Though, let us pay our attention at one feature, that have an important meaning for a practitioner during parameters estimation. For more simplification let us consider a static system with two inputs $u = (u_1, u_2)$and one output x. More than that, let us take as an example the simplest model like $\hat{x} = au_1 + bu_2$. If we have a sample of "input-output" measurements with random noise $\left\{x_i, u_1^i, u_2^i, i = \overline{1,s}\right\}$ it is easy to estimate parameters a and b. Let us consider such a case without community perturbation that $u_1 \in [0,1], u_2 \in [0,1], x \in [0,1]$, but the input variables are stochastically dependent, as we can see at the picture 1. Here $\Omega(u)$ - variables u_1 and u_2 presence area, $\Omega^H(u)$- real process behaviour area. It is clear that in a cube with side length 1, the area $\Omega^H(x,u)$ will have "tubular" structure, as we can see at picture 2.

It is clear from picture 2, that $\Omega^H(u) \subset \Omega(u)$, $\Omega^H(u,E) \subset \Omega(u,E)$. The areas $\Omega(u)$, $\Omega(x,u)$ are always known (in our case it is a cube with side length 1), and the areas $\Omega^H(u)$, $\Omega^H(x,u)$are never known. During the process $x = (u_1, u_2)$ building this fact is to be taken into account. Because of measurement mistakes or random facts, which influence to the process, and

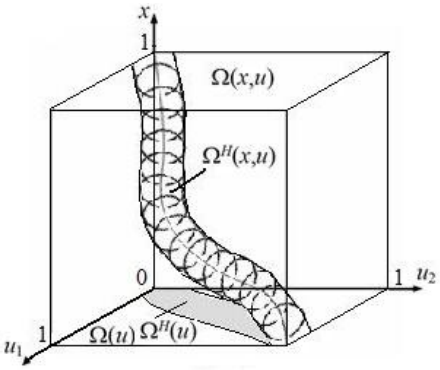


Figure 2

also during process research at model $\tilde{x}_\alpha(u) = A^\alpha(u, \alpha)$, where $\alpha = (a, b)$ the meanings $u \notin \Omega^H(u)$ may be defined (or selected). In this case estimation (forecast) x may be a value, which cannot be realized physically, i.e. $E \bar{\in} \Omega(E)$.

The parametric model of multi-dimensional statistic process may be taken in a view:

$$\tilde{x}(u_1, ..., u_n) = f(u_1, ..., u_n, \alpha_0, ..., \alpha_n), \tag{1}$$

where $x \in \Omega(x) \subset R^1$, $u \in R^n$, $\alpha_0, ..., \alpha_b$ – model parameters, which may be estimated with a help of sample we have out of s statistically independent observations of $(n+1)$-dimensional random value (u,x) - $\{(x_1, u_1), (x_2, u_2), ..., (x_s, u_s)\}$, f – model's chosen parametric structure with accuracy up to parameters vector.

When the dependence of vectors$u_1, ..., u_n$ from each other is stochastic, the researched process has "tubular" structure, which is similar to one at picture 2, and the model (1) transforms to non-parametric H-model like:

$$\tilde{x}^H(u_1, ..., u_n) = f(u_1, ..., u_n, \alpha_0, ..., \alpha_n)\theta^H(u_1, ..., u_n), \tag{2}$$

where $\theta^H(u_1, ..., u_n)$ – H-indicator. For example it may be defined like [2]:

$$\theta_s^H(u_1, ..., u_n) = sgn \sum_{i=1}^{s} \prod_{j=1}^{n} \Phi\left(\frac{u_j - u_{ji}}{C_s}\right), \tag{3}$$

where $\Phi(\cdot)$ – is a bell-like function, C_s – a parameter of blurriness [2]. As a bell-like function, for example, we can use the function

$$\Phi\left(\frac{u - u_i}{C_s}\right) = \begin{cases} 1, \text{ if } |u - u_i| \le C_s \\ 0, \text{ if } |u - u_i| > C_s \end{cases} \tag{4}$$

2 Non-parametric estimations of probabilities distribution density function

During identification of discrete-continuous processes we can build a mathematical model based upon non-parametric estimations of regression function according to observations. Let A - are the features which take place during variables monitoring, then appears a necessity of some correction of non-parametric estimations of probability distribution density and, as a consequence of that, non-parametric estimations of regression function.

Let (u, x) – is a random value with meanings in area $\Omega(u, x) \subset R^{n+1}$, and $p(u, x) > 0$ – is a distribution density of $(n + 1)$-dimensional random value (u, x), it is unknown. Let $(u_1, x_1), (u_2, x_2), ..., (u_s, x_s)$ – is a sample of s statistically independent observations of $(n + 1)$-dimensional random value $(u, x) \in \Omega(u, x)$.

During approximation of unknown function according to observations as a probability density non-parametric estimation $p(u) > 0 \ \ \forall u \in \Omega(u)$ usually the statistics [6,7] is taken:

$$p_s(u_1, ..., u_?) = \frac{1}{sC_s^n} \sum_{i=1}^{s} \prod_{j=1}^{n} \Phi\left(\frac{u_j - u_{ji}}{C_s}\right), \tag{5}$$

where the function which is integrated with square $\Phi(\cdot)$ and parameter Cs such, that they satisfy convergence conditions [4]:

$$0 < \Phi\left(\frac{u_j - u_{ji}}{!_s}\right) < \infty, \quad \lim_{s\to\infty} \frac{1}{C_s} \Phi\left(\frac{u_j - u_{ji}}{C_s}\right) = \delta(u_j - u_{ji}),$$

$$\frac{1}{C_s} \int_{\Omega(u)} \Phi\left(\frac{u_j - u_{ji}}{C_s}\right) du = 1, \ \ \frac{1}{C_s} \int_{\Omega(u)} u\Phi\left(\frac{u_j - u_{ji}}{C_s}\right) d\Omega(u) = u_i, \tag{6}$$

$$C_s > 0, s = 1, 2, \ldots, \lim_{s\to\infty} C_s = 0, \lim_{s\to\infty} sC_s^n = \infty.$$

When in a sample of "input-output" conditions observations there are some emptynesses, rarefactions and concentrations the new probability distribution density non-parametric estimation function is proposed:

$$\bar{p}_s(u_1, ..., u_?) = \frac{1}{sC_s^n} \sum_{i=1}^{s} \prod_{j=1}^{n} \Phi_1\left(\frac{u_j - u_{ji}}{C_s}\right) \Phi_2\left(\frac{u_j - u_{ji}}{C_s}\right), \tag{7}$$

where integrated with square and even relatively to u_i functions $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$, and parameter Cs are such that they satisfy the convergence conditions (6) and:

$$0 < \Phi_1\left(\frac{u_j - u_{ji}}{C_s}\right) \Phi_2\left(\frac{u_j - u_{ji}}{C_s}\right) < \infty, \lim_{s\to\infty} \frac{1}{C_s} \Phi_1\left(\frac{u_j - u_{ji}}{C_s}\right) \Phi_2\left(\frac{u_j - u_{ji}}{C_s}\right) = \delta(u_j - u_{ji}),$$

$$\frac{1}{C_s}\int_{\Omega(u)}\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)du=1, \tag{8}$$

$$\frac{1}{C_s}\int_{\Omega(u)}u\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)d\Omega(u)=u_i,$$

and also the probability distribution density function estimation is:

$$\tilde{p}_s(u_1,...,u_n)=\frac{1}{sC_s^n}\sum_{i=1}^{s}\prod_{j=1}^{n}\left(\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)+\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)\right), \tag{9}$$

where integrated with square and even relatively to u_i functions $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$, and parameter Cs are such, that they satisfy to convergence conditions:

$$0<\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)+\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)<\infty,\ \Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)<\infty,\ \Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)<\infty,$$

$$\lim_{s\to\infty}\frac{1}{C_s}\left(\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)+\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)\right)=\delta(u_j-u_{ji}),0\le\frac{1}{C_s}\int_{\Omega(u)}\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)du\le 1,$$

$$\frac{1}{C_s}\int_{\Omega(u)}\left(\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)+\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)\right)du=1,0\le\frac{1}{C_s}\int_{\Omega(u)}\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)du\le 1,$$

$$\lim_{s\to\infty}\frac{1}{C_s}\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)=\delta(u_j-u_{ji}),\ \lim_{s\to\infty}\frac{1}{C_s}\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)=\delta(u_j-u_{ji}), \tag{10}$$

$$C_s>0,s=1,2,\ldots,\lim_{s\to\infty}C_s=0,\lim_{s\to\infty}sC_s^n=\infty.$$

For proposed probability distribution density function estimation $\bar{p}_s(u_1,...,u_n)$ (7) the theorem takes place:

Theorem 1. Let $p(u)$is a twice differentiable, and functions $\Phi_1(\cdot)$, $\Phi_2(\cdot)$ and blurriness parameter Cs satisfies to convergence conditions (6) and (8), then:

$$\lim_{s\to\infty}\mathrm{M}\{(p(u)-\bar{p}_s(u_1,...,u_n))^2\}=0,\forall u\in\Omega(u).$$

The proof of theorem 1 is based upon following lemma.

Lemma 1.1. Non-parametric estimation of probability distribution density $\bar{p}_s(u_1,...,u_?)$ in conditions of theorem 1 is asymptotically unbiased i.e.:

$$\lim_{s\to\infty}M\left\{\bar{p}_s(u_1,...,u_?)\right\}=p(u).$$

More detailed proofs of theorem 1 and lemma 1.1 for non-parametric estimation of probability distribution density function $\bar{p}_s(u)$ (7) in a case $\Omega(u,x)\subset R^2$ are considered in [8].

For proposed estimation of probability distribution density function $\tilde{p}_s(u_1,...,u_n)$ (9) a theorem takes place:

Theorem 2. Let $p(u)$ is twice differentiable, and the functions $\Phi_1(\cdot)$, $\Phi_2(\cdot)$and blurriness parameter Cs satisfy convergence conditions (10), then:

$$\lim_{s\to\infty} \mathrm{M}\{(p(u) - \tilde{p}_s(u_1, ..., u_n))^2\} = 0, \forall u \in \Omega(u).$$

Proof of theorem 1 is based upon following lemma.

Lemma 2.1. Non-parametric estimation of probability distribution density $\tilde{p}_s(u_1, ..., u_n)$ in conditions of theorem 2 is asymptotically unbiased, i.e.:

$$\lim_{s\to\infty} M\left\{\tilde{p}_s(u_1, ..., u_n)\right\} = p(u).$$

More detailed proofs of theorem 2 and lemma 2.1 for non-parametric estimation of probability distribution density function $\tilde{p}_s(u)$ (9) in case $\Omega(u, x) \subset R^2$ are considered in [9].

2.1 Non-parametric estimations of regression function according to observations

Non-parametric estimation of regression x by u, $p(u) > 0$ with probability one looks like:

$$x_s(u_1, ..., u_n) = \sum_{i=1}^{s} x_i \prod_{j=1}^{n} \Phi\left(\frac{u_j - u_{ji}}{C_s}\right) \Bigg/ \prod_{j=1}^{n} \Phi\left(\frac{u_j - u_{ji}}{C_s}\right). \tag{11}$$

If (u, x) – is a random value with meanings in space $\Omega(u, x) \subset R^2$, and $p(u) > 0$ – is a two-dimensional random value distribution density (u, x), and $(u_1, x_1), (u_2, x_2), ..., (u_s, x_s)$ – is a sample of s statistically independent observations of two-dimensional random value $(u, x) \in \Omega(u, x)$, then non-parametric estimation of regression looks like:

$$x_s(u) = \sum_{i=1}^{s} x_i \Phi\left(\frac{u - u_i}{C_s}\right) \Bigg/ \Phi\left(\frac{u - u_i}{C_s}\right). \tag{12}$$

If there are some absences, rarefactions and condensations in sample of "input-output" observations it is proposed to use non-parametric estimations of regression curve like:

$$\bar{x}_s(u) = \frac{\sum_{i=1}^{s} x_i \Phi_1\left(\frac{u-u_i}{C_s}\right) \Phi_2\left(\frac{u-u_i}{C_s}\right)}{\sum_{i=1}^{s} \Phi_1\left(\frac{u-u_i}{C_s}\right) \Phi_2\left(\frac{u-u_i}{C_s}\right)}, \tag{13}$$

and

$$\tilde{x}_s(u) = \frac{\sum_{i=1}^{s} x_i \left(\Phi_1\left(\frac{u-u_i}{C_s}\right) + \Phi_2\left(\frac{u-u_i}{C_s}\right)\right)}{\sum_{i=1}^{s} \left(\Phi_1\left(\frac{u-u_i}{C_s}\right) + \Phi_2\left(\frac{u-u_i}{C_s}\right)\right)}. \tag{14}$$

When $\Omega(u, x) \subset R^{n+1}$, the new non-parametric estimations of regression function estimation (13) and (14) look like:

$$\bar{x}_s(u_1,...,u_n)=\frac{\sum_{i=1}^{s}x_i\prod_{j=1}^{n}\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)}{\sum_{i=1}^{s}\prod_{j=1}^{n}\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)},\tag{15}$$

where integrated with square and even relatively to u_i functions $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$, and blurriness parameter Cs are such, that they satisfy convergence conditions (6) and (8),

$$\tilde{x}_s(u_1,...,u_n)=\frac{\sum_{i=1}^{s}x_i\prod_{j=1}^{n}\left(\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)+\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)\right)}{\sum_{i=1}^{s}\prod_{j=1}^{n}\left(\Phi_1\left(\frac{u_j-u_{ji}}{C_s}\right)+\Phi_2\left(\frac{u_j-u_{ji}}{C_s}\right)\right)},\tag{16}$$

where integrated with square and even relatively to u_i functions $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$, and blurriness parameter Cs are such, that they satisfy convergence conditions (10).

The functions $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$, for example, look like:

$$\Phi_1(\frac{z}{C_s})=\left\{\begin{array}{l}1;\ if\ \ z<C_{\rm s}\\ 0;\ if\ \ z>C_{\rm s}\end{array}\right.,\Phi_2(\frac{z}{C_s})=\left\{\begin{array}{l}1;\ if\ C_{\rm s}/2<z\le C_{\rm s}\\ 1/2;\ if\ z\le C_{\rm s}/2\\ 0;\ if\ z>C_{\rm s}\end{array}\right.,\tag{17}$$

here $z=|u-u_i|$.

The new non-parametric estimations of regression curve (13) and (14), and also (15) and (16) are more steady to rarefaction, cavities and data missing in space of "input-output" meanings, than estimations (11) and (12), and in these conditions (13), (14), (15) and (16) in such conditions allows to approximate the required process more detailed, the results of numerical modeling are in [9].

For proposed non-parametric estimation of regression function $\bar{x}_s(u_1,...,u_n)$ (15) the following theorem takes place:

Theorem 3. Let $x(u)$ is twice differentiable and with probability one $p(u)>0$, $\forall u\in\Omega(u)$, and functions $\$_1(\cdot)$, $\$_2(\cdot)$ and blurriness parameter Cs satisfy convergence conditions (6) and (8), then:

$$\lim_{s\to\infty}\mathrm{M}\{(x(u)-\bar{x}_s(u_1,...,u_n))^2\}=0,\forall u\in\Omega(u).$$

The proof of theorem 3 is based upon the following lemma.

Lemma 3.1. Non-parametric regression estimation $\bar{x}_s(u)$ in conditions of theorem 3 is asymptotically unbiased, i.e.:

$$\lim_{s\to\infty}M\left\{\bar{x}_s(u_1,...,u_n)\right\}=x(u_1,...,u_n).$$

Such proofs of theorem 3 and lemma 3.1 for non-parametric estimation of regression function $\bar{x}_s(u)$ (15) in case $\Omega(u,x)\subset R^2$ are considered in [8].

For proposed non-parametric estimation of regression function $\tilde{x}_s(u_1,...,u_n)$ (16) the following theorem takes place:

Theorem 4. Let $x(u)$ is twice differentiable and with probability one $p(u) > 0, \forall u \in \Omega(u)$, and the functions $\Phi_1(\cdot)$, $\Phi_2(\cdot)$ and the blurriness parameter Cs satisfy the convergence conditions (10), then:

$$\lim_{s\to\infty} \mathrm{M}\{(x(u) - \tilde{x}_s(u_1, ..., u_n))^2\} = 0, \forall u \in \Omega(u).$$

The proof of theorem 4 is based upon following lemma.

Lemma 4.1. Non-parametric regression estimation $\tilde{x}_s(u)$ in conditions of theorem 4 is asymptotically unbiased, i.e.:

$$\lim_{s\to\infty} M\left\{\tilde{x}_s(u_1, ..., u_n)\right\} = x(u_1, ..., u_n).$$

Similar proofs of theorem 4 and lemma 4.1 for non-parametric regression function estimation $\tilde{x}_s(u)$ (16) in case $\Omega(u, x) \subset R^2$are considered in [9].

In order to define the mistake of gained non-parametric models the quadratic optimality criteria is used:

$$w(c_s) = \frac{1}{s}\sum_{j=1}^{s}(x_j - x_s(u_j, c_s))^2 \to \min_{c_s}, \tag{18}$$

where $i \neq j$.

Such criteria allows to estimate the closeness degree among true object's "input-output" measurements and model (estimation) of that object.

3 Oil decomposition process

3.1 Common data

In cooking practice usually vegetable oil is used for example sunflower oil, olive oil, hemp oil and other oils. Heating process lasts for a long time (1-3 hours) and with a high temperature (180-200 degrees celsius). During long heating of oil, there are some chemical processes in it. As a result of these processes some heavy and hard-soluble substances appear and accumulate in the oil. Most of these decomposition products are carcinogens and are dangerous for human health.

Data of Juravleva L.N. "Vegetable oils oxidizing research during long lasting heating and development the ways to stabilize oil. [10] "Russian science-research oil institute" were used in the work.

During experiment the oil was warmed up to 180° in a deep-fat fryer "Minutka" in an interval of 2-50 hours with tests in every 2 hours (in present work the interval was limited to 16 hours, because concentration of carcinogenic substances becomes dangerous for human health till this time). The equipment used: chromatograph "Kristall 2000M" with capillary column "Varian" CPF 420 with length 100 m with inner diameter 0,25 mm.

Here is a short characteristic of factors, which define the process:

1. Oil, technological space, fat – environment, which is being heated and in which the food is cooking.

2. Free fat acids, (u_1) – are the "easy to take" energy source, satisfies body energetic needs. Accumulates in fat tissues.

3. Hydro peroxide, (u_2) – chemically changed lipids or fat acids. The first products of fat autoxidation. They have no smell or taste.

4. Secondary decomposition products, (u_3) – carbonyl combinations, less-molecular acids, ethers, spirits etc. They say that every secondary oxidizing products appear as a result of some hydro peroxides transformations. Cherish one part of secondary products appears during hydro peroxides decomposition, and another part as a result of following reactions. They have unpleasant smell and taste.

5. Polymers, (u_4) – inorganic and organic, amorphous and crystal substances, which are gained by multiple atom-groups repeating.

6. Refraction coefficient (refraction index), (u_5) – equals to ratio between the falling ray angle sinus and refraction angle sinus. Characterizes oils purity, degree of their oxidation. It grows when there are some non-limited fat acids.

7. Density changing, (u_6) – relative density of vegetable oil may be defined as ratio between some defined volume mass and similar mass of distilled water during 20° or with a help of aerometer.

8. Peroxide number, (u_7) – witnesses about relative content of peroxide fat acids in researched fat.

9. Anizidine number, (u_8) – number, which defines content of secondary oxidizing products in oil (aldehydes).

The variable meanings belong to following intervals: $u_1 \in [4; 16]$, $u_2 \in [0,1; 0,2]$, $u_3 \in [15; 360]$, $u_4 \in [0,05; 0,4]$, $u_5 \in [1,4; 1,5]$, $u_6 \in [0,9; 0,95]$, $u_7 \in [3; 8]$, $u_8 \in [8; 356]$. Sample of "input-output" variables looks like $\left\{u_{1i}, u_{2i}, u_{3i}, u_{4i}, u_{5i}, u_{6i}, u_{7i}, u_{8i}, i = \overline{1,9}\right\}$.

3.2 Experimental researches

In order to research connection among the variables the correlation matrix was calculated according to data [10].

The correlation coefficients meanings among variables are quite high, so, variables $u_1, ..., u_8$ depend linearly among each other. As parametric models of oil thermal decomposition process the following equations were used:

Table 1: Correlation matrix

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
t	0.94	0.98	0.89	0.94	0.91	0.87	0.96	0.96
u_1	1	0.93	0.83	0.96	0.87	0.87	0.95	0.84
u_2	0.93	1	0.84	0.96	0.9	0.84	0.97	0.94
u_3	0.83	0.84	1	0.74	0.89	0.87	0.79	0.94
u_4	0.96	0.96	0.74	1	0.86	0.8	0.99	0.84
u_5	0.87	0.9	0.89	0.86	1	0.89	0.92	0.92
u_6	0.87	0.84	0.87	0.8	0.89	1	0.82	0.86
u_7	0.95	0.97	0.79	0.99	0.92	0.82	1	0.89
u_8	0.84	0.94	0.94	0.84	0.92	0.86	0.89	1

$$\begin{array}{l}
\tilde{u}_1 = \alpha_0 + \alpha_2 u_2 + \alpha_3 u_3 + \alpha_4 u_4 + \alpha_5 u_5 + \alpha_6 u_6 + \alpha_7 u_7 + \alpha_8 u_8 \\
\tilde{u}_2 = \beta_0 + \beta_1 u_1 + \beta_3 u_3 + \beta_4 u_4 + \beta_5 u_5 + \beta_6 u_6 + \beta_7 u_7 + \beta_8 u_8 \\
\tilde{u}_3 = \gamma_0 + \gamma_1 u_1 + \gamma_2 u_2 + \gamma_4 u_4 + \gamma_5 u_5 + \gamma_6 u_6 + \gamma_7 u_7 + \gamma_8 u_8 \\
\tilde{u}_4 = \eta_0 + \eta_1 u_1 + \eta_2 u_2 + \eta_3 u_3 + \eta_5 u_5 + \eta_6 u_6 + \eta_7 u_7 + \eta_8 u_8 \\
\tilde{u}_5 = \kappa_0 + \kappa_1 u_1 + \kappa_2 u_2 + \kappa_3 u_3 + \kappa_4 u_4 + \kappa_6 u_6 + \kappa_7 u_7 + \kappa_8 u_8 \\
\tilde{u}_6 = \lambda_0 + \lambda_1 u_1 + \lambda_2 u_2 + \lambda_3 u_3 + \lambda_4 u_4 + \lambda_5 u_5 + \lambda_7 u_7 + \lambda_8 u_8 \\
\tilde{u}_7 = \mu_0 + \mu_1 u_1 + \mu_2 u_2 + \mu_3 u_3 + \mu_4 u_4 + \mu_5 u_5 + \mu_6 u_6 + \mu_8 u_8 \\
\tilde{u}_8 = \rho_0 + \rho_1 u_1 + \rho_2 u_2 + \rho_3 u_3 + \rho_4 u_4 + \rho_5 u_5 + \rho_6 u_6 + \rho_7 u_7
\end{array} . \tag{19}$$

In order to estimate models parameters (equations coefficients of (19)) the least square method was used. Let's present obtained parametric models in matrix view:

$$U = A \times B, \tag{20}$$

where: $U = \begin{pmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \\ \tilde{u}_4 \\ \tilde{u}_5 \\ \tilde{u}_6 \\ \tilde{u}_7 \\ \tilde{u}_8 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ u_2 & u_1 & u_1 & u_1 & u_1 & u_1 & u_1 & u_1 \\ u_3 & u_3 & u_2 & u_2 & u_2 & u_2 & u_2 & u_2 \\ u_4 & u_4 & u_4 & u_3 & u_3 & u_3 & u_3 & u_3 \\ u_5 & u_5 & u_5 & u_5 & u_4 & u_4 & u_4 & u_4 \\ u_6 & u_6 & u_6 & u_6 & u_6 & u_5 & u_5 & u_5 \\ u_7 & u_7 & u_7 & u_7 & u_7 & u_7 & u_6 & u_6 \\ u_8 & u_8 & u_8 & u_8 & u_8 & u_8 & u_8 & u_7 \end{pmatrix}$,

$$A = \begin{pmatrix} -47,2 & -2,4 & 0,03 & 32,5 & 14,2 & 30,9 & 0,08 & 0,03 \\ 0,7 & 10 \cdot 10^{-4} & -8 \cdot 10^{-6} & -0,12 & -0,22 & -0,39 & 0,04 & -9 \cdot 10^{-6} \\ 2 \cdot 10^{3} & 26,1 & -62,9 & -815,21 & -378 & -1 \cdot 10^{3} & -4,7 & 1,08 \\ -0,4 & 2 \cdot 10^{-4} & 1,12 & 3 \cdot 10^{-4} & -0,46 & 0,93 & 0,08 & -89 \cdot 10^{-4} \\ 0,7 & 7 \cdot 10^{-3} & -0,05 & -2 \cdot 10^{-4} & -0,43 & 0,82 & 0,02 & 1 \cdot 10^{-4} \\ -1,2 & 0,02 & -2,41 & -1 \cdot 10^{-3} & -0,37 & 1,45 & -1 \cdot 10^{-5} & 2 \cdot 10^{-3} \\ -6 & 9 \cdot 10^{-3} & 4,03 & -6 \cdot 10^{-4} & 7,92 & 5,43 & -2 \cdot 10^{-3} & 3 \cdot 10^{-3} \\ -1 \cdot 10^{3} & -15,2 & 800,7 & 0,63 & 237,8 & 152,7 & 1 \cdot 10^{3} & 14,71 \end{pmatrix}.$$

These dependencies have cognitive meaning and may be used in further research. It is necessary to match that some variables are more essential during mathematical models building, and some are less essential (it is defined by model parameters).

Let's consider a parametric structure at an example of $\tilde{u}_1$ more detailed:

$$\tilde{u}_1 = -47,2 - 2,4u_2 + 0,03u_3 + 32,5u_4 + 14,2u_5 + 30,9u_6 + 0,08u_7 + 0,03u_8, \qquad (21)$$

where $\tilde{u}_1 \in \Omega(U_1) \subset R^8$, $U_1 = \{u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$.

Mean-square mistake w was calculated as a square taken from (18). $w_1 = 0,417$.

Analogous models were built also for vectors $u_2, u_3, u_4, u_5, u_6, u_7, u_8$.

In model $\tilde{u}_1$ (21) we shall place $u_2 = 0,1$, $u_3 = 15$, $u_4 = 0,05$, $u_5 = 1,4$, $u_6 = 0,9$, $u_7 = 3$, $u_8 = 8$ from acceptable intervals variables definition, and we gain $\tilde{u}_1 = 4$. Current value corresponds to acceptable borders of u_1 definition.

Now we shall place $u_2 = 0,2$, $u_3 = 360$, $u_4 = 0,4$, $u_5 = 1,5$, $u_6 = 0,9$, $u_7 = 3$, $u_8 = 8$ in (21) from acceptable intervals variables definition. We shall gain $\tilde{u}_1 = 26$. Such value runs out the acceptable borders of $u_1$definition, the maximum of u_1 is 16. Analogous results are gained for vectors $u_2, u_3, u_4, u_5, u_6, u_7, u_8$. Therefore, we can conclude about "tubular" process character.

3.3 H-models

As far as completed researches show, that process has "tubular" character, so the parametric models are not enough and there is need to use H-models. In order to build H-models, each gained parametric model was supplemented with appropriate H-indicator. H-models may be presented in matrix view like:

$$U^H = (A \times B) \times \Theta^H, \qquad (22)$$

where: $U^H = \begin{pmatrix} \tilde{u}_1^H \\ \tilde{u}_2^H \\ \tilde{u}_3^H \\ \tilde{u}_4^H \\ \tilde{u}_5^H \\ \tilde{u}_6^H \\ \tilde{u}_7^H \\ \tilde{u}_8^H \end{pmatrix}$, $\Theta^H = \begin{pmatrix} \theta_{s1}^H(u_2, ..., u_8) \\ \theta_{s2}^H(u_1, u_3, u_4, ..., u_8) \\ \theta_{s3}^H(u_1, u_2, u_4, ..., u_8) \\ \theta_{s4}^H(u_1, ..., u_3, u_5, ..., u_8) \\ \theta_{s5}^H(u_1, ..., u_4, u_6, u_7, u_8) \\ \theta_{s6}^H(u_1, ..., u_5, u_7, u_8) \\ \theta_{s7}^H(u_1, ..., u_6, u_8) \\ \theta_{s8}^H(u_1, ..., u_7) \end{pmatrix}$,

the matrix Θ^H elements are being calculated according to (3).

H-model in details at an example of $\tilde{u}_1$ is further:

$$\tilde{u}_1 = (-47,2 - 2,4u_2 + 0,03u_3 + 32,5u_4 + 14,2u_5 + \\ +30,9u_6 + 0,08u_7 - 0,03u_8)\theta_{s1}(u_2, ..., u_8), \quad (23)$$

where: $\tilde{u}_1 \in \Omega(U_1) \subset \Omega(U_1) \subset R^8$.

If we place in $\tilde{u}_1$ (23) the $u_2 = 0,2$, $u_3 = 360$, $u_4 = 0,4$, $u_5 = 1,5$, $u_6 = 0,9$, $u_7 = 3$, $u_8 = 8$, we shall gain $\tilde{u}_1 = 0$, or in current case $\tilde{u}_1 \in (\Omega(U_1) - \Omega(U_1))$, where the process doesn't exist.

So, the process "tubular" structure for the current case is considered.

Non-parametric H-models for $u_2, u_3, u_4, u_5, u_6, u_7, u_8$ were built by analogous way.

Conclusion

The new non-parametric estimations of regression function according to observations are considered. They work with presented gaps, condensations and rarefactions in sample of "input-output" observations and gain more accurate results, than usually used regression estimations, and also convergence theorems. Presented practical example of H-models shows existing of H-processes in practice. "Tubular" process researched quite slight and it needs to be much more deep researching.

References

[1] Medvedev A.V. (2010). Teorija neparametricheskih system. Processy (Non-parametric systems theory. Processes) *Vestnik SibSAU*. Krasnoyarsk, vol. **3**.

[2] Medvedev A.V. (1995). Analiz dannih v zadache identifikacii (Data analysis in identification) *Komp'uternij analiz dannih modelirovanija (Computer analysis of modeling data)*. Minsk, vol. **2**.

[3] Cypkin Ya.Z. (1984). *Osnovi informacionnoy teorii identifikacii (Basements of informational identification theory)*. Nauka, Moscow.

[4] Medvedev A.V. (1983). *Neparametricheskie sistemi adaptacii (non-parametric adaptation systems).* Nauka, Novosibirsk.

[5] *5. Metodi klassicheskoy I sovremennoy teorii avtomaticheskogo upravlenija. (Classic and modern methods of automatic direction theory).* V1: Matematicheskie modeli, dinamicheskie characteristiki I analiz system upravlenija (Mathematical models, dynamic characteristics and direction systems analysis) / (2004) under direction of Pupkov K.A., Egupova N.D. MGTU im. N.E. Baumana, Moscow.

[6] Parzen E. (1969). On estimation of a probability density. *Ann. Maht. Statistic.* N3, pp. 854-864.

[7] Epanechnikov V.A. (1969). Neparametricheskaya ozenka mnogomernoy plotnosti verojatnosti (Non-parametric estimation of multi-dimensional probability density) *Probability theory and it's applications.* Vol. **14**, rel. 1, pp. 156-162.

[8] Demchenko Ya.I. (2010). O nekotorih neparametricheskih ocenkah plotnosti veroyatnosti I krivoi regressii (article is about some non-parametric estimations of probability density and regression curve) Part 1. *Molodoj uchenij* . Chita, vol. **11**, Molodoj uchenij publishing house, pp. 10-20.

[9] Demchenko Ya.I. (2010). O nekotorih neparametricheskih ocenkah plotnosti veroyatnosti I krivoi regressii (article is about some non-parametric estimations of probability density and regression curve) Part 2. *Molodoj uchenij* . Chita, vol. **11**, Molodoj uchenij publishing house, pp. 21-32.

[10] Juravleva L.N. (2009). *Izuchenije okislenija rastitelnih masel pri visokotemperaturnom nagreve vo friture I razrabotka sposobov povishenija ih stabilnosti (researching of oxidizing of vegetable oils during high-temperature heating in deep-frying and development of the ways to increase oil's stability).* Gos. nauch. Uchrejd. "Vserossihskij nauchno-issledovatelskij institute jirov" Russian academy of agriculture sciencies.

Test of Goodness of Fit in Dose-Effect Model based on Finite Sample

Victor M. Kocheganov, Michail S. Tikhov
Nizhny Novgorod State University,
Nizhny Novgorod, Russia
e-mail: kocheganov@gmail.com

Abstract

The central limit theorem is proved for nonparametric estimator of distribution function $F(x)$. It is shown that the given estimation approaches for convolution of $F(x)$ and kernel $K_h(x)$ is better. The direct way of estimation of the distribution function $F(x)$ based on consistent estimate of characteristic function if offered.

Keywords: distribution function kernel estimator, asymptotic normality, integrated square error, summarized square error, convolution.

Introduction

The threshold model of a dose-effect dependence when the casual dose is entered into bioobject is considered and the alternative answer W is observed. The model basis is made by the assumption that there is latent quantity X— the threshold dose which is the biggest dose at which it is not observed effect in experiment. The quantity $W = I\{U > X\}$ is the event indicator $\{U > X\}$. The biological essence of the above-stated consists in the following. We will assume that it is a question of poison which gets to bioobject (all poison and all medicine, only doses divide them). For example, at biotest of water speech usually goes about Ceriodaphnia which are sensitive to pollution of water by substances and which for them are poison. For each poison theoretically there is a minimum dose which causes their destruction in test object. To estimate this dose for each bio-object it is labour-consuming enough. If the bioobject in experiment has survived, it has received a dose, obviously smaller the minimum lethal dose. For each bioobject this dose will be variously that is defined individual sensitivity of individuals of a biological kind to a tested preparation. However in homogeneous mass quantity X will be a random variable with unknown distribution function $F(x) = P\{X < x\}$. Such model is considered in [1].

Main results

Let $\{(X_i, U_i), 1 \le i \le n\}$ be stationary sequence of independent pairs random variables (X, U), where U has unknown distribution function $G(u)$ and density $g(u) > 0$ on $\mathbb{R}^1$. Sample $\mathcal{U}^{(n)} = \{(U_i, W_i),\ 1 \le i \le n\}$ is observed.

In works [1], [4] problems of estimation and test of goodness of fit $H_0 : F(x) = F_0(x)$ against alternative $H_1 : F(x) \neq F_0(x)$ were respectively considered. The following relation was used to estimate $F(x)$

$$\tilde{F}_n(x) = \sum_{i=1}^{n} W_i K\left(\frac{x-U_i}{h}\right) \Big/ \sum_{i=1}^{n} K\left(\frac{x-U_i}{h}\right),$$

where $K(x) \geq 0$ denote symmetric kernel and h is bandwidth converging to 0 with increasing sample size n. It has been shown that if $h = n^{-1/5}$ then

$$n^{2/5}\left(\tilde{F}_n(x) - F(x)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{N}\left(a(x), \sigma_0^2(x)\right),$$

$$a(x) = \frac{f'(x)g(x) + 2g'(x)f(x)}{g(x)},\ \sigma_0^2 = \frac{F(x)(1-F(x)) \parallel K \parallel^2}{g(x)},\ \parallel K \parallel^2 = \int K^2(x)dx$$

At concrete realisation of criteria and estimations on sample $\mathcal{U}^{(n)}$ of finite volume it has appeared that we receive an estimation which is closer not to function $F(x)$, and to function $F*K_h(x)$, where $F * K_h$ denote convolution of function $F(x)$ with function $K_h(x)$ and $K_h(x) = h^{-1} * K(xh^{-1})$. Convolution $F * K_h(x)$ differs from function $F(x)$, especially this difference is appreciable on tails of distribution $F(x)$. So, for example, if $F(x) = 1 - e^{-x}, x > 0$ (exponential distribution) and $K_h(x) = 1/(2h), -h < x < h$ then $F * K_h(x) = (x + h + 1 - e^{-(x+h)})/(2h)$ if $-h < x < h$ and $F * K_h(x) = 1 - e^{-x} \operatorname{sh} h/h$ if $x > h$. Thus, if initial distribution $F(x)$ exponential with support on an interval $(0, \infty)$ then distribution $F * K_h(x)$ has support on an interval $(-h, \infty)$ and is not exponential. Thus $n^{2/5}\left(\tilde{F}_n(x) - F * K_h(x)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{N}(0, \sigma_0^2(x))$.

Hence, at finite samples it is necessary for us to test a hypothesis concerning not of function $F_0(x)$, and a hypothesis concerning function $F_0 * K_h(x)$. For this purpose we can use either the integrated square error $I_n = \int \left(\hat{F}_n(x) - F_0 * K_h(x)\right)^2 \omega(x)dx$ or the summarized square error $S_{n,m} = \sum_{j=1}^{m} \left(\hat{F}_n(x_j) - F_0 * K_h(x_j)\right)^2 \omega(x_j)$.

It is shown that asymptotic distributions of these statistics are normal. As to estimation of function of distribution $F(x)$ we offer following procedure. Let $\varphi_K(t)$ be a characteristic function of density $K(x)$, and $\hat{\varphi}(t)$ be consistent estimation of characteristic function constructed on $\hat{F}_n(x)$. As an estimation of distribution function at finite n the following function is offered

$$\hat{F}_n(\beta) - \hat{F}_n(\alpha) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-it\alpha} - e^{-it\beta}}{it} \frac{\hat{\varphi}(t)}{\varphi_K(t/h)} dt$$

For example, the Epanechnikov's kernel $K(x) = (3/4)(1 - x^2)$ has characteristic function $\varphi_K(t) = (3\sin t - t\cos t)/t^3$, and for a kosinus-kernel $K(x) = (\pi/4)\cos(\pi x/2)$ we have $\varphi_K(t) = \pi^2 \cos t/(\pi^2 - 4t^2)$. Both kernels have the support on $[-1, 1]$.

Under the assumptions of regularity these estimations $\hat{F}_n(x)$ are asymptotically normal. That is

$$\frac{n^{2/5}\left(\hat{F}_n(x) - \mathbf{E}\left(\hat{F}_n(x)\right)\right)}{\sigma_n(x)} \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{N}(0,1),\ \sigma_n^2 = \mathbf{D}\left(\hat{F}_n(x)\right).$$

References

[1] Krishtopenko S.V., Tikhov M.S., Popova E.B. (2008). *Dose-effect*. Medicina, M., 288 p.

[2] Finney D.J. (1980). *Probit Analysis*. University Press, Cambridge, 333 p.

[3] Dette H., Neumeyer N., Pilz K.F. (2005). A note on nonparametric estimaton of the effective dose in quantal bioassay. *Journal of the American Statist. Assoc.*. Vol. **100**, No.470, pp. 303-307.

[4] Tikhov M.S. (2004). Statistical Estimation based on Interval Censored Data. *Param. and Semiparam. Models with Appl. to Rel., Surv. Analisys, and Qual. of Life:Springer-Verlag: Theor.&Meth.*, pp. 209-215.

Semi-Recursive Nonparametric Algorithms of Identification and Forecasting

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

Abstract

A class of semi-recursive kernel plug-in algorithms of identification and forecasting is considered. The main parts of the asymptotic mean square errors (AMSE) of the estimates are found. The algorithms of identification and forecasting are applied to investigate the dependence of Russian Federation's Industrial Production Index on the dollar exchange rate, direct investment, and export for the period from September 1994 till March 2004.

Keywords: identification, forecasting, kernel recursive estimator, mean square convergence.

Introduction

Numerous statistical problems of identification and prediction is connected to estimation of certain characteristics of the following expressions:

$$J(x) = H\left(\{a_i(x)\},\, i = \overline{1,s}\right) = H\left(a(x)\right). \tag{1}$$

Here $x \in \mathsf{R}^m$, $H(\cdot) : \mathsf{R}^{ms} \to \mathsf{R}^1$ is a given function,

$$a(x) = \left(a_1(x), \ldots, a_s(x)\right), \quad a_i(x) = \int g_i(y) f(x,y)dy, \quad i = \overline{1,s},$$

where $g_1, \ldots, g_s$ are the known Borel functions, $\int \equiv \int\limits_{\mathsf{R}^1}$, $f(\cdot,\cdot)$ is an unknown probability density function (p.d.f.) for the observed random vector $Z = (X, Y) \in \mathsf{R}^{m+1}$.

If $g_i(y) \equiv 1$, then $a_i(x) = \int f(x,y)dy = p\left(x\right)$, where $p\left(\cdot\right)$ is the marginal p.d.f. of the random variable X, and $f(y|\, x) = f(x,y)/p\left(x\right)$ is the conditional p.d.f.

Here are the well known examples of such functions:

— the conditional initial moments

$$\mu_m(x) = \int y^m f(y|x)dy,\ H(a_1,a_2) = \frac{a_1}{a_2},\ m \geq 1,$$

$g_1(y) = y^m,\ g_2(y) = 1;\ \mu_1(x) = r(x)$ is the regression line;

— the conditional central moments

$$V_m(x) = \int (y - r(x))^m f(y|x)dy,\ g_1(y) = y,\ \ g_2(y) = y^2, \ldots,\ g_m(y) = y^m,\ g_{m+1}(y) = 1;$$

$V_2(x) = D(x)$ is the conditional variance or volatility function.

1 Problem statement

Take the following expression as an estimate of the functional $a(x)$ at a point x:

$$a_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{g(Y_i)}{h_i^m}\mathbf{K}\left(\frac{x - X_i}{h_i}\right). \tag{2}$$

Here $Z_i = (X_i, Y_i)$, $i = \overline{1,n}$, is the $(m+1)$-dimensional random sample from p.d.f. $f(\cdot,\cdot)$, (h_i) is a sequence of positive bandwidths tending to 0 as $i \to \infty$, the kernel $\mathbf{K}(u) = \prod_{i=1}^{m} K(u_i)$ is a m-dimensional multiplicative function which does not need to possess the properties of p.d.f., $g(y) = (g_1(y), \ldots, g_s(y))$.

Note that (2) can be computed recursively by

$$a_n(x) = a_{n-1}(x) - \frac{1}{n}\left[a_{n-1}(x) - \frac{g(Y_n)}{h_n^m}\mathbf{K}\left(\frac{x - X_n}{h_n}\right)\right]. \tag{3}$$

This property is particularly useful when the sample size is large since (3) can be easily updated with each additional observation.

The recursive kernel estimate of $p(x)$ (the case when $m = 1$, $s = 1$, $g(y) = 1$, $H(a_1) = a_1$) was introduced by Wolverton and Wagner in [1] and apparently independently by Yamato [2], and has been thoroughly examined in [3].

For estimation of (1) we use the plug-in estimate

$$J_n(x) = H\left(a_n(x)\right). \tag{4}$$

2 Mean square error

Denote: $\sup\limits_{x} = \sup\limits_{x\in \mathsf{R}^m}$, $T_j = \int u^j K(u)du$, $j = 1, 2, \ldots$.

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

Definition 2. A Borel function $K(\cdot) \in \mathcal{A}$, if $\int |K(u)|\,du < \infty$, and $\int K(u)\,du = 1$.

Definition 3. A Borel function $K(\cdot) \in \mathcal{A}_\nu$, if $K(\cdot) \in \mathcal{A}$, $T_j = 0$, $j = 1, \ldots, \nu - 1$, $T_\nu \neq 0$, $\int |u^\nu K(u)|du < \infty$, and $K(u) = K(-u)$.

Definition 4. A sequence $(h_n) \in \mathcal{H}(m)$ if

$$(h_n + 1/(nh_n^m)) \downarrow 0, \quad \frac{1}{n}\sum_{i=1}^{n} h_i^\lambda = S_\lambda h_n^\lambda + o(h_n^\lambda),$$

where λ is a real number, S_λ is a constant independent on n.

Definition 5. Let t_n, $X_1, \ldots, X_n$ are vectors, and $t_n = t_n(X_1, \ldots, X_n)$. A sequence of functions $\{H(t_n)\}$ belongs to the class $\mathcal{M}(\gamma)$ if for any possible values $X_1, \ldots, X_n$ the sequence $\{|H(t_n)|\}$ is dominated by a sequence of numbers $(C_0 d_n^\gamma)$, $(d_n) \uparrow \infty$ as $n \to \infty$, $0 \le \gamma < \infty$, C_0 is a constant.

Put for $t, p = 1, \ldots, s;\ j = 1, \ldots, m:\ a = (a_1, \ldots, a_s) = a(x);\ H_t = \partial H(a)/\partial a_t;\ a_n = (a_{1n}, \ldots, a_{sn});\ a^{s+}(x) = \int |g^s(y)| f(x,y) dy;\ a_{t,\,p}(x) = \int g_t(y) g_p(y) f(x,y)\, dy;$

$a_{t,\,p}^{1+}(x) = \int |g_t(y) g_p(y)|\, f(x,y)\, dy;\ L = \int K^2(u)\, du;\ \omega_{i\nu}(x) = \dfrac{T_\nu}{\nu!} \sum\limits_{l=1}^{m} \dfrac{\partial^{\,\nu} a_i(x)}{\partial x_l^\nu}.$

Theorem 1 (the AMSE of the estimate $J_n(x)$). If for $t, p = \overline{1, s}$:

1) a kernel $K(\cdot) \in \mathcal{A}_\nu$, $\sup\limits_x |K(x)| < \infty$; 2) a sequence $(h_n) \in \mathcal{H}(m)$;

3) functions $a_{t,\,p}(\cdot) \in \mathcal{N}_0(\mathsf{R})$, $\sup\limits_x a_{t,\,p}^{1+}(x) < \infty$, $\sup\limits_x a_t^{1+}(x) < \infty$, $\sup\limits_x a_t^{4+}(x) < \infty$;

4) $a_t(\cdot) \in \mathcal{N}_\nu(\mathsf{R})$, $\sup\limits_x |a_t(x)| < \infty$, $\sup\limits_x \left| \dfrac{\partial^{\,\nu} a_t^{(rj)}(x)}{\partial x_l \partial x_t \ldots \partial x_q} \right| < \infty,\ \ l, t, \ldots, q = \overline{1, m}$;

5) $H(\cdot) \in \mathcal{N}_2(a)$; 6) $\{H(a_n)\} \in \mathcal{M}(\gamma), \quad 0 \le \gamma \le 1/4$.

Then the AMSE of the estimate $J_n(x)$ as $n \to \infty$

$$\mathsf{u}^2(J_n(x)) = \sum_{t,\,p=1}^{s} \sum_{j,\,k=1}^{m} H_t H_p \left[\frac{S_{-m} L^m a_{t,\,p}(x)}{n h_n^m} + S_\nu^2 \omega_{t\nu}(x) \omega_{p\,\nu}(x) h_n^{2\nu} \right] + O\left(\left[\frac{1}{n h_n^m} + h_n^{2\nu} \right]^{\frac{3}{2}} \right).$$

The proof is given in [4].

3 Nonparametric semi-recursive identification of the production function

Apply the results to estimate the macroeconomic production function.

Let $r(x)$, $x = (x_1, x_2, x_3) \in \mathsf{R}^3$ be the regression model of the three-factor production function, $a(x) = (a_1(x), a_2(x))$, $a_1(x) = \int yf(x,y)dy$, $a_2(x) = \int f(x,y)dy = p(x)$. Here $x_1 > 0$ is the capital input, $x_2 > 0$ is the labor input, $x_3 > 0$ is the nature input, $y > 0$ is a product, and $f(x,y) > 0$ only if $x_1 > 0,\ x_2 > 0,\ x_3 > 0,\ y > 0$. Then

$$r_n(x) = \sum_{i=1}^{n} \frac{Y_i}{h_i^3} \mathbf{K}\left(\frac{x - X_i}{h_i} \right) \Bigg/ \sum_{i=1}^{n} \frac{1}{h_i^3} \mathbf{K}\left(\frac{x - X_i}{h_i} \right) = \frac{a_{1n}(x)}{p_n(x)}. \tag{5}$$

Estimate (5) is called semi-recursive because it can be updated sequentially by adding extra terms to both the numerator and denominator when new observations became available [5], [6].

Let the kernel $\mathbf{K}(u) = K(u_1)K(u_2)K(u_3)$, $K(\cdot) \in \mathcal{A}_\nu$, $\sup\limits_{u \in \mathsf{R}^1} |K(u)| < \infty$, and $(h_n) \in \mathcal{H}(3)$. To find the AMSE of the estimate $r_n(x)$, we use Theorem 1. In view of 3) and 4) conditions of this theorem functions $a_i(z)$, $i = 1, 2$, and their derivatives are continuously differentiable up to the order ν for any $z \in \mathsf{R}^3$, and the function $\int y^4 f(x, y)dy$ is bounded on R^3. If $p(x) > 0$, then condition 5) is fulfilled. It seems impossible to find a majorizing sequence (d_n) (condition 6) of Theorem 1), since the denominator in (5) may be equal to zero. In some cases we can find a majorizing sequence according to Definition 5 with $\gamma = 0$ under $\nu = 2$ if, for example, $K(\cdot) \geq 0$, and $Y < \infty$ [7]. For $\nu > 2$ we can use the following piecewise smooth approximation $\widetilde{r}_n(x)$ [8]:

$$\widetilde{r}_n(x) = \frac{r_n(x)}{(1 + \delta_{n,\nu} |\, r_n(x)|^{\,\tau})^\rho},$$

where $\tau > 0$, $\rho > 0$, $\rho\tau \geq 1$, $\delta_{n,\nu} = O\left(h_n^{2\nu} + 1/(nh_n^3)\right)$, $(\delta_{n,\nu}) \downarrow 0$ as $n \to \infty$.

In this case it is enough to take even $\tau \geq 4$, and as $n \to \infty$ (see [8])

$$\mathsf{u}^2(\widetilde{r}_n(x)) = \sum_{i,\,p=1}^{2} H_i H_p \left(S_{-3} \frac{LB_{i,p}}{nh_n^3} + S_\nu^2\, \omega_{i\nu}(x)\omega_{p\nu}(x) h_n^{2\nu} \right) + O\left(\left[\frac{1}{nh_n^3} + h_n^{2\nu} \right]^{3/2} \right),$$

where $H_1 = 1/p(x)$, $H_2 = -r(x)/p^2(x)$; $B_{1,1} = \int y^2 f(x, y)dy$, $B_{2,2} = p(x)$; $B_{1,2} = B_{2,1} = \int y f(x, y)dy$, $\omega_{1\nu}(x) = \dfrac{T_\nu}{\nu!} \left(\dfrac{\partial^{\,\nu} a_1(x)}{\partial x_1^\nu} + \dfrac{\partial^{\,\nu} a_1(x)}{\partial x_2^\nu} + \dfrac{\partial^{\,\nu} a_1(x)}{\partial x_3^\nu} \right)$,

$\omega_{2\nu}(x) = \dfrac{T_\nu}{\nu!} \left(\dfrac{\partial^{\,\nu} p(x)}{\partial x_1^\nu} + \dfrac{\partial^{\,\nu} p(x)}{\partial x_2^\nu} + \dfrac{\partial^{\,\nu} p(x)}{\partial x_3^\nu} \right)$.

4 Nonparametric semi-recursive dynamic models

Generalize the above results, given for independent observations (random samples), to time series.

4.1 Identification

In [9] an autoregressive heteroscedastic model satisfying geometric ergodicity conditions is considered. The approach allows us to estimate dynamic production functions with lagged values of the output.

Suppose that a sequence $(Y_t)_{t=\ldots,-1,0,1,2,\ldots}$ is generated by a nonlinear homoscedastic ARX process of order (m, s)

$$Y_t = \Psi(Y_{t-i_1}, \ldots, Y_{t-i_m}, X_t) + \xi_t = \Psi(U_t) + \xi_t, \tag{6}$$

where $X_t = (X_{1t}, \ldots, X_{st})$ are exogenous variables, $U_t = (Y_{t-i_1}, \ldots, Y_{t-i_m}, X_t)$, $1 \leq i_1 < i_2 < \ldots < i_m$ is the known subsequence of natural numbers, (ξ_t) is a sequence of independent identically

distributed (with density positive on R^1) random variables with zero mean, finite variance, zero third, and finite fourth moments, $\Psi(\cdot)$ is an unknown non-periodic function bounded on compacts. Assume that the process is strictly stationary.

Criteria for geometric ergodicity of a nonlinear heteroscedastic autoregression and ARX models which in turn imply α-mixing have been given by many authors (see, for example, [10]–[12]).

Let $Y_1, \ldots, Y_n$ be observations generated by process (6). The conditional expectation $\Psi(x, z) = \Psi(u) = \mathsf{E}(Y_t|U_t = u) = \mathsf{E}(Y_t|u)$, $(x, z) = u \in \mathsf{R}^{m+s}$ we estimate by the statistic, which is a semi-recursive counterpart of the Nadaraya–Watson estimate [13], [14] (similarly to (5)):

$$\Psi_{n,\,m+s}\,(u) = \sum_{t=2}^{n} \frac{Y_t}{h_t^{m+s}} \mathbf{K}\left(\frac{u - U_t}{h_t}\right) \Bigg/ \sum_{t=2}^{n} \frac{1}{h_t^{m+s}} \mathbf{K}\left(\frac{u - U_t}{h_t}\right). \tag{7}$$

Since the observations are dependent, investigation of the estimates properties becomes much harder. For example, the main part of the Nadaraya-Watson estimate's AMSE for strongly mixing (s.m.) sequences was found only in 1999 [15].

According to [9], if the observed sequence satisfies the s.m. condition with a s.m. coefficient $\alpha(\tau)$ such that

$$\int_0^\infty \tau^2 [\alpha(\tau)]^{\frac{\delta}{2+\delta}} d\tau < \infty \tag{8}$$

for some $0 < \delta < 2$, then Theorem 1 holds. Note that a s.m. coefficient with the geometric rate satisfies condition (8).

We will examine the dependence of Russian Federation's Industrial Production Index (IPI) Y on the dollar exchange rate X_1, import X_2, and direct investment X_3 for the period from September 1994 till March 2004. The data are available from: http://www.gks.ru and http://sophist.hse.ru/. Apply (7) under

$$U_t = (Y_{t-1}, X_{1t}, X_{2t}, X_{3t}, X_{3(t-1)}). \tag{9}$$

The identification results one can be seen in Fig. 1.

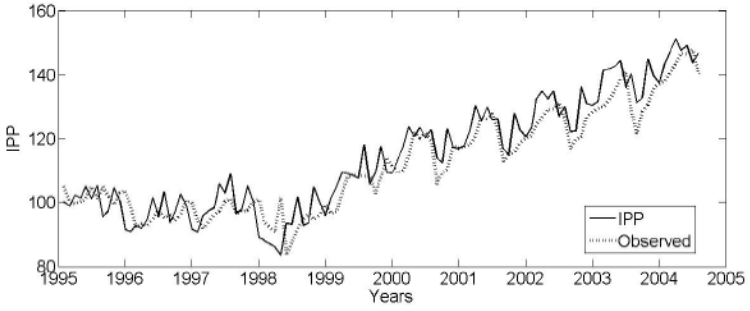


Figure 1: Identification

The kernel $K(u)$ used is the Gaussian kernel and the bandwidths $h_{jt} = 1.1\hat{\sigma}_j t^{-1/9}$, where $\hat{\sigma}_j$, $j = 1, 2, 3, 4, 5$ are the corresponding sample mean square deviations, the constant 1.1 is chosen subjectively. To compare nonparametric algorithms (7) and the least-squares estimators, we has calculated the relative errors A_n and relative average annual errors $A(t)$, $t = 1994, \ldots, 2004$ for both the approaches:

$$A_n = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|, \quad A(t) = \frac{1}{12}\sum_{i=1}^{12}\left|\frac{Y_i(t) - \hat{Y}_i(t)}{Y_i(t)}\right|,$$

where Y_i is the true value of the IPI and $\hat{Y}_i$ is its estimator. The results of such a comparison are given in Fig. 2.

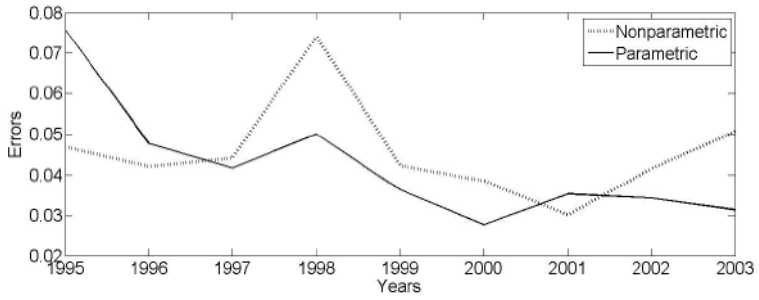


Figure 2: Identification relative errors $A(t)$

The result of 1998 can be explained by 1998 Russian financial crisis ("Ruble crisis") in August 1998.

4.2 Forecasting

To predict the IPI Y, we will apply (7) under

$$U_t = (Y_{t-1}, X_{1(t-1)}, X_{2(t-1)}, X_{3(t-1)}, X_{3(t-2)}). \tag{10}$$

The structure of data (10) provides the following forecast for Y_n :

$$\hat{Y}_n = \Psi_{n,5}\left(Y_{n-1}, X_{1(n-1)}, X_{2(n-1)}, X_{3(n-1)}, X_{3(n-2)}\right) = \sum_{t=2}^{n-1} Y_t \frac{K_t}{H_t} \Bigg/ \sum_{t=2}^{n-1} \frac{K_t}{H_t}, \tag{11}$$

where $H_t = \prod_{j=1}^{5} h_{jt}$, and the five-dimensional kernel K_t is defined by the formula

$$K_t = K\left(\frac{Y_{n-1} - Y_{t-1}}{h_{1t}}\right)\prod_{j=1}^{3} K\left(\frac{X_{j(n-1)} - X_{j(t-1)}}{h_{(j+1)t}}\right) K\left(\frac{X_{3(n-2)} - X_{3(t-2)}}{h_{5t}}\right).$$

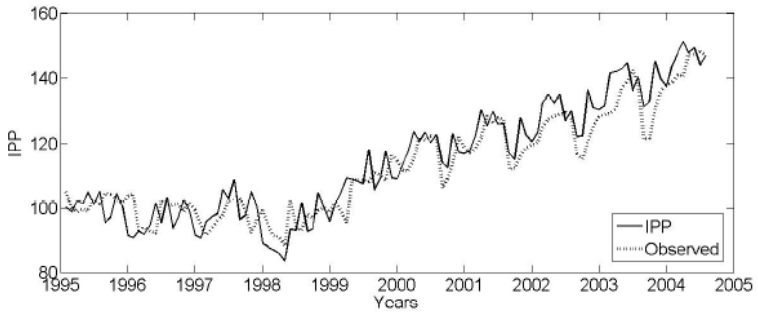


Figure 3: Forecasting

Statistic (11) may be interpreted as the predicted value based on the past information.

To find the AMSE of the estimate $\Psi_{n,5}(u)$ we use Theorem 2 [9].

The relative error of the forecast (REF) A_n obtained with $\Psi_{n,5}(\cdot)$ is 5%. Here the bandwidths $h_{jt} = 0.9\hat{\sigma}_j t^{-1/9}$, where $\hat{\sigma}_j,\ j = 1, 2, 3, 4, 5$. In Table 1 the REFs for each year from 1995 are given.

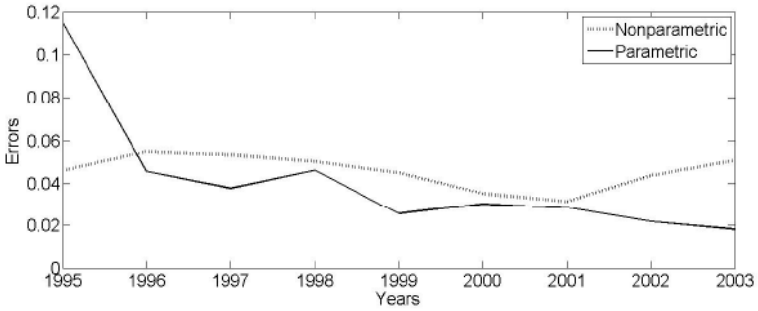


Figure 4: Forecasting relative errors $A(t)$

Table 1: Relative errors A_n

	Parametric	Nonparametric
Identification	0.0414	0.0448
Forecasting	0.045	0.05

Acknowledgements

This investigation was supported by the Russian Foundation for Basic Research (project № 09-08-00595-a).

References

[1] Wolverton C.T., Wagner T.J. (1969). Recursive Estimates of Probability Densities. *IEEE Trans. Syst. Sci. and Cybernet.* Vol. **5**, no. **3**, pp. 246-247.

[2] Yamato H. (1971). Sequential Estimation of a Continuous Probability Density Function and Mode. *Bulletin of Mathematical Statistics*. Vol. **14**, pp. 1-12.

[3] Wegman E.J., Davies H.I. (1979). Remarks on Some Recursive Estimates of a Probability Density Function. *Ann. Statist.* Vol. **7**, no. **2**, pp. 316327.

[4] Kitaeva A.V., Koshkin G.M. (2009). Recurrent Nonparametric Estimation of Functions from Functionals of Multidimensional Density and Their Derivatives. *Autom. Remote Control.* No. **3**, pp. 389-407.

[5] Ahmad J.A., Lin P.E. (1976). Nonparametric Sequential Estimation of a Multiple Regression Function. *Bull. Math. Statist.* Vol. **17**, no. **1-2**, pp. 63-75.

[6] Buldakov V.M., Koshkin G.M. (1977). On the Recursive Estimates of a Probability and a Regression Line. *Problems Inform. Trans.* Vol. **13**, pp. 41-48.

[7] Collomb G. (1976). *Estimation non Parametrique de la Regression par la Methode du Noyau.* Thèse Docteur Ingénieur. Univ. Paul-Sabatier, Toulouse.

[8] Koshkin G.M. (1999). Deviation Moments of the Substitution Estimate and Its Piecewise Smooth Approximations. *Siberian Math. J.* Vol. **40**, no. **3**, pp. 515-527.

[9] Kitaeva A.V., Koshkin G.M. (2010). Semi-Recursive Nonparametric Identification in the General Sense of a Nonlinear Heteroscedastic Autoregression. *Autom. Remote Control.* Vol. **71**, no. **2**, pp. 257-274.

[10] Masry E., Tjøstheim D. (1995). Nonparametric Estimation and Identification of Nonlinear ARCH Time Series. *Econom. Theory*. Vol. **11**, pp. 258-289.

[11] Doukhan P., Tsybakov A. (1993). Nonparametric Robust Estimation in Nonlinear ARX Models. *Problems of Information Transmission*. Vol. **29**, no. **4**, pp. 318327.

[12] Doukhan P. (1994). Mixing: properties and examples. *Lecture Notes in Statistics*. Vol. **85**. Springer-Verlag. 1994.

[13] Nadaraya E.A. (1964). On Regression Estimates. *Theor. Prob. App.* Vol. **19**, no. **1**, pp. 147149.

[14] Watson G.S. (1964). Smooth Regression Analysis. *Sankhya. Indian J. Statist.* Vol. A**26**, pp. 359-372.

[15] Bosq D., Cheze-Payaud N. (1999). Optimal Asymptotic Quadratic Error of Nonparametric Regression Function Estimates for a Continuous-Time Process from Sampled-Data. *Statistics*. Vol. **32**, pp. 229-247.

Nonparametrical Estimation of Survival Functions by Censored Data

Abdurahim Abdushukurov
Department of Probability Theory and Mathematical Statistics,
National University of Uzbekistan, Tashkent, Uzbekistan
e-mail: a_abdushukurov@mail.ru

Incomplete observations occur in survival analysis, especially in clinical trials and engineering when we partially observe death in biological organisms or failure in mechanical systems.

From statistical literature one can learn that incomplete observations arise in two ways: by censoring and truncation. Note that truncation is sampling an incomplete population, while censoring occurs when we are able to sample the complete population, but the individual values of observations below and/or above given values are not specified. Therefore censoring should not be confused with truncation. In this work we deal only with right censoring model which is easily described from methodological point of view.

Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables (i.i.d.r.v.-s) (the lifetimes) with common distribution function(d.f.) F. Let X_j be censored on the right by Y_j, so that observations available for us at the n-th stage consist of the sample $S^{(n)} = \{(Z_j, \delta_j), 1 \le j \le n\}$, where $Z_j = min(X_j, Y_j)$ and $\delta_j = I(X_j \le Y_j)$ with $I(A)$ meaning the indicator of the event A. Suppose that Y_j are again i.i.d.r.v.-s, the so-called censoring times with common d.f. G, independent of lifetimes X_j.

The main problem consist of nonparametrically estimating F with nuisance G based on censored sample $S^{(n)}$, where r.v.-s of interest X_j-s observed only when δ_j=1. Kaplan and Meier (1958) were the first to suggest the product-limit (PL) estimator F_n^{PL} defined as

$$F_n^{PL}(t) = \begin{cases} 1 - \prod_{\{j:Z_{(j)} \le t\}} \Big[1 - \frac{\delta_{(j)}}{n-j+1}\Big], & t \le Z_{(n)}, \\ 1, & t > Z_{(n)}, \delta_{(n)} = 1, \\ undefined, & t > Z_{(n)}, \delta_{(n)} = 0, \end{cases}$$

where $Z_{(1)} \le ... \le Z_{(n)}$ are the order statistics of Z_j and $\delta_{(1)}, ..., \delta_{(n)}$ are the corresponding δ_j. In statistical literature there are different versions of this estimator. However, those do not coincide if the largest Z_j is a censoring time. Gill (1980) redefined the F_n^{PL} setting $F_n^{PL}(t) = F_n^{PL}(Z_{(n)})$ when $t > Z_{(n)}$. Further, we use the Gill's modification of PL-estimator. At present there is an enormous literature on properties of the PL-estimator (see, for example [3]-[9]) and most of work on estimating incomplete observation are concentrated on PL-estimator. However F_n^{PL} is not unique estimator of F.

The second, closely related with the F_n^{PL}, nonparametrical estimator of F is the exponential hazard estimator

$$F_n^E(t) = 1 - exp\Big\{ - \sum_{j=1}^{n} \frac{\delta_{(j)} I(Z_{(j)} \le t)}{n-j+1} \Big\}, -\infty < t < \infty.$$

F_n^E plays an important role in investigating the limiting properties of the estimator F_n^{PL}.
Abdushukurov (1998,1999) proposed another estimator for F of power type:

$$F_n(t) = 1 - (1 - H_n(t))^{R_n(t)} = \begin{cases} 0, & t < Z_{(1)}, \\ 1 - (\frac{n-j}{n})^{R_n(t)}, & Z_{(j)} \le t < Z_{(j+1)}, 1 \le j \le n-1, \\ 1, & t \ge Z_{(n)}, \end{cases}$$

where

$$H_n(t) = \frac{1}{n} \sum_{j=1}^{n} I(Z_j \le t)$$

is empirical estimator of d.f. $H(t) = P(Z_j \le t) = 1 - (1 - F(t))(1 - G(t))$ and

$$R_n(t) = \frac{-log(1 - F_n^E(t))}{\sum_{j=1}^{n} \frac{I(Z_{(j)} \le t)}{n-j+1}}.$$

As we see, estimator F_n is defined on whole line. Let

$$a_n(t) = \sum_{j=1}^{n} \frac{I(Z_{(j)} \le t)}{(n-j)(n-j+1)}.$$

Note that $\sup\{a_n(t), t \le T\} \le [n(1-H_n(T))]^{-1} = \mathbb{O}(\frac{1}{n})$ with probability 1, where $T < Z_{(n)}$.

Following inequalities show that all three estimators are closely related (Abdushukurov [1-5]): *For $t < Z_{(n)}$ with probability 1*

(I) $0 < -log(1 - F_n^{PL}(t)) + log(1 - F_n^E(t)) < a_n(t)$;
(II) $0 \le F_n^{PL}(t) - F_n^E(t) < \frac{1}{2} a_n(t)$;
(III) $0 < -log(1 - F_n(t)) + log(1 - F_n^E(t)) < a_n(t)$;
(IV) $|-log(1 - F_n^{PL}(t)) + log(1 - F_n(t))| < a_n(t)$;
(V) $|F_n^{PL}(t) - F_n(t)| < a_n(t)$;
(VI) $|F_n^E(t) - F_n(t)| < a_n(t)$.

Thus one can expect the stochastic equivalences of these estimators in the sense of their weak convergence to the same Gaussian process (Abdushukurov (1998)). Let d.f.-s F and G be continuous and $T < T_H = \inf\{t : H(t) = 1\}$. Then one can define the sequence of Wiener processes $\{\mathbb{W}_n(x), 0 \le x < \infty\}_{n=1}^{\infty}$ such that when $n \to \infty$

$$\sup_{t \le T} |n^{\frac{1}{2}}(F_n^*(t) - F(t)) - (1 - F(t))\mathbb{W}_n(d(t))| \stackrel{a.s.}{=} O(n^{-1/2} \log n),$$

where F_n^* stands for one of estimators F_n^{PL}, F_n^E, F_n and

$$d(t) = \int_{-\infty}^{t} [(1 - F)^2(1 - G)]^{-1} dF.$$

Here we state the convergence result in the form of strong approximation by the sequence of appropriate copies of limiting Gaussian process, with optimal rate. As consequence from here we obtain that $n^{\frac{1}{2}}(F_n^* - F)$ converges weakly in the Skorochod's space $\mathbb{D}(-\infty, T]$ to the mean-zero Gaussian process with covariance function $\sigma(t;s) = (1-F(t))(1-F(s))d(min(t,s))$, $t, s \leq T$. Thus we see that all three estimators are equivalent in the asymptotic sense. But as we see in [3-5] the estimator F_n has some peculiarities and even a better properties than F_n^{PL} and F_n^E do for all $n \geq 1$. Let's consider the following exponential representation for any right continuous d.f. (Gill (1980)):

$$1 - F(t) = exp\Big\{ - \int_{-\infty}^{t} \frac{dF(u)}{1-F(u-)} \Big\} \prod_{s \leq t} (1 - \Delta\Lambda(s)),$$

where $\Delta\Lambda(s) = (F(s) - F(s-))/(1 - F(s-))$ and $F(s-) = lim_{u \uparrow s} F(u)$. Then we see that F_n^{PL} is a natural estimator for $\prod_{s \leq t}(1 - \Delta\Lambda(s))$, that is a discrete d.f.. On the other side, F_n^E and F_n are a natural estimators for continuous d.f. $F(t) = 1 - exp\Big\{ - \int_{-\infty}^{t} (1-F)^{-1} dF \Big\} = 1 - (1 - H(t))^{R(t)}$, where $R(t) = -log(1-F(t))/[-log(1-H(t))]$ - relative risk function. Obviously, the relative risk estimators $F_n(t)$ and $G_n(t) = 1 - (1 - H_n(t))^{1 - R_n(t)}$ of $F(t)$ and $G(t)$ satisfy the empirical analogy of equality $(1 - F(t))(1 - G(t)) = 1 - H(t)$, $-\infty < t < \infty$, that is $(1 - F_n(t))(1 - G_n(t)) = 1 - H_n(t)$, $-\infty < t < \infty$. But for exponential hazard estimators $F_n^E(t)$ and $G_n^E(t) = 1 - exp\{ - \sum_{j=1}^{n} (1 - \delta_{(j)}) I(Z_{(j)} \leq t)/(n - j + 1) \}$ of $F(t)$ and $G(t)$, we have

$$(1 - F_n^E(t))(1 - G_n^E(t)) = exp\Big\{ - \sum_{j=1}^{n} \frac{I(Z_{(j)} \leq t)}{n - j + 1} \Big\} \neq 1 - H_n(t).$$

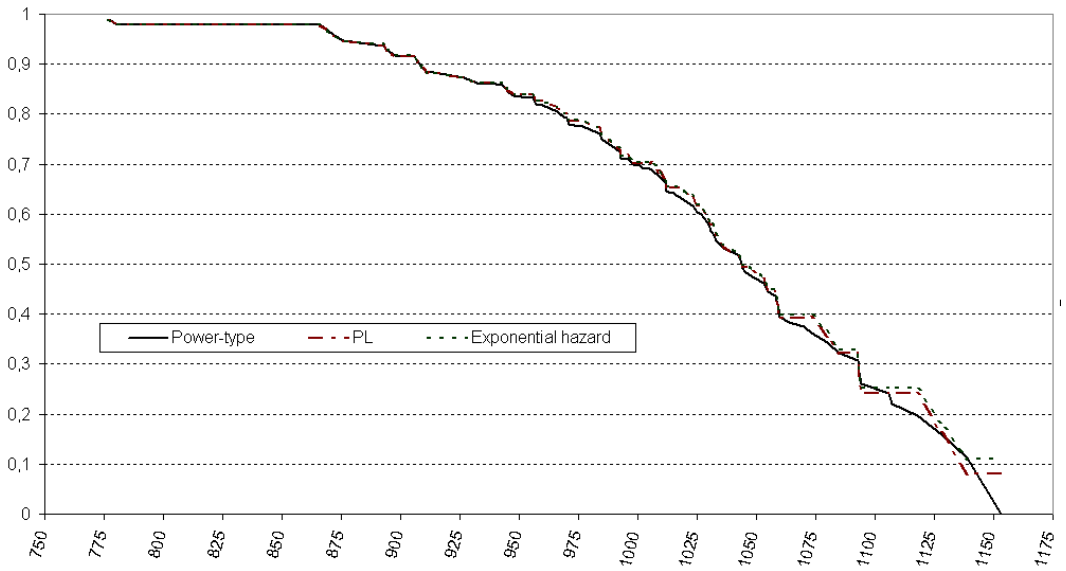
Moreover, for $t \geq Z_{(n)}$, $F_n(t) = 1$, but $F_n^E(t) < 1$. Therefore F_n is a correct estimator of continuous d.f. F than F_n^{PL} and F_n^E. In picture below we demonstrate plots of estimators $1 - F_n$, $1 - F_n^{PL}$ and $1 - F_n^E$ of survival function $1 - F$ using well-known Channing House data of size n=97(see [3-5]). Here, thin-solid line stands for $1 - F_n^E$, medium-one for $1 - F_n^{PL}$ and thick-solid line stands for $1 - F_n$. In monographies [3-5] of author one can find several extensions of estimators F_n, F_n^{PL} and F_n^E with full asymptotical results theory (weak convergence, law of itherated logarithm type strong consistency, weak and strong approximation, empirical Bayes approach ...) in competing risks models with random censorship from the right and both sides.

References

[1] Abdushukurov, A.A. (1998). Nonparametric estimation of the distribution function based on relative risk function. *Commun. Statist.:Theory and Methods*, **27**, N.8, pp. 1991-2012.

[2] Abdushukurov, A.A. (1999). On nonparametric estimation of reliability indices by censored samples. *Theory Probab. Appl.*, **43**, N.1, pp. 3-11.

[3] Abdushukurov, A.A. (2009). *Statistics of Incomplete observations: Asymptotical Theory of Estimation for Nonclassical Models.* University Press. Tashkent.



[4] Abdushukurov, A.A.(2011). Nonparametric estimation based on incomplete observation. International Enciclopedia of Statistical Sciences. Springer. Part 14. pp. 962-964.

[5] Abdushukurov, A.A.(2011). Estimation of unknown distributions from incomplete data and its properties. LAMBERT Academic Publishing.

[6] Akritas, M.G. (2000). The central limit theorem under censoring. *Bernoulli, 6,* N.6, pp. 1109-1120.

[7] Csörgő, S. (1996). Universal Gaussian Approximations under Random Censorship, *Ann. Statist.*, **24**, N.6, pp. 2744-2778.

[8] Gill, R.D. (1980). *Censoring and Stochastic Integrals*, Mathematical Centre Tracts, **124**. Amsterdam.

[9] Gill, R.D. (1994). Glivenko-Cantelli for Kaplan-Meier. *Math. Meth. of Statist.*, **3**, N.1, pp. 76-87.

[10] Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from Incomplete observations. *J.Amer.Statist.Assoc.*, **58**, pp. 457-481.

Estimation of Density from Indirect Observation

V.Solev
St.Petersburg Department of Steklov Mathematical Institute RAS
St.Petersburg, Russia
e-mail: solev@pdmi.ras.ru

Abstract

In this paper we consider an observational scheme, as data are interval censored. We suggest a simple nonparametric estimator $\widehat{f}_n$ for unknown density f and under some appropriate condition prove the consistency of this estimator.

Introduction

Let τ be a random partition of real line, that is such countable collection of intervals:

$$\tau = \{[a_j, b_j),\ j \in \mathbb{Z}\} \tag{1}$$

that

$$\mathbb{R} = \bigcup_{j\in\mathbb{Z}} [a_j, b_j),\ \text{and}\ [a_j, b_j) \cap [a_i, b_i) = \emptyset,\ \text{as}\ i \neq j. \tag{2}$$

For simplicity we assume that

$$0 \in [a_0, b_0),\ \text{and}\ a_{j+1} = b_j.$$

For a point x we denote by $[L(x), R(x))$ the interval of the partition τ such that

$$x \in [L(x), R(x))\,.$$

Let X be the random variable with density f. Denote by $\Delta(X) = [L(X), R(X))$ the interval of the partition τ which contains X. We assume that X and τ are independent. Suppose that instead of X we observe the interval $\Delta(X)$. The main goal in this paper is to construct an estimator $\widehat{f}_n$ for unknown density f of the distribution of X from indirect observations.

Namely, let $\Delta_1(X), \dots, \Delta_n(X)$ be independent copy of random interval $\Delta(X)$. We need to construct an estimator $\widehat{f}_n$ on observations $\Delta_1(X), \dots, \Delta_n(X)$. Here

$$\Delta_j(X) = [L_j(X), R_j(X))\,.$$

This problem vas investigated by Turnbull, B. W. (1976) and many other authors. In our paper we use approaches which were proposed in Huber-Carol, C. and Vonta, F. (2004), and Huber-Carol, C., Solev, V., and Vonta, F. (2006).

1 The density of random vector $[L(x), R(x))$

We assume that the distribution of random vector (a_j, b_j) has density $p_j(u, v)$. Clearly,

$$p_j(u, v) = p_j(u, v)\mathbb{1}_{(u,\infty)}(v),$$

where $\mathbb{1}_A(x)$ is the indicator function of the set A. Since

$$\mathbb{P}\{x \in [a_j, b_j)\} = \iint\limits_{u<v} \mathbb{1}_{[u,v)}(x)\, p_j(u, v)\, dudv,$$

then

$$\iint\limits_{u<v} \sum_{j\in\mathbb{Z}} \left\{\mathbb{1}_{[u,v)}(x)\, p_j(u, v)\right\}\, dudv = 1.$$

Therefore almost everywhere

$$\sum_{j\in\mathbb{Z}} \left\{\mathbb{1}_{[u,v)}(x)\, p_j(u, v)\right\} < \infty. \tag{3}$$

Let $\psi(u, v)$ be a nonnegative function. Let us calculate the value $\mathbf{E}\psi\left(L(x), R(x)\right)$. We have

$$\mathbf{E}\psi\left(L(x), R(x)\right) = \mathbf{E}\left[\sum_{j\in\mathbb{Z}} \psi\left(a_j, b_j\right) \mathbb{1}_{[a_j,b_j)}(x)\right] = \sum_{j\in\mathbb{Z}} \mathbf{E}\,\psi\left(a_j, b_j\right) \mathbb{1}_{[a_j,b_j)}(x).$$

It is clear, that

$$\mathbf{E}\psi\left(a_j, b_j\right) \mathbb{1}_{[a_j,b_j)}(x) = \iint\limits_{u<v} \psi(u, v)\mathbb{1}_{[u,v)}(x)\, p_k(u, v)\, dudv.$$

Therefore,

$$\mathbf{E}\psi\left(L(x), R(x)\right) = \iint\limits_{u<v} \psi(u, v) \left\{\sum_{j\in\mathbb{Z}} p_j(u, v)\mathbb{1}_{[u,v)}(x)\right\}\, dudv. \tag{4}$$

So, since (see (3)) for fixed x and almost for all $u \le x < v$ the value

$$\sum_{j\in\mathbb{Z}} p_j(u, v) < \infty, \tag{5}$$

then we obtain, that the random vector $[L(x), R(x))$ has density

$$p_x(u, v) = \sum_{j\in\mathbb{Z}} \left\{p_j(u, v)\mathbb{1}_{[u,v)}(x)\right\} \tag{6}$$

It is evident, that

$$p_x(u, v) = p(u, v)\mathbb{1}_{[u,v)}(x),$$

where the function $p(u, v)$ does not depend on x. The function $p(u, v)$ is called the basic density of the partition τ.

2 The density of random vector $[L(X), R(X))$

Now suppose that random variable X and partition τ are independent. Let $F(x)$ be the distribution function of X and $f(x)$ be the density function. For a nonnegative function $\psi(u, v)$ let us calculate the value $\mathbf{E}\psi\left(L(X), R(X)\right)$. Since X and τ are independent we have

$$\mathbf{E}\psi\left(L(X), R(X)\right) = \int \left\{\mathbf{E}\left[\psi\left(L(x), R(x)\right)\right]\right\} f(x)\, dx.$$

Since

$$\mathbf{E}\left[\psi\left(L(x), R(x)\right)\right] = \iint\limits_{u<v} \psi\left(u, v\right) p(u, v) \mathbb{1}_{[u,v)}(x)\, dudv,$$

and

$$\int \mathbb{1}_{[u,v)}(x)\, f(x)\, dx = F(v) - F(u),$$

we obtain

$$\mathbf{E}\psi\left(L(X), R(X)\right) = \iint\limits_{u<v} \psi\left(u, v\right) p(u, v) \left(F(v) - F(u)\right)\, dudv. \tag{7}$$

So, for the density $k(u, v)$ of random vector $[L(X), R(X))$ we have the relation

$$k(u, v) = p(u, v)\left(F(v) - F(u)\right) = p_*(u, v)\, \frac{F(v) - F(u)}{v - u}, \tag{8}$$

where

$$p_*(u, v) = p(u, v)(v - u). \tag{9}$$

3 The estimating of density

Let $X, X_1, \ldots, X_n, \ldots$ independent identically distributed random variables with common density f, and $\tau, \tau_1, \ldots, \tau_n, \ldots$ independent among themselves and independent on $X_1, \ldots, X_n, \ldots$ random partitions with common basic density $p(u, v)$. Suppose that random variables $X_1, \ldots, X_n, \ldots$ are inaccessible to observation. Instead of them we observe random vectors $W_j = (L_j(X_j), R_j(X_j))$. Here we denote by $[L_j(x), R_j(x))$ the interval of the partition τ_j which contains x. The problem consists in estimating of unknown function $f(x)$ on observations $W_1, \ldots, W_n$, as the basic density $p(u, v)$ is known.

For a point x and $\varepsilon > 0$ denote

$$A(x; \varepsilon) = \left\{x - \varepsilon \le L(X) \le x \le R(X) < x + \varepsilon\right\}. \tag{10}$$

Clearly,

$$\mathbb{P}\{A(x;\varepsilon)\} = \iint\limits_{x-\varepsilon\le u\le x\le v\le x+\varepsilon} p_*(u,v)\,\frac{F(v)-F(u)}{v-u}\,du\,dv,$$

and for small ε and smooth f

$$\mathbb{P}\{A(x;\varepsilon)\} \approx f(x) \iint\limits_{x-\varepsilon\le u\le x\le v\le x+\varepsilon} p_*(u,v)\,du\,dv.$$

It suggests to use as the estimator $\widehat{f}_n(x)$ for value $f(x)$ the ratio

$$\widehat{f}_n(x) = \frac{\mathbb{P}_n\{A(x;\varepsilon)\}}{\mu(x;\varepsilon)}, \tag{11}$$

where $\mathbb{P}_n\{A(x;\varepsilon)\}$ is the empirical version of $\mathbb{P}_n\{A(x;\varepsilon)\}$,

$$\mathbb{P}_n\{A(x;\varepsilon)\} = \frac{1}{n}\sharp\{j:\, x-\varepsilon\le L_j(X_j)\le x\le R_j(X_j)<x+\varepsilon\}, \tag{12}$$

and

$$\mu(x;\varepsilon) = \iint\limits_{x-\varepsilon\le u\le x\le v\le x+\varepsilon} p_*(u,v)\,du\,dv. \tag{13}$$

Here $\sharp\{A\}$ denotes the number of element of A.

By Hoeffding inequality for $y>0$

$$\mathbb{P}\{|\mathbb{P}_n\{A(x;\varepsilon)\} - \mathbb{P}\{A(x;\varepsilon)\}| > y\} \le 2e^{-2ny^2}. \tag{14}$$

Now we assume that function f satisfy to the condition

$$|f(x)-f(y)| \le C|x-y|. \tag{15}$$

Denote

$$f_\varepsilon(x) = \frac{1}{\mu(x;\varepsilon)} \iint\limits_{x-\varepsilon\le u\le x\le v\le x+\varepsilon} p_*(u,v)\,\frac{F(v)-F(u)}{v-u}\,du\,dv. \tag{16}$$

Here we suppose that for all x and $\varepsilon>0$

$$\mu(x;\varepsilon) > 0. \tag{17}$$

Lemma 1. *Under the conditions (15), and (17)*

$$|f_\varepsilon(x) - f(x)| \le 2C\varepsilon. \tag{18}$$

Proof. At the beginning we shall estimate the value

$$\left|\frac{F(v)-F(u)}{v-u}-f(x)\right|,$$

as $u \le x < v$. It is clear, that

$$\left|\frac{F(v)-F(u)}{v-u}-f(x)\right| \le \frac{1}{v-u}\int\limits_u^v |f(y)-f(x)|\,dy.$$

Hence, by (15) we obtain

$$\left|\frac{F(v)-F(u)}{v-u}-f(x)\right| \le \frac{C}{v-u}\int\limits_u^v |y-x|\,dy \le C(v-u).$$

Therefore,

$$|f_\varepsilon(x)-f(x)| \le \frac{1}{\mu(x;\varepsilon)} \iint\limits_{x-\varepsilon\le u\le x\le v\le x+\varepsilon} p_*(u,v)\left|\frac{F(v)-F(u)}{v-u}-f(x)\right| du\,dv \le$$

$$\frac{C}{\mu(x;\varepsilon)} \iint\limits_{x-\varepsilon\le u\le x\le v\le x+\varepsilon} p_*(u,v)\,(v-u)\,du\,dv \le 2C\varepsilon.$$

So, we obtain (18).■

It can be observed that

$$f_\varepsilon(x) = \mathbf{E}\widehat{f}_n(x) = \frac{\mathbb{P}_n\{A(x;\varepsilon)\}}{\mu(x;\varepsilon)}.$$

Thus, in accordance with (14)

$$\mathbb{P}\left\{\left|\widehat{f}_n(x)-f_\varepsilon(x)\right|>y\right\} \le 2e^{-2ny^2\mu^2(x;\varepsilon)}. \tag{19}$$

The inequality (19) suggests us to choose $\varepsilon=\varepsilon_n$ by this way:

$$\varepsilon_n \to 0, \ \text{ and } \ \sqrt{n}\,\mu(x;\varepsilon_n)\to\infty, \ \text{as } n\to\infty. \tag{20}$$

Theorem 1. *Suppose that value ε_n satisfy to the conditions (20). Then under conditions (15), and (17)*

$$\mathbf{E}\left|\widehat{f}_n(x)-f(x)\right| \to 0, \ \textit{as } n\to\infty. \tag{21}$$

Proof. At the beginning we shall prove that

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| \to 0, \text{ as } n \to \infty.$$

Together with lemma 1 this implies (21). It is known that

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| = \int\limits_0^\infty \mathbb{P}\left\{\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| > y\right\} dy.$$

The change of variable

$$2\sqrt{n}\, y\mu(x;\varepsilon) = s$$

in last integral gives

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| = \frac{1}{2\sqrt{n}\mu(x;\varepsilon)} \int\limits_0^\infty \mathbb{P}\left\{\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| > s/2\sqrt{n}\mu(x;\varepsilon)\right\} ds. \tag{22}$$

From the inequality (19) we deduce

$$\int\limits_0^\infty \mathbb{P}\left\{\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| > s/2\sqrt{n}\mu(x;\varepsilon)\right\} ds \le 2\int\limits_0^\infty e^{-s^2/2}\, ds = \sqrt{2\pi}.$$

Together with (22) this implies

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| \le \frac{\sqrt{2\pi}}{2\sqrt{n}\mu(x;\varepsilon)}$$

Since

$$\sqrt{n}\,\mu(x;\varepsilon_n) \to \infty, \text{ as } n \to \infty,$$

we obtain

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| \to 0, \text{ as } n \to \infty.$$

Acknowledgements

Research partially supported by Grant RFBR 11-01-00577-a, Grant RFBR-DFG 09-01-91331-a, Grant NSh-4472.2010.1.

References

[1] Alioum, A. and Commenges, D. (1996) A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, **52**, 512-524.

[2] Frydman, H. (1994) A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society, Series B*, **56**, 71-74.

[3] Huber-Carol, C., Solev, V., and Vonta, F. (2006) Estmation of density for arbitrarily censored and truncated data. In *Probability, Statistics and Modelling in Public Health*, Nikulin, M. S., Commenges, D., and Huber-Carol, C. eds., Springer (Kluwer Acad. Publ.), New York, pp. 246-265.

[4] Huber-Carol, C. and Vonta, F. (2004) Frailty models for arbitrarily censored and truncated data. *Lifetime Data Analysis*, **10**, 369-388.

[5] Kosorok, M., Lee, B., and Fine, J. (2004) Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.*, **32**, No. 4, 1448-1491.

[6] Slud, E. and Vonta, F. (2005) Efficient semiparametric estimators via modified profile likelihood. *Jour. Statist. Planning and Inference*, **129**, 339-367.

[7] Turnbull, B. W. (1976) The empirical distribution function with arbitrary grouped, censored and truncated data. *Journal of the Royal Statistical Society*, **38**, 290-295.

Part V
Application of Statistical Methods

Planning Seismic Networks

O. Omelchenko
Institute of Computational Mathematics and Mathematical Geophysics of SB RAS
Novosibirsk, Russia
e-mail: oko@omzg.sscc.ru

Abstract

This paper is dealt with a brief statement of a basis of the theory of optimal planning of seismic networks. Some concepts of such a planning of seismic networks are given. Some specific formulations of problems of planning of seismic networks are presented.

Keywords:optimal planning, seismic networks, hypocenters of earthquakes.

Introduction

A source of the primarily observed seismic data in seismology is an observation system (OS), i.e., a network of spatially distributed seismic stations equipped with instruments for recording seismic waves.

The basic OS parameters in seismology are as follows [1,2]: a). the number of seismic stations, the geometric configuration of a network, and individual station sites; b). the frequency responses of the recording instruments, their dynamic range and amplification.

Definition: We shall call a given number of seismic stations deployed at fixed sites a seismograph network (SN).

Seismograph networks are usually divided into several categories by their spatial dimensions: local networks ranging in size from a few hundred meters to a few tens of kilometers; zonal networks, from a few tens to a few hundreds of kilometers; regional networks, from a few hundreds to a few thousands of kilometers; global networks, which are deployed all over the world or a large part of it.

For the purpose of economy the number of stations in a network should be kept to a minimum without affecting the quality of records; this naturally calls for an optimum network design.

The modem worldwide tendency to optimize seismic networks is to minimize errors in determining basic kinematic hypocenter parameters. The mathematical planning of experiment uses methods of mathematical statistics and optimization techniques [3-5].

The problem of designing a SN has arisen from the parameter estimation problem Hypocenters of earthquakes, setting and methods of solution by which one reduce here briefly.

1 The hypocenter location problem

The basic data for this problem are the coordinates of existing seismic stations and those of possible sites for the new stations to be added to the network, the velocity structure of the

region under study, and the positions of the seismic-prone zones which are to be studied using the network in question. This information related to arrival times of the waves excited by an earthquake, is obtained by means of nonlinear equations of condition [5-7]:

$$\vec{T} = \vec{\eta}(X, \vec{\theta}) + \vec{\varepsilon}, \tag{1}$$

where $\vec{T} = (T_1, T_2, \ldots, T_N)^T$ is the arrival time vector, $\vec{\eta}(X, \vec{\theta})$ is the N-dimensional vector of theoretical arrival times or the regression function, $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T$ is the vector of residuals, $\vec{\theta} = (\varphi, \lambda, h, t)^T$ is the vector of estimated parameters, $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ is the matrix of stations' coordinates, N is the number of recorded arrival times, n is the number of stations.

The estimation of $\vec{\theta}$ is treated by the regression analysis: the solutions are the least squares (LS) estimates

$$\vec{\theta} = \arg\min_{\vec{\theta}\in\Omega} Q(\vec{\theta}), \qquad Q(\vec{\theta}) = \sum_{i=1}^{N} \sigma_i^{-2} \big(T_i - \eta(\vec{x}_i, \vec{\theta})\big)^2. \tag{2}$$

The functional $Q(\vec{\theta})$ has usually been minimized in seismology since Geiger's times using the iterative Gauss-Newton method with regularization based on a linear fit to the regression function around the point $\vec{\theta}^k$:

$$J(X, \vec{\theta}^k)\Delta\vec{\theta}^k + \vec{\eta}(X, \vec{\theta}^k) - \vec{T} + \vec{\varepsilon} = 0, \tag{3}$$

where

$$J(X, \vec{\theta}) = \left(\frac{\partial\eta(\vec{x}_i, \vec{\theta})}{\partial\theta_1}, \frac{\partial\eta(\vec{x}_i, \vec{\theta})}{\partial\theta_2}, \ldots, \frac{\partial\eta(\vec{x}_i, \vec{\theta})}{\partial\theta_m}\right), \quad i = 1, 2, \ldots, n. \tag{4}$$

Multiplying both parts of the linearized equations (3), by $J^T(X, \vec{\theta}^k)$ from the left, one obtains the following normal equations is:

$$(J^T(X, \vec{\theta}^k)J(X, \vec{\theta}^k) + \alpha^2 I)\Delta\vec{\theta}^k = J^T(X, \vec{\theta}^k)\vec{y}(X, \vec{\theta}^k), \tag{5}$$

where α - parameters of regularization, I - unique matrix, $\vec{y}(X, \vec{\theta}) = (\vec{T} - \eta(X, \vec{\theta}))^T$.

The estimates of $\vec{\theta}$ are found by iteration ($\vec{\theta} = \lim_{k\to\infty}\vec{\theta}^k$):

$$\vec{\theta}^{k+1} = \vec{\theta}^k + \Big[J^T(X, \vec{\theta}^k)J(X, \vec{\theta}^k)\Big]^{-1} J^T(X, \vec{\theta}^k)\vec{y}(X, \vec{\theta}^k), \quad k = 0, 1, 2, \ldots, \tag{6}$$

where the starting fit $\vec{\theta}^0$ should be chosen as close to the true values.

Another approach to solving (1)-(4) is to abstain from using normal equations, but solve the iterative process for (3) directly at each step. The most popular recent method to do this is singular value decomposition (SVD) or the generalized inversion [8, 9]. The computational scheme of the Gauss–Newton singular value decomposition is to decompose (4) into a product of three matrices at each step of the iterative process:

$$J(X, \vec{\theta}^k) = U_k\Sigma_k V_k^T, \tag{7}$$

where U_k is an orthogonal $n \times n$ matrix, V_k is an orthogonal $m \times m$ matrix, Σ_k is a diagonal $n \times m$ matrix having the structure $\Sigma_k = \binom{S_k}{0}$, where $S_k = diag(\rho_1, \rho_2, \ldots, \rho_m)$ is a diagonal matrix of singular values arranged in nonincreasing order $\rho_i \geq \rho_{i+1}$.

The method also provides a so-called singular value analysis, which consists in the elimination of zero singular values and the respective columns in U and V. The iterative process then becomes

$$\vec{\theta}^{k+1} = \vec{\theta}^{k} + V_k S_k^{-1} \vec{d}^{k}, \quad k = 0, 1, 2, \ldots, \tag{8}$$

where $\vec{d}^{k}$ is a vector which consists of the first m components of $U_k^T \vec{y}(X, \vec{\theta}^{k})$.

It can be shown that, at each step of the iterative process (11), the vector $\Delta\vec{\theta}^{k} = V_k S_k^{-1} \vec{d}^{k}$ minimizes not only the functional $Q(\vec{\theta})$, in which the vector $\vec{\eta}(X, \vec{\theta})$ has been replaced by its linear part as given by (4), but also the norm of the parameter vector, which ensures the uniqueness of the solution. The advantage of this process, as compared with (6), (9), is that one easily obtains as a side result not only the covariance matrix of the parameter space but also the matrix of the data space [13].

2 The necessity of planning SN

To sum up, using any of the above methods based on the Jacoby matrix (4) of the linearized equations of condition (3), (4), one can estimate earthquake hypocenter parameters and the associated uncertainties. The solution of this problem is discussed, for example, in [6, 10]. However good the iterative techniques (9), (11) for hypocenter location may be, they are unsatisfactory for a poorly conditioned matrix (4). The regularization techniques recommended for such cases often fail to give the desired effect in practice. The cause of matrix (4) being poorly conditioned lies in poor observational arrangements, namely, in poor network geometries with respect to seismic source zones. The necessary conclusion is that observations should be planned beforehand; that is, network geometries should be chosen so that matrix (4) should be as well conditioned as possible to improve in parameter estimation. The design of seismograph networks is thus to remove the cause why matrix (4) is poorly conditioned, rather than trying to mend the matters by using various regularizations.

It is now generally recognized in the theory of experimental design that well-advised preliminary planning is required for costly experiments (e.g., explosions) or experiments that cannot be reproduced (e.g., natural phenomena such as earthquakes).

All this shows once more that observational arrangements should be planned beforehand. This also concerns seismic networks.

3 A concept of plan, and network design

Definition: A design of an optimum seismic network is defined here as a network having a fixed

constant list of wave types recorded at each site. For example,

$$\xi_n = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n; \vec{P}, \vec{S}\}. \tag{9}$$

Definition: The SN design problem is to find an optimum design that satisfies one of the criteria of optimal planning [4].

4 Optimal design criteria.

The quality of a design is determined by applying certain criteria to it. These can be classified into statistical and nonstatistical criteria. There three statistical criteria (A-, D-, and E-) of an optimum design and one more nonstatistical C-criterion of an optimum design. They are used in seismology [3],[5],[7],[11],[13-15].

The criterion was formulated by Burmin [3] who elaborated Marchuk's ideas [12] as to the need for experimental design, provided the matrix of the normal equations has the highest condition number possible. We used this criterion in planning many seismic networks [5],[15].

The above criteria are related in one way or another to the matrix of the linearized equations of condition $J(\xi, \vec{\theta})$ (4). It would then be reasonable to carry out a theoretical search for designs that minimize a given function Ψ of matrix (4). As has been mentioned above, this optimization problem is the problem of optimal planning of seismic networks, the results of its solution being optimum designs (optimum seismic networks), and the function Ψ being the optimality criterion.

5 Specific formulations of seismic network design problems

The design of seismic networks in many seismological problems can be stated in either of the following formulations [5]: a). A region has a network of $k > m$, where m is the number of unknown hypocenter parameters. It is required to select an optimum subnetwork of $k_1 < k$ stations for recording the earthquakes occurring in a given source zone (or zones). b). A region has a network of $k \geq m$ stations,equipment being available (expected or planned) for k_2 stations more. There are $l > k_2$ sites in the region chosen from geological and geographical considerations for the deployment of this equipment. It is required to supplement the existing network in an optimum manner to have $k + k_2$ stations for recording the earthquakes occurring in a given source zone (zones). c). A region has no seismic network. It is required to plan (design) an optimum network consisting of a fixed number $k \geq m$ of stations for recording the earthquakes occurring in a fixed source zone (zones). The new network can be designed both using the sites that have been chosen beforehand from geological and geographical considerations and the sites that cover the region uniformly.

Each of the above problems divides in its turn into several subproblems, depending on what set of parameters is planned to be determined using the future network.

6 Analysis of seismic networks

An important constituent of SN design is a network analysis intended to characterize in quantitative terms the performance of a seismograph network in hypocenter location in given regions, to plot contour maps of confidence intervals for earthquake hypocenter parameters and find correlations between them based on the elements of the covariance matrix of the parameter space, and plot similar contour maps of the elements of the covariance matrix of the data space characterizing an excess or deficit of the information provided by the network. To sum up, SN analysis can provide conclusions on hypocenter location capability for earthquakes occurring in different source zones and on the importance of an individual stations in the network, these parameters being then used to compare different networks.

7 Planning seismic networks in the Kemerovo region

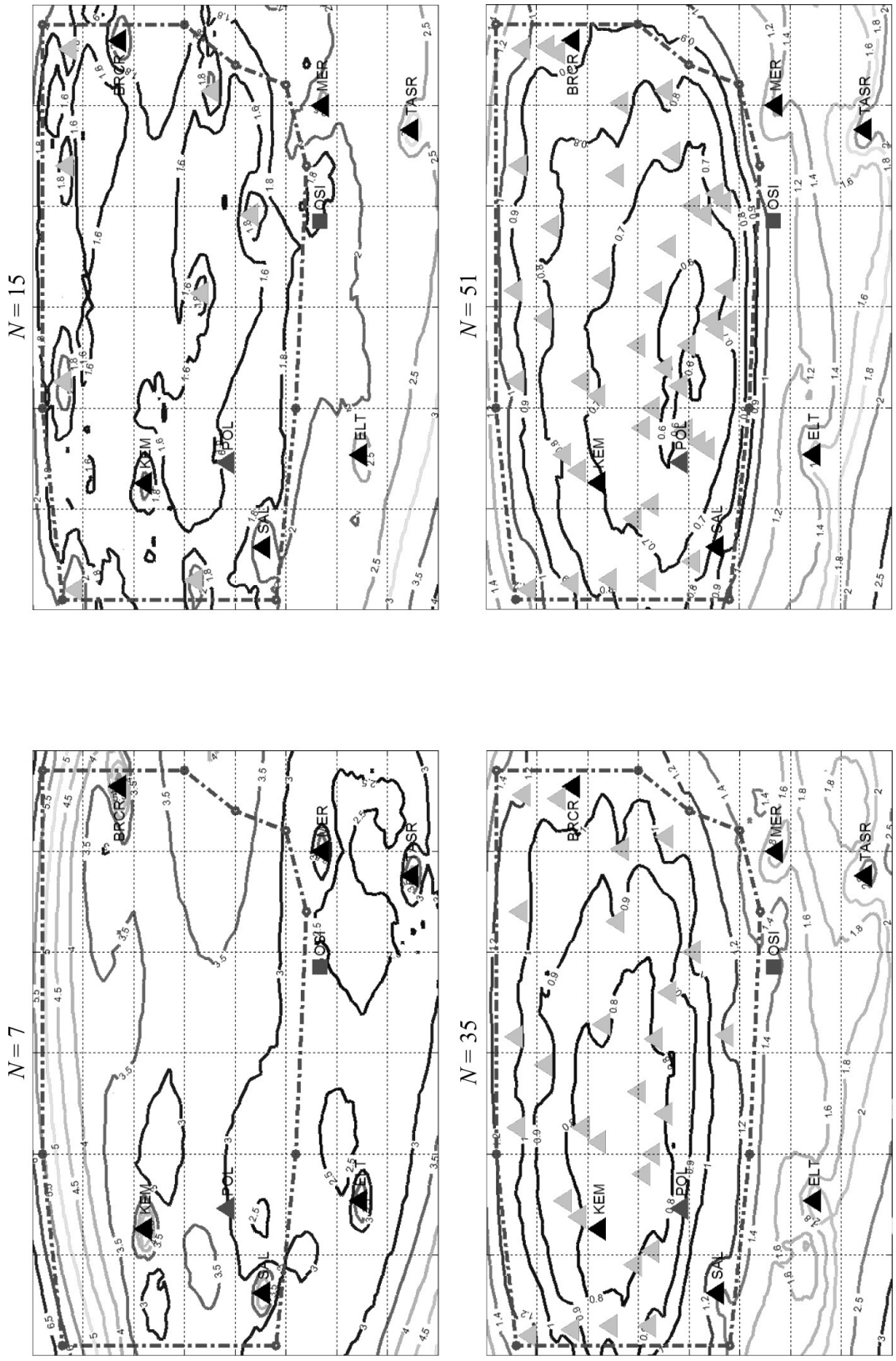
The objective of this research is planning of zonal network of seismic stations (NSS) which could provide the solution to the following tasks [15]: 1). Seismic monitoring of a section of intensive man-caused impacts actions (on an example of the city of Kemerovo and adjoining territories) at a "weak seismicity" level, i.e., under the condition of the lower limit of representativeness of seismic events not exceeding 4-5 energy classes. 2). Recognition of records of industrial explosions and earthquakes. 3). Estimation of a depth of the focus of seismic events independent of their nature.

The problem of planning is currently actual due to the seismic activation near to the towns of Osinniki (2005) and Polysaevo (2007). With a well-planned zonal network of observation, it would be possible to trace the previous history of the seismic activity, to calculate more reliably its nature, and to assess until very recently consequences at various scenarios of its development.

With allowance for the fact that Kuznetsk Basin according to the current maps of seismic zoning is attributed to 7-8-magnitude territories by the level of seismic activity, it is impossible to exclude a possibility of natural tectonic activity, along with the man-caused factors. With network capable of fixing even minor changes in seismic conditions, it is possible to exceed the detection, say, of foreshocks activity followed by a strong seismic event. In the case of a man-caused activity, it is apparently possible to correct the factors which have caused it in due time.

In this paper, the current regional sub-net, consisting of 7 stations, is optimally supplemented with the planned NSSs containing 15, 20, 25, 30, 35, 40, 45 stations. Seismic stations of the existing network are installed in Salair, Kemerovo, Mezhdurechensk, Berchikul, Tashtagol, Eltsovka, and Verkh-Baza (they are marked with black triangles in Figure 1).

Optimal addition is achieved within 44 stations in predetermined built-up areas (they are marked with grey triangles in Figures). A-optimum additions of 7 existing regional stations to 15, 20, 25, 30, 35, 40, 45 stations are received. For each of the received networks its detailed analysis consisting in build-up of maps of isolines of energy representation, confidential radius of an error of epicentres and a confidential interval of an error on depth is carried out. Averages of the termed quantities on all observed area and the selected hypocentral zone for each received



Maps of confidence radius of the epicenter isolines for the network of N stations

Figure 1: The level of confidence radius at the epicenter on all observed area.

network are computed also. All it allows to select for realization a subnet of certain number of the stations, corresponding to available material possibilities. The value of 0.05 s for the accuracy of recording arrival times and the value of 0.1 km/s for an error of speed of seismic waves velocity were chosen for all the planned stations.

Figure 1 shows the level of confidence radius at the epicenter for networks of 7, 15, 35, and 51 stations. From these figures it is seen that its mean in the zone sharply decreases and is equal to 3.6, 1.7, 1, and 0.8 km, respectively.

The above-mentioned characteristics of the regional network and of all the planned networks a change of characteristics in the number of network stations is demonstrated with the diagrams shown in Figure 2.

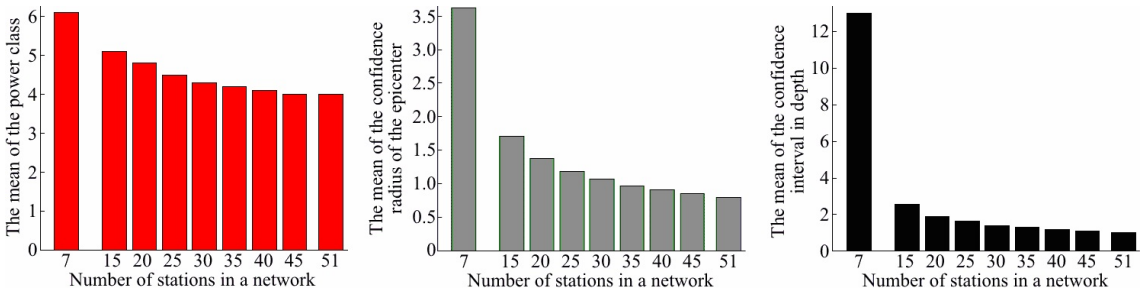


Figure 2: A change of characteristics in the number of network stations.

Analyzing Figures 1-2 , it is possible to make a conclusion that all the planned networks in all their parameters essentially surpass the regional network. Beginning with a network of 35 stations, their characteristics can provide a solution of the tasks stated.

References

[1] Nikolaev A.V. (1977). *Optimization of seismic networks and automation of seismic observations*. Metsniereba, Tbilisi: pp. 81-89 (in Russian).

[2] Kondorskaya N.V., Aranovich Z.I. (1971). Methodical basis of optimization of seismic observation. *Fizika Zemli*. *N^o* **7**. pp. 14-30 (in Russian).

[3] Burmin V.Yu. (1976). The problem the experiment design and condition of a system of linear algebraic equations. *Tekhnicheskaya Kibernetika*. *N^o* **2**, pp. 195-200 (in Russian).

[4] Fedorov V.V. (1971). *The Theory of the Optimum Experiment*. Nauka, Moscow. (in Russian).

[5] Omelchenko O.K., Gusyakov V.K. (1996). Planning of seismic networks for tsunami prediction. *Volcanology and Seismology*. Vol. **18**. pp. 211-230 (in Russian).

[6] Omelchenko O.K., Beloborodov V.N. (1983). Modeling of seismic wave fields. *Mathematical Problems of Geophysics*. Novosibirsk: VTs SO AN SSSR, pp. 71-81 (in Russian).

[7] Omelchenko O.K., Beloborodov V.N. (1986). *On optimal design of a regional seismic network*. (Preprint / VTs SO AN SSSR; 668) Novosibirsk, (in Russian).

[8] Forsythe G.E., Malcolm M.A., Moler C.B. (1977). *Computer Methods for Mathematical Computations*. Englewood Cliffs, N.J.

[9] Hatton L., Worthington M., Makin J. (1986). *Seismic Data Processing. Theory and Practice*. Blackwell, Oxford.

[10] Lukk A.A., Pivovarova N.., and Pukhnacheva T.P. (1974) *Matematicheskie problemy geofiiiki (Mathematical problems of geophysics)*. **5, Part 1** (Novosibirsk: VTs SO AN SSSR,). pp 159-167.

[11] Kijko A. (1978). *Methods of the optimal planning of regional seismic networks*. Polish Academy of Sciences. Institute of Geophysics. A-7 **(119)**. Panstwowe wydawnictwo naukowe, Warszawa-Lodz.

[12] Marchuk G.I. (1977). *Methods of Computational Mathematics*. Nauka, Moscow (in Russian).

[13] Uhrhammer R.A. (1982). The optimal estimation of earthquake parameters. *Physics of the Earth and Planetary Interiors*. Vol. **30**. pp. 105-118.

[14] E.F. Savarenskiy, V.V. Safronov, A.B. Peshkov, et al. (1979). *Fiuka Zemli*. N **8**. pp. 64-71.

[15] A.F. Emanov, O.K. Omelchenko. (2010). Planning seismic networks in the Kemerovo region. *Bulletin of the Novosibirsk Computing Center, series: Mathematical Modeling in Geophysics*. issue: **13**, pp. 31-40.

Practical Application of Forecasting Method Based on Universal Measure*

P. Pristavka[1], B. Ryabko[2]
Siberian State University of Telecommunications and Information Sciences
Novosibirsk, Russia
e-mail: [1]ppa@ngs.ru, [2]boris@ryabko.net

Abstract

In this article we describe and experimentally investigate a method to construct forecasting algorithms for stationary and ergodic processes based on universal measures (or the so-called universal data compressors). By the example of predicting the sunspot numbers and some other solar characteristics we show that the precision of thus obtained predictions is higher than for known methods.

Keywords: time series, nonparametric methods, universal measure, universal coding, solar activity, sea level, cross-rates.

Introduction

The problem of forecasting is important for many applications. In this paper we develop and experimentally investigate an efficiency of prediction methods which are based on so-called universal measures (or universal data compressors).

By definition, universal codes, or universal methods of lossless data compression, are intended to "compress" texts, i.e. encode them in such a way that the length of an encoded text is shorter than the length of the initial one (and, of course, the original text can be recovered from the encoded one). It is important that the text statistics is unknown beforehand, that is why such codes are called universal. In fact, universal codes implicitly estimate unknown characteristics of the processes and use them for data compression. The universal codes for stationary and ergodic sources with finite alphabets have been known since 1980's; see [4].

It is clear that universal codes can be considered as a tool of mathematical statistics and it is natural to try to apply them for solving traditional problems of this science (like hypothesis testing, parameter estimation, etc.) First it was recognized in 1980's (see [3] and [5]) and nowadays it is shown that the universal codes can be efficiently used for hypothesis testing, parameter estimation and prediction of time series with finite and real-valued alphabets; see [6], [7]. However, there are only preliminary results which concern practical applications of prediction methods to real data, see [8]. These results were rather aimed to illustrate the possibility of such applications than give information about the precision of the methods.

*Research was supported by Russian Foundation for Basic Research (grant № 09-07-00005.) and federal target scientific-technical programme (governmental contract no. 02.740.11.0396)

The goal of this work is the construction, implementation and experimental estimation of the methods of prediction suggested, which are based on universal codes and deeply connected with them so-called universal measures. To this aim, we have considered several types of time-series data: the indices of solar activity (SA), water level, and cross-rate of certain currency. It is clear that such processes are of great theoretical and practical importance. For example, nowadays the statistical connections between climate and SA are widely investigated.

The obtained experimental results show that the forecasting methods based on universal codes possess a high precision.

The outline of the paper is as follows. The next part contains necessary definitions and some information about universal codes and measures. The part three is devoted to prediction of real time series and the part four is a short conclusion.

1 Method and its implementation

1.1 Description of the method

Here we will describe the forecasting method studied, and also provide required theoretical information. It will be convenient at first to describe briefly the prediction problem. We consider a stationary and ergodic source which generates sequences $x_1x_2...$ of elements (letters) from some set (alphabet) A, which is either finite or real-valued. It is supposed that the probability distribution (or distribution of limiting probabilities) $P(x_1 = a_{i_1}, x_2 = a_{i_2}, ..., x_t = a_{i_t})$ (or the density $p(x_1, x_2, ..., x_t)$) is unknown. Let the source generate a message $x_1, ...x_{t-1}x_t$, $x_i \in A$ for all i, and the next letter needs to be predicted.

Now let us describe the forecasting method. In what follows we give a definitions of universal code and universal measure and describe the connection between them.

Roughly speaking, a code maps words from the set A^t, $t \geq 1$, into the set of words over alphabet $\{0, 1\}$. By definition, a code U is universal (for the set of stationary ergodic sources), if for any stationary and ergodic source P the following equalities are valid:

$$\lim_{t\to\infty} |U(x_1...x_t)|/t = H(P),$$

with probability 1, and

$$\lim_{t\to\infty} |E_P(U(x_1...x_t)|)/t = H(P),$$

where $|v|$ is the length of the word v, $E_P(f)$ is the mean of f with respect to P, $H(P)$ is the Shannon entropy of P, i.e.

$$H(P) = \lim_{t\to\infty} -t^{-1} \sum_{u\in A^t} P(u) \log(P(u)).$$

A universal code is called optimal if it encodes a sequence of letters generated by a finite-alphabet source in such a way that the length of the encoded sequence is asymptotically minimal.

By definition a measure μ is universal if for any stationary and ergodic source P the following equalities are valid:

$$\lim_{t\to\infty} t^{-1}(\log P(x_1...x_t) - (\log \mu(x_1...x_t))) = 0,$$

with probability 1, and

$$\lim_{t\to\infty} t^{-1} \sum_{u\in A^t} P(u) \log(P(u)/\mu(u))) = 0.$$

(Here and below $\log x = \log_2 x$.) These equations show that, in a certain sense, the measure μ is an estimate of (unknown) measure P. That is why the universal measures can be used for estimation of process characteristics and prediction.

The following statement shows that any universal code determines a universal measure.

Theorem 1. *Let U be a universal code and*

$$\mu_U(w) = 2^{-|U(w)|} / \sum_{u\in A^{|w|}} \mu_U(u).$$

Then μ_U is a universal measure.

(The simple proof of this theorem can be found in [9]. So, we can see that, in a certain sense, the measure μ_U is a consistent (nonparametric) estimate of the (unknown) measure P.

Now we a going to describe the universal measure R which will be used as a basis for forecasting in this paper. For this purpose we first describe the Krichevsky measure K_m, which is universal for the set of Markov sources of memory, or connectivity, m, $m \le 0$, if $m = 0$, the source is i.i.d. In a certain sense this measure is optimal for this set (see for details [2], [9].) By definition,

$$K_m(x_1...x_t) = \begin{cases} \frac{1}{|A|^t}, & \text{if } t \le m; \\ \frac{1}{|A|^m} \left(\frac{(|A|/2)}{(1/2)^{|A|}}\right)^{|A|^m} \prod_{v\in A^m} \frac{\prod_{a\in A}((v_x(va)+1/2))}{((\bar{v}_x(v)+|A|/2))}, & \text{if } t > m, \end{cases} \tag{1}$$

where $x = x_1...x_t$, $\nu_x(v)$ is the count of word v , occurring in the sequence $x_1...x_{|v|}$, $x_2...x_{|v|+1}$, ..., $x_{t-|v|+1}...\ x_t$, $\bar{\nu}_x(v) = \sum_{a\in A} \nu_x(va)$, $\Gamma()$ is the gamma function.

We also define a probability distribution $\{\omega = \omega_1, \omega_2, ...\}$ on integers $\{1, 2, ...\}$ by

$$\omega_1 = 1 - 1/\log 3,\ ...\ ,\ \omega_i\ = 1/\log(i+1) - 1/\log(i+2),\ ...\ , \tag{2}$$

$i = 1, 2, ...$. (In what follows we will use this distribution, but the theorem described below is true for any distribution with nonzero probabilities.)

The measure R is defined as follows

$$R(x_1...x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1...x_t)\,. \tag{3}$$

It is important to note that the measure R is a universal measure for the class of all stationary and ergodic processes with a finite alphabet; [4]. Hence, R can be used as a consistent estimator of probabilities.

Let us describe the scheme of the forecasting method based on the measure R for the sequences generated by the sources of different types.

1.2 Finite-alphabet case

As we mentioned above the measure R can be applied for prediction. More precisely we may use R for defining the following conditional probability as:

$$R(a|x_1...x_t) = R(x_1...x_t a)/R(x_1...x_t)\,,$$

$a \in A$. In the finite-alphabet case the scheme of the prediction algorithm is quite simple. Let $x_1...x_t$ be a given sequence. For each $a \in A$ we construct the sequence $x_1...x_t a$ and compute the value $R(a|x_1...x_t)$. Having the set of such conditional probabilities we use them as estimations of the unknown probabilities $P(a|x_1...x_t)$, $a \in A$.

1.3 Real-valued case

Let (Ω, F, P) be a probability space and let $X_1, X_2, ...$ be a stochastic process with each X_t taking values in a standard Borel space. Suppose that the joint distribution P_n for $(X_1, X_2, ..., X_n)$ has a probability density function $p(x_1 x_2 ... x_n)$ with respect to the Lebesgue measure L. (A more general case is considered in [7]. In particular, it is shown that any universal measure can be used instead of R.) Let Π_n, $n \geq 1$, be an increasing sequence of finite partitions Ω that asymptotically generates the Borel sigma-field F, and let $x^{[k]}$ denote the element of Π_k that contains the point x. For integers s and n we define the following approximation of the density:

$$p^s(x_1, ..., x_n) = P(x_1^{[s]}, ..., x_n^{[s]})/L(x_1^{[s]}, ..., x_n^{[s]})\,.$$

Now we define the corresponding density r as follows:

$$r(x_1...x_t) = \sum_{s=1}^{\infty} \omega_s R(x_1^{[s]}, ..., x_t^{[s]})/L(x_1^{[s]}, ..., x_t^{[s]})) \tag{4}$$

It is shown in [7] that the density $r(x_1...x_t)$ estimates the unknown density $p(x_1, ..., x_t)$, and the conditional density

$$r(a|x_1...x_t) = r(x_1...x_t a)/r(x_1...x_t) \tag{5}$$

is a reasonable estimation of $p(a|x_1...x_t)$.

1.4 Implementation of the method

Consider next some aspects of the implementation of the investigated method. Suppose there is a certain source which generates values from some real-valued interval $[A; B]$ and we have time series $x_1...x_t$ generated by this source. The next value x_{t+1} needs to be predicted. For the purpose of simplicity we will consider computations based on r.

Step 1. Calculate $r(x_1...x_t)$ using (4). We will describe this step in more details. Divide the interval $[A; B]$ into two equal partitions, called bins, and transform $x_1...x_t$ into a sequence of symbols each of which is equal to the index of the bin, that contains the appropriate point x_i. (I.e. if 0 and 1 are the bins indexes, we will then obtain from $x_1...x_t$ the sequence consisting only of these symbols). Then calculate the first item of the sum from the formula (4). As the value of L we take a product of lengths of all bins containing x_i , and the measure R is computed for the sequence obtained as a result of transforming $x_1...x_t$ for the current quantization. After that again we divide each of the existing bins (there are two) into two equal bins (there will be four) and for this new quantization do analogous operations to compute the second term from the right-hand side of the equation (4). Go on in this way until getting the quantization for which each of the distinct time series values (including those added at the next step) belongs to distinct bins. Summing all terms, obtain $r(x_1...x_t)$ from (4). It should me mentioned that at this step any other algorithm for achieving the increasing sequence of finite partitions of $[A; B]$ may be applied.

Step 2. Consider the set P consisting of points A,$A + h$, $A + 2h$, ..., B , where h is a certain small constant. (In this paper $h = 0.01$ was used.) For each element a of this set we construct the sequence $x_1...x_ta$ and compute the value $r(x_1...x_ta)$ similarly. Here we use the quantization that is the same as for previous step. Then by formula (5) for each element from the set P calculate the estimations of the appropriate conditional probabilities and find the corresponding prediction. In this paper the forecast value was considered to be equal to the element from P with the biggest estimate of conditional probability. But any other adequate approach may be used.

2 Experimental results

In this section the results of experimental estimation of the method, described in the previous section, are given. The computations in this section can be divided into the two independent parts. In the first part we considered only the forecasting method, while in the second part some preprocessing was used.

2.1 Simple time series forecasting

As the target here we chose the time series consisting of the following indices: monthly and smoothed monthly means of sunspot numbers, absolute daily and monthly solar flux values. All datasets used in the experiments of this subsection, can be found at the National Geophysical

Data Center (NGDC) Internet site in the "Space Weather & Solar Events" section [10] . In this subsection one-step ahead forecasting was considered.

The content of each experiment can be described as follows. Given t successive values of certain time series, we tried to forecast its $(t+1)$th element. For each length of the time series related to a certain process there were 25 experiments on independent datasets. After doing all appropriate computations, the precision estimation was made by considering the differences between the forecast and actual values. The results of the experimental computations of this stage are given in Table 1 below. The first column contains the name of the investigated time series, the second gives the range of its values. The rest of the columns contain the mean absolute error (MAE), obtained when using the corresponding length of the time series. The "n/d" ("no data") text means that no calculations were carried out, because there was no data. For example, the information in the third row of the table indicates that for the time series on absolute daily solar flux the MAE corresponding to the calculations based on the 4000 known values, is 1.45. All time series values belong to the range [50; 300].

Table 1: Experimental results for SA prediction

Time series	Range	500	700	1000	1200	2000	3000	4000
Monthly sunspot number means	[0; 256]	6,54	2,56	9,58	15,85	21,7	19,63	n/d
Smoothed monthly sunspot number means	[0;210]	1,5	1,1	1,99	0,77	3,36	2,56	n/d
Absolute daily solar flux	[50;300]	1,17	1,17	2,71	5,52	8,35	1,72	1,45
Absolute monthly solar flux	[580;2540]	211,29	45,88	n/d	n/d	n/d	n/d	n/d

2.2 Forecasting with preprocessing of time series.

It is known that statistical methods of forecasting are often used with the prior transformation of input data. The combination of this approach and the investigated method may improve the precision of prediction. In this subsection experimental estimation of the application of the forecasting method based on the measure R together with the preparatory differentiation of the time series was accomplished. Having such forecast value we can easily add it to the last known elements of time series and obtain the desired prediction.

One-step ahead forecasts comparison. Here as the target of research time series, consisting of the 15-minute sea level indices, were considered. All datasets used in the experiments of this item, can be found at the British Oceanographic Data Centre (BODC) Internet site in the "UK

Tide Gauge Network" section [11]. The web interface of the site allows any registered user to download data file with the history of 15-minute sea level indices for the certain gauge from some location. In this file for each timestamp for the certain period of time the appropriate actual and residual values are given. The latter are calculated from the observed sea level values minus the predicted sea level values. According to the provided by BODC support information predicted tide values are produced at the National Oceanography Centre's (NOC), using their harmonic tidal analysis. This is based on the TIRA tidal analysis programs following the Doodson method. We considered the data related to the Bubbler tide gauge from Dover and captured in 2005 year, because all values for that period are available and there are no missed or interpolated points there. Again the one-step ahead forecasting was decided here. Furthermore, the scheme of the calculations was the same, but each estimation series contained 30 experiments. After the accomplishing of all appropriate computations, the precision estimation was made by considering the differences between corresponding forecast and actual values. As we mentioned above the history contains residuals for every element of time series. So, in the purpose of the objectivity for each MAE of NOC software we took into account residuals corresponding only to the values predicted by investigated method. The results of the investigation are situated in Table 2.

Table 2: Forecasting results 15-minutes sea level indices

Forecast	500	1000	2000	5000
R	0.034	0.038	0.0396	0.037
NOC	0.207	0.2133	0.09883	0.0779

The second row contains MAE for R measure based method and the appropriate input data. The third row includes MAE for computations produced at the NOC. So it is seen from the table that for the forecast using the described method and the time series of size 5000 the MAE is 0.037 for 30 experiments. And if we consider the given by NOC residuals for each of that 30 predicted by the investigated method values we will obtain 0.0779.

Comparison with the simplest method The simplest test to explore a new forecasting method is a comparison with so-called inertial prediction, where the last actual value is supposed to be the next forecast value. Here we will accomplish this kind of comparison, regarding the FOREX cross-rates daily currency of Great Britain pound to Euro. We used data for cross-rates from January 2001, 03 to January 2011, 17. The dataset used in this investigation item was taken from FXHISTORICALDATA.COM [1]. Again we considered the one-step ahead forecasting. Each estimation series contained 10 experiments. After the accomplishing of all appropriate computations, the precision estimation was made by considering the differences between corresponding forecast and actual values. The comparison results are summarized in Table 3.

Table has the following structure. The second and the third rows contain the MAE for forecasting method based on R measure and inertial prediction respectively when using the appropriate length of input data. As a whole we may say that the forecasting method based on the universal measure R showed the better than the inertial prediction results.

Table 3: Comparison with the inertial method for GBP/EUR cross-rate

Foracast	500	1000
R	0.0019	0.00089
Inertial	0.0025	0.0010

Conclusions

In this article the implementation and experimental estimation of forecasting method, based on the universal measure R were considered. Analysis of the investigation outcomes has shown the quite high precision of the obtained results. Good results were also achieved in the combine with preparatory transformation of time series. We found parameters for which the consistent superiority of the considered method in comparison with UK NOC prediction was detected in the one-step ahead forecasting. As a result we may conclude that universal codes are believed to be the effective tool for the forecasting methods construction in practical application.

References

[1] FXHISTORICALDATA.COM: URL: http://www.fxhistoricaldata.com/

[2] Krichevsky R. (1993). *Universal Compression and Retrival*, Kluver Academic Publishers, Norwell, MA

[3] Rissanen J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, Vol. **30**, pp. 629-636.

[4] Ryabko B. (1984). Twice-universal coding. *Problems of Information Transmission*, Vol. **20**, pp. 173-177.

[5] Ryabko B. (1988). Prediction of random sequences and universal coding. *Problems of Inform. Transmission*, Vol. **24**, pp. 87-96.

[6] Ryabko B., Astola J. (2006) Universal Codes as a Basis for Time Series Testing *Statistical Methodology*, Vol. **3**, pp. 375-397.

[7] Ryabko B. (2009). Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series. *IEEE Transactions on Information Theory*, Vol. **55**, pp. 4309-4315.

[8] Ryabko B., Monarev V. (2005). Using Information Theory Approach to Randomness Testing. *Journal of Statistical Planning and Inference*, Vol. **133**, pp. 95-110.

[9] Ryabko B. (2010). Applications of Universal Source Coding to Statistical Analysis of Time Series. In: Isaac Woungang et al. (Eds.), *Selected Topics in Information and Coding Theory*, World Scientific Publishing, pp. 289 - 338. (see also http://arxiv.org/pdf/cs/0701036v2)

[10] Space Weather & Solar Events. National Geophysical Data Center: URL: http://www.ngdc.noaa.gov/stp/spaceweather.html

[11] UK National Tide Gauge Network. British Oceanographic Data Centre: URL : $https$: $//www.bodc.ac.uk/data/online_delivery/ntslf/$

Applying Statistical Methods to Text Steganography

I. Nechta, A. Fionov
Siberian State University of Telecommunication and Information Sciences
Novosibirsk, Russia
e-mail: www@inbox.ru

Abstract

This paper presents a survey of text steganography methods used for hiding secret information inside some covertext. Widely known hiding techniques (such as translation based steganography, text generating and syntactic embedding) and detection are considered. It is shown that statistical analysis has an important role in text steganalysis.

Keywords: Steganography, steganalysis, linguistic stegosystem, statistical attacks.

Introduction

Steganography is the art and science of writing hidden messages in such a way that no one, apart from the sender and intended recipient, suspects the existence of the message. In steganography, it is very important to find a good covertext suitable for embedding hidden messages. This paper provides a basic introduction to steganography and steganalysis, with a particular focus on text steganography. Information hiding techniques are discussed, providing motivation for moving toward text steganography and steganalysis.We will show some of the problems inherent in text steganography as well as issues with existing solutions.

1 Steganography

In 1984, Gustavus Simmons illustrated what is now widely known as the prisoners' problem: Let us consider Fig1. two accomplices in a crime, Alice and Bob, are arrested in separate cells. They want to coordinate an escape plan, but their only means of communication is by way of messages conveyed for them by Wendy the warden. Should Alice and Bob try to exchange messages that are not completely open to Wendy, or ones that seem suspicious to her, they will be put into a high security prison. Alice and Bob will have to deceive the warden by finding a way of communicating secretly in the exchanges. It can be done such way: Alice gets any text (covertext) which does not arise the warden suspicion and embeds (using steganographic method) secret message into it. Then she sends the covertext with message to Bob. This covertext is available to both Warden and Bob, but it contains different information to Wendy than to Bob.

Many types of covertexts are based on data having redundancy, such as video, audio, or image files. In this article we'll discuss one of the areas of steganography, which uses text files as a covertext. A file with covertext is called container.

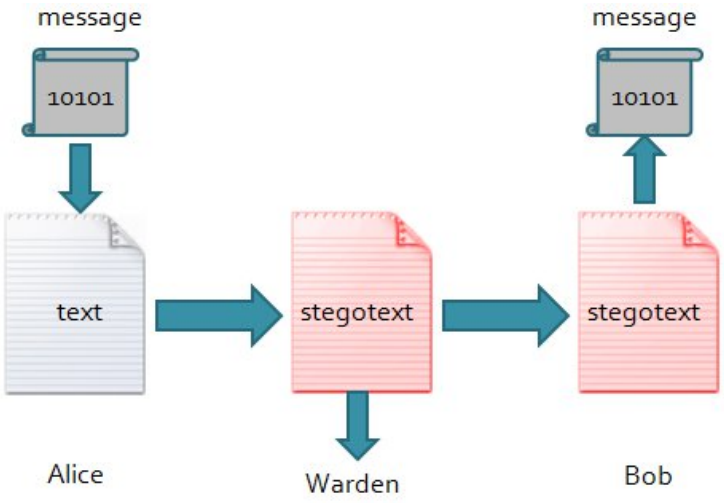


Figure 1: Basic steganography protocol

2 Information hiding methods

Existing methods of embedding secret messages in the text data could be divided into three groups:

1. **Syntactic methods** The early methods of text information hiding are based on the physical formatting of text. One of such methods, for example, proposed in [1], uses the extra space between words. One space means that the transmitted information bit is "0", and two spaces mean "1". This technique is widely used in HTML files (web pages) because space presence does not affect on the web page appearance.The disadvantage of this method is easy detectability, extra spaces are not used in text. It is possible to use special characters instead of spaces wich do not appear in commonly used text editors.

   Another method proposed in [1] uses a syntax error when writing words such as:
   "This is the end"
   "This iz the end"
   The second version has a misprint. The presence of errors in certain words (in particular "is") means that the transmitted information bit is "0", and errors absence means that bit is "1" . Thus, there is a transfer of information in the text. This method is not easily detectable, because some errors may occur in the message.

2. **Semantic methods** This group includes Tyrannosaurus Lex(T-Lex), published in [2], which uses the replacement words in the sentence on their synonyms, for example:

   Message embeded by synonym selection. Sentence *"Tobolsk is a decent little town"* contains message - *"01"*. This method requires a large synonyms dictionary. The below examples

	excellent		**city**
Tobolsk is a	(0) decent (1) fine	little	(0)metropolis (1) town

illustrate two shortcomings of the T-Lex system. First, it sometimes replaces words with synonyms that do not agree with correct English usage, as seen in the phrase *"soon subsequently dispatched"*. Second, T-Lex also substitutes synonyms that do not agree with the genre and the author style of the given text.

An invitation to dinner was soon afterwards dispatched
An invitation to dinner was soon subsequently dispatched

. . . and make it still better, and say nothing of the bad belongs to you alone.
. . . and make it still better, and say nada of the bad belongs to you alone.

It is clear that the word *"nada"* does not belong to Jane Austen's style. Furthermore, the string *"say nada"* of is not part of typical English usage.

There is another approach, proposed in [3], of generating sentence level paraphrases for information hiding. Example:
The caller identified the bomber as Yussef Attala, 20, from the Balata refugee camp near Nablus.
The caller named the bomber as 20-year old Yussef Attala from the Balata refugee camp near Nablus.
This method has a high degree of secrecy.

3. Linguistically-driven generation methods

Let's consider the method proposed in [4], using a context-free grammar to generate a natural like text.

Grammar Rules:
$S \rightarrow ABC$
$A \rightarrow She(0) \mid He(1)$
$B \rightarrow likes(0) \mid hates(1)$
$C \rightarrow milk(0) \mid apples(1)$

This approach produces stegotext that looks similar to the real structure of the original text. It is used a set of grammatical rules to generate stegotext and the choice of each word determines how secret message bits are encoded. The quality of the resulting stegotext directly depends on the quality of the grammar. Today's most popular stegosystems are Nicetext [5], Texto [6] and Markov-Chain-Based [7], because they have high ratio of the input message size to the generated text size. Also, resulting stegotext looking like natural text but it should be noted that, as usually, such text is meaningless.

The next one approach was proposed at [8]. The key idea is to hide information in a noise than occurs invariably in natural languages transformation. When translation a non-trivial text between a pair of natural languages, there are typically many possible translations. Selecting one of these translations can be used to encode information. For example:

"Джек нанес краску на стену" can be translated as:
"Jack spayed paint on the wall", or *"Jack sprayed the wall with paint"*.

3 Steganalysis methods

There converse problem of steganography is steganalysis. Its goal is to identify suspected container, determine whether or not they have embedded message in it, and, if possible, recover that message. Statistical attacks are commonly used for stegotext detection. For example, widely known support vector machines (SVMs) [9] are a set of related supervised learning methods that analyze data and recognize patterns, used for classification. There are two types of errors uses for evaluating the steganalysis methods reliability:
False Positive errors occur when the method mistakenly flags an natural text as stegotext.
False Negative errors occur when the method mistakenly flags stegotext as natural text.

The most easily detectable methods are syntactic because they could be detected by simple analyzer. Presence of double spaces in text might cause suspicion. It was noted earlier that methods of natural-like text generation have one disadvantage — resulting text is meaningless. It requires a human intervention to determine the meaningfulness of the text. However, it is not always possible, because of the large volume of messages transmitted in the network. It is necessary to create automated methods for steganalysis. Nowadays, there are a large number of different steganalysis methods.Let us consider in more detail the following method. A method using semantic shortcomings of methods published in [10]. When you replace the words on their synonyms can break semantic rules, for example:

"What time is it ?" Word *"time"* could be replaced as *"period"* or *"duration"*that do not agree with correct English usage. False positive ratio is 38.6%. False negative - 15.1%. Low reliability level makes it difficult to the practical application of this method. In addition, it is requires a lot of time working, and a large database of language rules. Method, proposed in [11], uses word frequency and its variance in the analyzed text. Obtained data and The Support Vector Machine (SVM) used for identify stegotext Nicetext, Texto or Markov-Chain-Based presence when container size more than 5Kb. Sum of errors less than 7.05%.

The most effective steganalysis for Nicetext stegosystem proposed in [12]. In virtue of the concepts in area of information theory, the method uses an information entropy-like statistical variable of words in detected text segment together with its variance as two classification features for SVM. The method was centered on detection for small size text segments estimated in the hundreds in words. The experimental accuracy of the method on classification of generated text and normal text exceeds 99% when text size is larger than 400 bytes. Even for sentences, the experimental accuracy exceeds 85%.

New method of statistical analysis was suggested in [13]. The compression used for stegotext detection. It is known that an embedding message breaks statistical structure of the container, increasing its entropy. Consequently, the full container will compress worse than empty. Let us consider the example:

A and B are empty and full containers, respectively.

Table 1: Container size before and after compression

container	before compr.	after compr.
A	500	320
B	500	300

Add content of suspected container C into A and B. Compare added content sizes before and after compression.

Table 2: Content size of C

container	before compr.	after compr.
C	50	45
C	50	20

It could be asserted that container C is statistically depend with B, which ensures good compression. This principle used in attack on Texto. Accuracy of detection exceeds 99.98% when text size is larger than 400 bytes.

The next one [14], statistical method used for attack on stegotext Nicetext, Texto, and Markov-Chain-Based. The average length of words, frequency of spaces, letter distribution of the words, first letter of word distribution used as features in SVM classificator. Detection accuracy exceeds 84.42% for text segments larger than 500 bytes.

Table 3: The effectiveness of existing steganalysis methods

stegosystem	400 bytes	1 Kb	5 Kb
Nicetext	99.61%	99.61%	99.61%
Texto	99.98%	99.98%	99.98%
MCB	84.42%	87.61%	99.46%

Analysis of translation-based steganography was published in [15]. This method has improved on a previously proposed linguistic steganalysis method based on word distribution which

is targeted for the detection of linguistic steganography like nicetext and texto. The newmethod aims to detect the application of TBS and uses none of the related information about TBS, its only used resource is a word frequency dictionary obtained from a large corpus, or a so called natural frequency dictionary, so it is totally blind. It is known, that stegotext consist less high frequency word then in natural text or translated text. This Method needs to know the machine translator set and covertext language. The experimental results show that the method accuracy is 87.7% when container size - 20 Kb.

The most effective method was proposed in [16]. It is used two features (words frequency and variance of the words distances as sentences structures) for SVM classification.

Let us consider distances of words in *"What is a web browser?"*.

Table 4: Word distances

	what	web	browser
what	0	1	2
web	1	0	1
browser	2	1	0

It should be noted than the sentences structures of stegotext looks more "noisy" than natural text. The total detection accuracy are 97.65%, 98.88% and 99.69% respectively when the text size is 10Kb, 15Kb and 20 Kb.

Conclusion

This paper presents a background on the major algorithms of text steganography and steganalysis. It is shown that overwhelming majority of effective steganalysis methods are based on statistical analysis. Most existing embedding methods could be detected with high probability.

References

[1] James C. Judge (2009). Steganography: Past, Present, Future. [URL] http://www.sans.org/reading_room/whitepapers/stenganography/steganography _past_present_future_552

[2] Winstein K. Tyrannosaurus lex (1999). [URL] http://alumni.imsa.edu/~keithw/ tlex/

[3] Barzilay R., Lee L. (2003). Learning to paraphrase: An Universal approach using multiple-sequence aligment. [URL] http://www.aclweb.org/anthology/N/N03/N03-1003.pdf

[4] Chapman M. , Davida G. . (1997). Hiding the hidden: A software system for concealing ciphertext in innocuous text. *Proceedings of the International Conference on Information and Communications Security*. Vol. **LNCS 1334**. Beijing, China.

[5] [URL] ftp://ftp.eenet.ee/pub/FreeBSD/distfiles/nicetext-0.9.tar.gz

[6] [URL] http://www.nic.funet.fi/pub/crypt/steganography/texto.tar.gz

[7] [URL] http://www.eblong.com/zarf/markov/chan.c

[8] Grothoff C.,Grothoff K.,Stutsman R.,Alkhutova L. (2009). Translation-based steganography. *Journal of Computer Security*. pp. 269-303.

[9] Chang C., Lin. C. (2001) LIBSVM : a library for support vector machines.(2001).[URL] http://www.csie.ntu.edu.tw/ cjlin/libsvm

[10] Taskiran C., Topkara U., Topkara M., Delp E. (2006). Attacks on Lexical Natural Language Steganography Systems // *Proceedings of the SPIE Security, Steganography, and Watermarking of Multimedia Contents VIII*. Vol. **6072**. pp. 97-105.

[11] Chen Z., Huang L., Yu Z., Zhao X., Zheng X. (2008).Effective Linguistic Steganography Detection. *IEEE 8th International Conference on Computer and Information Technology Workshops*. pp. 224-229.

[12] Meng P., Huang L.,Chen Z., Yang W., Li D. (2008). Linguis-tic Steganography Detection Based on Perplexity. [URL] http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=5089098&isnumber =5089035

[13] Нечта И.В. (2010). Эффективный метод стегоанализа базирующийся на сжатии №1. стр. 50-55.

[14] Hao Y., Xianbin C. (2010). Linguistic Steganalysis Based on Meta Features and Immune Mechanism *Chinese Journal of Electronics*. Vol. **19**, pp. 661-666.

[15] Meng P. (2010). Analysis and Detection of Translation-Based Steganography. *Lecture Notes in Computer Science*. Vol. **6387/2010**, Springer-Verlag. pp. 208-220.

[16] Chen Z. (2011). Blind linguistic steganalysis against translation based steganography. *Lecture Notes in Computer Science*. Vol. **6526/2011**, Springer-Verlag, Berlin. pp. 251-265.

Gradient Statistical Attack at Block Cipher RC6

A.S. Lysyak
Institute of Computational Technologies SB RAS
Novosibirsk, Russia
e-mail: accemt@gmail.com

Abstract

This work covers an experimental research of statistical methods in cryptoanalysis on the example of the block cipher RC6. The given gradient attack is based on the statistical test "book stack", developed by B.Ya. Ryabko. The attack's circuit allows to reduce considerably a labour of input private key finding. The earlier known variants of attacks based on the test of hi-square made the big complexity. In the given work the efficiency researches of gradient attacks are conducted, the limits of its modern practical and theoretical applicability are shown (up to 9 rounds of cipher RC6), the mathematical dependences between effectively cracked rounds and quantity of demanded computing resources are received. Also in this operation the trial and error method of optimal parameters for the test is given, and also their influence on attack is shown; the time estimation of an attack and its dependence on test and size parameters of cipher text is researched; theoretical requirements for the computational capabilities necessary for realization of attack are shown.

Introduction

Block ciphers are widely applicated in systems of transmission and information storage. Many experts are engaged in research of their theoretical and practical stability to any attacks. both new ciphers, and new attacks to them are permanently being developed. We will mark that those attacks which allow to find a key in time less than a method of direct exhaustive search of a private key are of interest.

Cipher RC6 - one of the challengers for the rank of AES, the competition carried out in 2001. There are a lot of operations devoted to the analysis of this cipher. The majority of them used results of the attack offered in [1], and are applicable to cipher RC6, but in some cut form, in other words, without operation of so-called "bleaching" ("post-whitening" and "pre-whitening"). These attacks are grounded on research of statistical properties of cipher RC6. For example, the result according to which the output sequence of cipher RC6 can be distinguished from casual sequence in the presence of suitable quantity selected cipher text (for r-th round is received it is necessary 2^{8r+10} text), has been obtained. Thus, the complexity for 4th round will already be considerable - 2^{74} that makes the attack almost impossible.

In the given work a new attack which allows to find a private key for 5th round is offered, thus time complexity is equal 2^{46}, and generally: $T(R) = 2^{8\cdot\left[\frac{9r-2}{7}\right]}$. The attack is grounded on the research of statistical properties of cipher RC6 with the assistance of the test "a pile of books" developed by B.Ya. Ryabko, which, according to many researches, is the best statistical test by the current moment.

In the given work the set of experimental data is resulted, and also efficiency of gradient statistical attack with an example of cipher RC6 is shown.

1 The description of gradient statistical attack

The described method belongs to the class of attacks with a selected ciphered text (chosen ciphertexts attack). At implementation of this attack a cryptography analyst can submit any text on an input of the cipher, parse the received ciphered message, then, knowing the results of this analysis, to submit the new message etc. The aim of the attack is to find a (confidential) key, and thus it is supposed that cryptography analyst knows all characteristics of the cipher, except this key.

For the whole class modern block ciphers an initial private key K will be transformed to so-called keys of rounds $K_1, K_2, ..., K_r$ which are used sequentially for enciphering at the different stages named as rounds. Schematically for such class of ciphers it is possible to present the process of enciphering as sequence of stages of enciphering in which all process is divided into some stages - rounds, thus at each following round there is an enciphering of the block received at the previous round.

$$x_1 = E(x_0, K_1),\ x_2 = E(x_1, K_2), ..., x_r = E(x_{r-1}, K_r),$$

where x_0 - an initial bulk which is necessary for ciphering, E - operation (function)of enciphering on i-th round, K_i - a key used on i-th round, x_i - a bulk, being "output" of i-th round and "input" $(i+1)$-th. In different ciphers this procedure is carried out differently, and it depends not only on the cipher, but also from values of block lengths and a key and number of rounds r which for many ciphers are parametres. For example, for cipher RC6 the block length can accept values 32, 64 or 128 bits, quantity of rounds can be any integer, and the key length should be multiple 8 and can accept any value, since 8 bits. Deciphering is spent under the circuit, return to enciphering:

$$x_{r-1} = D(x_r, K_r),\ x_{r-2} = D(x_{r-1}, K_{r-1}), ..., x_0 = D(x_1, K_1),$$

where the same keys of rounds, and operation D - inverse function to E are used.

For our task on an input of ciphers blocks α_i, which length is in binary notation equal to a block length of the data of the cipher of the data moved. Blocks move in sequences $\alpha_1, \alpha_2,\ \alpha_m$, i.e. blocks of a following sort: 00 ... 001, 00. 0010, 000. 0011 **(1")**.

Let's enter concept "randomness measures" a data series. We will accept there is set sequence α then statistical test $\Gamma(\alpha)$ is applicable to it and will take the value, return to $\Gamma(\alpha)$. This numerical value and we will name as "a randomness measure" sequences α.

One of the main requirements to modern block to ciphers consists that what data series on an input would not move to the cipher, on an output the sequence which measure of randomness aspires to infinity or very big should be received. Almost realized requirement consists that the sequence on an output would admit casual any statistical test **(1')**.

If on a cipher input to submit the data of sort (1) (that is essential, with lengths of everyone α_1 equal to a block size) the probability will detect on an output sequence with deviations from "randomness" considerably above since "randomness" of this sequence very low and simultaneously coincidence with a block size reduces "miscibility" of sequents and gives less tangled on an output of the cipher sequence. As a result of experiments described in [1] it has been shown that the measure of randomness of the sequence received on an output after i-th round in any by the block the cipher, increases with growth of number of rounds that completely corresponds to the requirement (1'). During too time, during time deciphering a randomness measure (or, in abbreviated form, "randomness") sequences decreases.

Let's short describe the circuit gradient attacks. Let there is any i-th round of the cipher: sequence of bats $x_1,\ x_2, ..., x_m$ on an input of operation of deciphering of a round i and the same sequence on an output of the decipherer of a round i: $y_1,\ y_2, ..., y_m$. Then it is possible to present the deciphering circuit in sort: $y_m = D(x_m, K_i),\ y_{m-1} = D(x_{m-1}, K_i), ..., y_1 = D(x_1, K_i)$. We estimate a measure of randomness of sequence further: $y_1,\ y_2, ..., y_m$. At deciphering with a correct key it should decrease strongly whereas deciphering with a wrong key to similarly enciphering, i.e. it on the contrary increases a measure of randomness of sequence. As a result, $\Gamma(y_1,\ y_2, ..., y_m) < \Gamma(x_1,\ x_2, ..., x_m)$ the key of round K_i is picked up correctly, differently - it is wrong. The sense of attack is reduced to finding of such key $K_i$which will minimize $\Gamma(y_1,\ y_2, ..., y_m)$. Theoretically for this purpose it is necessary to sort out all $2^{|K_i|}$ variants, but recognizing that a correct key gives essentially smaller â value $(y_1,\ y_2, ..., y_m)$, it is possible to try keys before finding of it of "an essential minimum".

Efficiency of the given attack in comparison with exhaustive search of initial key K consists that we can find separately keys of each round. As a rule, the size of keys in rounds is constant and equal in all rounds, thus, it usually much less on length, than the initial key K. Using the above described circuit, we can try keys K_i that gives the quantity of operations equal $\sum_{i=1}^{m} m_i \cdot 2^{|K_i|}$ **(1")**. At the equal sizes round keys, it is possible to define the upper estimation, as $r \cdot m_{\max} \cdot 2^{|K_0|}$, where K_0 - the size round a key. Quantity of operations at exhaustive search equally $2^{|K|}$. Thus, gradient attack is effective at performance of a following condition: $\sum_{i=1}^{m} m_i \cdot 2^{|K_i|-|K|} < 1$ **(2)** or, in the nearest approach: $r \cdot m_{\max} \cdot 2^{|K_0|-|K|} < 1$ **(2').** The formula (2') can be copied in a short form: $\log_2(m) + |K_0| < |K|$ **(2")**.

Thus, if there is a statistical test which effectively finds deviations from randomness on $(r-1)$ a round, at recommended r rounds the key of such cipher can be found essentially faster, than a method of direct exhaustive search. It is the main idea "gradient statistical attack".

As the statistical test the test "a pile of books", described in [4] has been taken. As this test has proved, as one of the most effective tests for check of cryptography generators (see [4]).

2 Research and optimization of the statistical test "a pile of books" in gradient to attack in application to cipher RC6

As the statistical test the test "a pile of books" has been taken. The given test has three key parametres on which efficiency of its application depends: a word length quantity of degree of freedoms and the size of each part (in case of 1 degree of freedom - the top size). All researches were spent on the test with 1 degree of freedom, and the remained two parametres varied. Test parametres we will designate, as to steam (u, w), where u - a test word length in bits, and w - the top size in bits (i.e. $\log_2(W)$, where W - the actual size of a top (quantity of units)).

During operation stability of cipher RC6 in relation to gradient to attack has been checked up. 128-bit mode RC6 with an 128-bit key. Cipher RC6 is arranged in such a manner that it ciphers 128-bit number in two stages. At first the first 32 bits are ciphered by means of the first key of the so-called semiround, then following 32 bits, using other key of the semiround, the remained 64 bits on the given round do not vary (they are ciphered on a following round). Outgoing it, it is possible to spend attack not on all round, and on each semiround separately.

In table 1 results of testing of the cipher c by the help of "a pile of books" are resulted. 100 tests with 100 various casually picked up keys have been spent.

Table 1

Round	Sample size	Test parametres	$> Q_{0.95}$	$< Q_{0.05}$	$E(Xi^2)$
3	2^{18}	$(24, 10)$	100	0	$6.05 * 10^7$
4	2^{18}	$(24, 10)$	100	0	14775.6
5	2^{18}	$(32, 22)$	16	14	1.72

It is shown that deviations from randomness in 16 % of cases can be fixed on 5 round. It means that gradient attack can effectively be applied on 6-th round of the cipher at the size of sampling 2^{18}. But it is possible to tell of 100 % about efficiency of application of attack to all only in case of 5 rounds since on 5 round fixing of deviations from randomness occurs only in 16 % of cases, and average Xi^2 value aloud is less quantile than a degree of belief 0.95. As under tests for a round 5 it is well visible that 14 % of keys give values Xi^2 less quantile a degree of belief 0.05 that speaks about high level of randomness of output sequence. It is actually possible to assume that there is a certain class of "weak" keys to which the given attack can effectively be applied on 6 rounds, and there are "strong" keys for which attack is effective only on 5 rounds.

Research of dependence of efficiency of considered attack from the size of sampling and from test parametres has been conducted. Results of research are presented in the following table.

From the given table it is visible that on each separate file of the set size there is a optimal set of parametres. This set essentially depends on a file size. The more the size, the there

Table 2

No the test	Round	Sample size	Test parametres	$> Q_{0.95}$	$E(Xi^2)$
1	4	2^{14}	(8, 7)	50	133.4
2	4	2^{14}	(8, 6)	50	210.9
3	4	2^{14}	(16, 15)	100	1265.0
4	**4**	2^{14}	**(16, 10)**	**100**	**18484.9**
5	4	2^{18}	(8, 7)	100	1653.5
6	4	2^{18}	(8, 6)	78	1204.7
7	4	2^{18}	(16, 15)	100	21142.5
8	**4**	2^{18}	**(16, 8)**	**100**	**118515.0**
9	4	2^{18}	(24, 20)	65	170.4
10	4	2^{18}	(24, 10)	100	14775.7
11	4	2^{18}	(32, 24)	70	36.3
12	4	2^{18}	(32, 16)	11	1.02
13	**4**	2^{22}	**(16, 8)**	**100**	**123552**
14	**4**	2^{22}	**(24, 10)**	**100**	**83179.5**
15	4	2^{22}	(32, 18)	11	1.68
16	4	2^{22}	(32, 24)	68	51.2

should be a word length in the test more. From above the word length is limited by a condition of applicability of criterion $Xi^2 : S \cdot 2^{w-u} \geq 5$ (S - the sampling size in blocks), but even at performance of this condition, but at too great value w *the* test yields bad results. In the same way and at too small value w or too small word length (u) the test badly recognizes randomness of sequence. Proceeding from the presented data, the optimal word length w is equal on the average: $w_{opt} = 8 \cdot \left[\frac{\log_2(L)}{8}\right]$ **(3)**, where L - length of entry sequence bits ($L = 2^b \cdot S$). In the table fat optimal parametres for each of the sequence sizes are selected. The optimal size of a top varies and precisely it is impossible to define it, but on the average the optimum w is, how $\frac{u}{2} \pm 10\%$ **(3 ')**. Besides, the task of selection of pair optimal parametres for the test is very essential that is brightly visible on an example of 14 and 15 tests, and as 8 and 12 tests. At optimally and correctly picked up under the concrete size of sampling (not exceeding 2^{22}) parametres of the test cipher RC6 can be cracked to 5-6 rounds, and at incorrectly picked up parametres this number decreases to 3-4 rounds, and in certain cases even more low.

Let's consider dependence of time of the analysis of sampling on its size and test parametres. In the following table the data for 4 rounds of the test and casually selected 100 keys is cited.

Table 3

No the test	Sample size	Test parametres	$E(Xi^2)$	Time
1	2^{14}	(16, 8)	11752.6	0.14 s
2	2^{18}	(16, 8)	191344.2	1.6 s
3	2^{22}	(16, 8)	1364622.9	24.5 s
4	2^{20}	(24, 23)	693.4	7.2 s
5	2^{20}	(24, 16)	6008.0	5.7 s
6	2^{20}	(24, 12)	24056.4	5.1 s

In table 3 the linear dependence of the size of sampling from time of the analysis of a file the test is well visible strictly. Besides, given results say that with growth of the size of sampling $E(Xi^2)$ value at what it grows grows the same as also time, is linear from a file size that is essential. Thus, it is possible to make the supposition about possibility of rise of number of rounds cracked by the given statistical test (number of rounds in which the test recognises deviations from randomness) with growth of the size of files. It is quite clear that with growth of the size of files as the optimal size of a word that demands additional and essential expenditure of memory will grow also. Tests 4-6 in table 3 confirm idea of optimal size of a top, as the radical square of size of all alphabet of the test: in case of a top equal on size to the radical square from u, value $E(Xi^2)$ in 34.7 times is more. Besides, the less top, the occurs the sampling analysis faster. Accordingly, except correct selection of parametres of the test, for cipher breaking it is required to define also the optimal size of sampling which is necessary for breaking of the set quantity of rounds as at too big sampling time of the analysis of sampling, and accordingly, and each selected key, becomes too big.

In work [2] research of the given attack in a 64-bit mode of cipher RC6 has been conducted. The received results are presented in table 4.

Table 4

Round	Number of tests	$Q_{0.99}$	Sample size
3	100	100	2^9
4	100	100	2^9
5	100	100	2^{18}
6	20	16	2^{29}

From the resulted data it is well visible that in this mode quantity of rounds in which the statistical test "the pile of books" reveals deviations from randomness on the average above on 1, i.e. in 64-bit variant RC6 number of cracked rounds equally 6-7. Thus, it becomes clear that efficiency of attack essentially depends on a block size in the cipher: the it is more, the attack is less effective.

3 Practical implementation and an efficiency estimation gradient statistical attack

Till now operation of cipher RC6 got to consideration on samplings in length to 2^{22}. From table 2 it is obviously visible that the test overall performance, and accordingly, and all attack essentially depends on a sample size: at a sample size 2^{10} the number of cracked rounds equally already 4 (for the 4-th round $Q_{0.95}$ is 25). Proceeding from it, it is possible to make the supposition that for rise of number of cracked rounds on 1, it is necessary to increase sampling in 2^8 times **(4)**. Nevertheless, there is a lawful question: whether it is possible to crack gradient statistical attack 128-bit cipher RC6 for more, than 5-6, number of rounds? The answer to this point in question contains in researches of American standards institute NIST. With application of less effective, than "a pile of books" statistical tests, they have obtained the following data, concerning breaking RC6 with the help gradient statistical attack.

Information given tables 5 can be interpreted as follows. The number of cracked rounds of 128-bit cipher RC6 on the house computer makes 5-6. Generally, it is possible to crack all cipher, i.e. all 20 rounds, but for this purpose is required at present technical inaccessible memory size. In spite of the fact that in the research which results are presented in table 5, the statistical test, on the metrics conceding to "a pile of books", nevertheless proceeding from the data received on "a pile of books" took part, it is possible to draw output that on high rounds a difference with standard statistical tests not so essential. Nevertheless, for the exact answer to this point in question it is required to conduct research of considered attack on a supercomputer then to compare results to other tests.

Table 5

Round	Key size	Sample size	Memory expenditure
12	128	2^{94}	2^{42}
14	128	2^{118}	2^{112}
14	192	2^{110}	2^{42}
14	256	2^{108}	2^{74}
15	256	2^{119}	2^{138}

Let's answer now a question: what quantity of rounds can be cracked at the present stage of development of computer facilities, and also, what interpolational dependence of number of rounds on requirements for memory if to consider, what given tables 5 correspond to the test "a pile of books"?

Let's name the cipher with n rounds effectively cracked gradient statistical attack if with application of this attack the quantity of operations on cipher breaking is essential less exhaustive search, i.e. if to it a condition (2), (2') or (2").

As rounds of cipher RC6 can be divided into semirounds realizing on it considered attack, it is possible to try keys for each of semirounds separately. In a 128-bit variant of cipher RC6 each key of a semiround has length of 32 bits, i.e. for selection of a key to a round r it is required at worst $m_r \cdot 2^{32}$ operations, where m_r - the size of the sampling submitted on an input of the decipherer of r th round. Proceeding from the supposition (4) and as table 5 data, it is possible to draw output that the sample size at which the statistical test "the pile of books" can successfully be applied at implementation gradient attacks, is really equal $\approx 2^{8\cdot\left[\frac{9R-80}{7}\right]}$ **(5)**. Number R thus can be potentially any. From here it is easy to make an estimation for number of operations: $T(R) = 2^{8\cdot\left[\frac{9R-30}{7}\right]} \cdot 2^{32} = 2^{8\cdot\left[\frac{9R-2}{7}\right]}$. It is necessary to notice that the above-stated estimation is fair only for breaking of the most high round of the cipher whereas attack should be applied on each round, since the senior to the low. Nevertheless, this estimation can be applied and to all attack, and not just for a high round. We will explain why. Proceeding from the formula (1") it is visible that the number of operations linearly depends on the sampling size m_i. Simultaneously, from the formula (5) follows that on each following - more low - a round the sample size for breaking will decrease in $\simeq 2^{10}$ time. Thus, already - following - a member of the sum of a number in (1") in 1000 times is less than second, i.e. its contribution to the sum is too small. Memory expenditure will be equal in bytes $2^{u+\log_2(w)} = w \cdot 2^u$ **(5')**, where (u, w) - test parametres. Such estimation of memory has proved to be true experimentally and is exact. From (3), (3 ') and (5) follows that $w_{opt}(R) = 8 \cdot \left[\frac{9R-30}{7}\right]$ and from (5) it is received: $\Pi(R) = 8 \cdot \left[\frac{9R-30}{7}\right] \cdot 2^{4\cdot\left[\frac{9R-80}{7}\right]}$. Under these formulas and as to the data of tables 5 and 2 we will make following table 6.

Table 6

Rounds	Quantity of operations	Expenditure memories (RAM)	Expenditure memories (ROM)
5	2^{46}	2^{19}	2^{20}
7	2^{68}	2^{25}	2^{28}
9	2^{87}	2^{35}	2^{44}
12	2^{126}	2^{52}	2^{98}
14	2^{142}	2^{66}	2^{114}
15	2^{151}	2^{70}	2^{123}
20	2^{204}	2^{70}	2^{176}

Conclusions

The results can be interpreted as follows. At computational capabilities accessible by the current moment gradient statistical attack with test application "the pile of books" can be applied to 9-raundovomu to cipher RC6, for this purpose it is required: 32 Gb of the RAM, 32 Tb of ROM, and also 2^{87} exhaustive search operations. Whereas in case of attack by exhaustive search it is required 2^{128} operations. Since 13 rounds to apply gradient statistical attack it makes sense only in case of usage of 192-/256-bit keys is follows from (2 "). For 19-20 rounds - only in case of 256-bit keys. Also it is possible to interpret these results in another way: at a 128-bit key and accessible resources of 32 Gb of the RAM, 32 Tb of ROM, quantity effectively cracked gradient attack of rounds equally 10 (87+32=119 <128). At a 192-bit key the number of effectively cracked rounds increases to 12 (87 + 32*3 = 183 <192). And in case of a 256-bit key this number is equal 14 (87 + 5*32 = 247 <256).

In the perspective, it is supposed to hold this statistical attack on the other block ciphers: AES, 3DES, GOST 28147-89.

References

[1] Knudsen L., Meier W. (2000). Correlations in RC6 with a reduced number of rounds. *FSE 2000. LNCS 1978*. Springer-Verlag. 94-108.

[2] A.A. Menjashev, V.A. Monarev, B.Ya. Ryabko, A.N. Fionov (2006). Application gradient statistical attack to block to ciphers RC5, RC6 And SAFER. *Scientifically-practical log "Informational counteraction to terrorism"*, 7.

[3] V.A. Monarev, A.M. Lubkin (2010). Effective attack to block cipher RC6. *Vestnik SibGuTi.*, 4.

[4] B.Ya. Ryabko, A.N. Fionov (2006). Cryptography methods of protection of the information, Moscow.

[5] B.Ya. Ryabko, V.A. Monarev, J.I.Shokin (2005). The new type of attacks to block ciphers. *"Information transfer Problems", permanently delete.*, 41, 4, 181-128.

[6] B. Schneier (1996). Applied Cryptography. *N.-Y. Wiley*.

[7] B.Ya. Ryabko, V.A. Monarev (2005). Using information theory approach to randomness testing. *Journal of Statistical Planning and Inference*, Vol. **133**, 1, 95-110.

Non-Linear Probability Models and Problems of Their Application

V.V. Gubarev
Novosibirsk state technical university,
Novosibirsk, Russia

Abstract

Concepts of the "linear" and "non-linear" random signals, as phisical information carriers, and their mathematical models in the form of random variables, vectors, the continuous in time processes and the discrete series are considered. Concepts are coordined with methods of defining and characteristics of static and dynamic signals models. The examples underlining unfitness or small suitability of the "linear" characteristics in a non-linear reality are resulted; characteristics and models, suitable for non-linear situations and also examples of their application for identification of non-linear systems and estimation of communication parameters of multidimentional distributions of random vectors, processes, series.

1 Introduction. Problem statement

The theory of probabilistic models of real processes and objects, i.e. models, represented as the form of random variables, vectors, scalar and vector functions with one (random processes, time series – RP) or several (fields) arguments, is well enough theoretically developed. However their application in the majority of theoretical and practical implementation used us one-dimensional or bivariate distributions of probabilities, linear, rarely non-linear, correlation and regression analysis, static (for values and vectors) and linear correlation-spectral dynamic (for processes, fields) analyses. However rapid development and propagation of signal and data collection and processing tools leads to necessity of use non-linear description of real processes and objects since linear approximation of description is not effective.

The goul of the present report is reviewing of non-linear probabilistic models of signals as information carriers about real processes and objects.

2 Probabilistic models decription

2.1 Initial definitions and designations

To concretize approaches stated in paper, we consider only static and "dynamic" probabilistic models, random vectors $X = (X_1, X_2, ..., X_n)$ as an example of the static model /the first model/ and stationary processes $X(t) = (X_1(t), X_2(t), ..., X_n(t))$ as an example of the dynamic model (the second). For abridgement let be $X_i = X(t_i)$ or $X_i = X_i(t_i)$, $i = \overline{1,n}$. Let $F(x_1, ..., x_n; t_1, ..., t_n)$ and $W(x_1, ..., x_n; t_1, ..., t_n)$ probabilistic function and absolutery relative

density distribution, $M\{f[X_i(t_i), X_j(t_j); X_k(t_k) = x_k]\}$, $i, j, k = \overline{1, n}$, $i, j \neq k$, the operator of mathematical expectation (average) $f[X_i(t_i), X_j(t_j)]$ at $X_k(t_k) = x_k$, supposing, certainly, that it exists and has finite quantities.

2.2 Methods of the random processes description

Definition of non-linear random processes as signals models, it is necessary to connect methods of their mathematical description, representation. Let's review the most popular. Two methods based on the main and selective probability spaces, are general-purpose, suitable for random variables, vectors and processes. For random variables and vectors this method means the defining through distribution laws, for example, $F(\cdot)$ or $W(\cdot)$. The concept of "nonlinearity" in this case is similar to what will be considered for the defining of random processes (RP) through family of finitedimensional distributions (see the third method of the defining for RP). The second method of the defining for the random processes is connected with probabability measures on set of process trajectory (implementations, sample functions). Linearity or nonlinearity of RP is thus defined by linearity or nonlinearity (in some sense) of process trajectory characteristics or average characteristics. This method is not widely used. Therefore we won't consider it in detail.

In practice the third method – the defining of RP by means of family of finite dimensional distributions, i.e. a set of distributions $F(x,t)$, $F(x_1, x_2; t_1, t_2)$,..., $F(x_1, \ldots, x_n; t_1, \ \ldots, t_n)$ for different n = 1,2,3, ... and $t_1, t_2, ..., t_n \in (-\infty, \infty)$ [1] is most often used. We will consider it in detail.

Besides the explicit RP defining by means of a probability measure their representation in the form of the mathematical description of output signals of some given dynamic system with in advance known, simply described input signals often is used. It is the description in the form of systems of the integro-differential or finite-difference equations, when RP is their decision. For example, random processes as autoregression or moving average. Linearity or nonlinearity of RP thus is a consequence of linearity or nonlinearitigs in the standard sense of equations system and therefore doesn't need special reviewing.

The same can be said about other two types of random processes – canonical and not canonical [1, 3]. As soon as in practice their modeling representation is implemented through the same characteristics, as for the models defined by finitedimensional distributions, we will consider terms the "linear" or "non-linear" model first, start, stage only obtained by the characteristics received through distribution laws.

3 Static non-linear probability models

The idea of probability models of real processes and objects is connected to the description with not separate specific quantities of random values, vectors and functions, but with mass behavior of quantities in the same conditions. Therefore "linearity" or "nonlinearity" model concepts should be considered through the characteristics reflecting this mass character in any sense: on the average, quantiles, in particular on median, in mean-square, on them interquantiles. We consider

in the present paper only by linearity or nonlinearity of model at averages level. We say, that model is linear, if its regression functions, for example, $m_{\mathrm{X}}(y;\tau) = M\{X(t)/X(t+\tau) = y\}$ or $m_X(y;\tau) = M\{X(t)/Y(t+\tau) = y\}$, are linear functions with respect to all arguments. We will consider otherwise its non-linear.

It is clear that for random variables and vectors, which development in time, and, hence, its dynamic, is not considered, it is logical to suppose all models as static models. But it's not wright for random processes $X(t)$. For them, it is possible, to consider static of processes, i.e. their characteristics for the fixed moments of time $t_1, t_2, ..., t_n$, as well as dynamics, i.e. behavior of characteristics with changing $t_1, t_2, ..., t_n$ and their arrangement among themselves.

Since for stationary processes

$$F(x_1, \ldots, x_n; t, t+\tau_1, \ldots, t+\tau_{n-1}) = F(x_1, \ldots, x_n; \tau_1, \ \ \ldots, \tau_{n-1}),$$

characteristics of stationary RP are independence from t. They depend only from $\tau_1, ..., \tau_{n-1}$ relative positioning of of time moments $t, t+\tau_1, ..., t+\tau_{n-1}$, and the static description is similar to the description of a random vector $X_1 = X_1(t),\ X_2 = X_2(t+\tau_1), ...,\ X_n = X_n(t+\tau_{n-1})$ with one stipulation connected with not each regression function of a random vector can be regression function of random process, especially autoregression function [1].

Necessary and sufficient conditions at which functions of a regression of two random variables or random processes samples will be linear, are presented in [1, 4]. In many sources examples of bidimentional distributions with nonlinear regression are presented. In [2] there are examples of three discrete and 45 absolutely continuous bidimensional and 9 multidimentional distributions, from which only small part has the linear regressions, are resulted. From sets of bidimensional and multidimentional distributions a particular to be interested such, which distribution densities can be presented in the form of expansions in orthogonal series (rows) on system of the functions, which weight coincide with onedimensional distribution densities [1, 2, 4], and also noncentral distributions generalizing them [2]. Among them well-known distributions which examples are discussed in the report.

For random vectors it is repeatedly shown, that in case of non-linear regression the application of traditional Pearson correlation coefficients as numerical index of presence, direction and closeness (tightness) parameters of communication of random variables can mislead (see examples in [1]). Many new closeness parameters (indexes of communication) for non-linear regression were offered (see, for example, [2]). There are among them correlation relations, concorrelation coefficients [1, 2] and others, the part from which is resulted in the report.

4 Dynamic non-linear probability models

Now we will consider bi- and multidimensional models reflecting of random processes dynamics of development, i.e. its parameters and characteristics dependence on samples arrangements $\tau_1, ..., \tau_{n-1}$. We will name such RP and, hence, random signals described by them, non-linear, if regression functions of process samples $X(t),\ X(t+\tau)$ (at $\tau \neq 0$) or (at $X(t),\ Y(t+\tau)$ any τ) are non-linear.

Table 1: Table

$W(x,y)$	$R_{XY}(\tau)$	$\chi_{XY}(\tau)$	Is the model lineary?
1. Gaussian (normal) $N(a_1,a_2;\lambda_1,\lambda_2;\psi)$			
$\left[2\pi\lambda_1\lambda_2\sqrt{1-\psi^2}\right]^{-1}\times$ $\times\exp\left\{-\frac{1}{2(1-\psi^2)}\times\right.$ $\left.\times\left(z^2-2\psi zu+u^2\right)\right\}$; $z=(x-a_1)/\lambda_1, u=(y-a_2)/\lambda_2$; $\lvert x\rvert,\lvert y\rvert,\lvert a\rvert<\infty$; $\lambda>0$; $\lvert\psi\rvert\le 1$.	$\lambda_1\lambda_2\psi(\tau)$	$\frac{6}{\pi}\arcsin\left[\frac{\psi(\tau)}{2}\right]\approx$ $\approx\psi(\tau)=\rho_{XY}(\tau)$	yes
2. $\Im$-distribution $\Im(a_1,a_2;\lambda_1,\lambda_2;\alpha_1,\alpha_2;\beta;\psi)$			
$\alpha_1\alpha_2\left[\lambda_1\lambda_2\Gamma(\beta)\psi^{\beta-1}\left(1-\psi^2\right)\right]^{-1}\times$ $\times z^{[\alpha_1(\beta+1)-2]/2}u^{[\alpha_2(\beta+1)-2]/2}\times$ $\times\exp\left\{-\frac{1}{1-\psi^2}\left(z^{\alpha_1}+u^{\alpha_2}\right)\right\}\times$ $\times I_{\beta-1}\left(\frac{2\psi\, z^{\frac{\alpha_1}{2}}u^{\frac{\alpha_2}{2}}}{1-\psi^2}\right)$; $z,u\ge 0$, i.e. $x\ge a_1$, $y\ge a_2$, $\lvert a_1\rvert,a_2\rvert<\infty$; $\psi=\psi(\tau)$; $0\le\lvert\psi\rvert\le 1$; $\alpha_1,\alpha_2,\beta>0$; $\Gamma(\cdot)$ – is a gamma-function; $I_\beta(\cdot)$ – is a Bessel function	$\lambda_1\lambda_2\Gamma\left(\beta+\frac{1}{\alpha_1}\right)\times$ $\times\Gamma\left(\beta+\frac{1}{\alpha_2}\right)\Gamma^{-2}(\beta)\times$ $\times\left[\left(1-\psi^2\right)^{\beta+\frac{1}{\alpha_1}+\frac{1}{\alpha_2}}\times\right.$ $\times F\left(\beta+\frac{1}{\alpha_1};\right.$ $\left.\left.\beta+\frac{1}{\alpha_2};\beta,\psi^2\right)-1\right]$; $F(\alpha,\beta;\gamma;x)$– giper-geometrical function	$\frac{3\psi^2(\tau)}{4-\psi^2(\tau)}$	yes at $\alpha_1=1$, $\alpha_2=1$, non at $\alpha_1\ne 1$, $\alpha_2\ne 1$
Particular cases of $\Im$-distribution are: Nakagamy ($\alpha_1=\alpha_2=2$); Maxwell ($\alpha_1=\alpha_2=2$; $\beta=3/2$); Rayleigh ($\alpha_1=\alpha_2=2$; $\beta=1$); Waybull ($\beta=1$); exponential ($\alpha_1=\alpha_2=1$; $\beta=1$); gamma ($\alpha_1=\alpha_2=1$)			
3. Pearson I $\pi_I(a_1,a_2;\lambda_1,\lambda_2;\alpha,\beta;\psi)$			
$\Gamma(\alpha+\beta)(1+z)^{\alpha-1}(1-z)^{\beta-1}\times$ $\times(1+u)^{\alpha-1}(1-u)^{\beta-1}\times$ $\times\left[\lambda_1\lambda_2\Gamma(\alpha)\Gamma(\beta)2^{2(\alpha+\beta-1)}\right]^{-1}\times$ $\times\sum_{n=0}^{\infty}A_nP_n^{(\alpha,\beta)}(z)\times$ $\times P_n^{(\alpha,\beta)}(u)\psi^{n(n+\alpha+\beta-1)}$, $A_n=\frac{(2n+\alpha+\beta-1)\Gamma(n+\alpha+\beta-1)}{\Gamma(n+\alpha)\Gamma(n+\beta)n!}$	$\frac{\alpha\beta\lambda_1\lambda_2\psi^{\alpha+\beta}(\tau)}{(\alpha+\beta+1)(\alpha+\beta)^2}$	$\psi^2(\tau)$	yes
4. Arcsin $A(a,\lambda;\psi)$; $X(t)=a+\lambda\sin(\nu t+\Xi)$, Ξ – uniform on $(0,2\pi)$			
$\left(2\pi\sqrt{\lambda^2-(x-a)^2}\right)^{-1}\times$ $\times[\delta\{y-a-\lambda\cos[\nu\tau+\arccos(x-a)]\}+$ $\delta\{y-a-\lambda\cos[\upsilon\tau-\arccos(x-a)]\}]$ $\lvert x-a\rvert\le\lambda;\lvert y-a\rvert\le\lambda$; $\psi=\cos(\upsilon\tau)$; $\delta(\cdot)$– delta-function	$\frac{\lambda^2}{2}\cos(\nu\tau)$	$96\pi^{-4}\times$ $\times\sum\limits_{n=0}^{\infty}\frac{\cos[(2n+1)\nu\tau]}{(2n+1)^4}\approx$ $\approx\cos(\nu\tau)$	yes

Usually for the closeness characteristic of these counting's are used auto

$$R_{XX}(\tau) = M\left\{\overset{\circ}{X}(t)\,\overset{\circ}{X}(t+\tau)\right\}, \overset{\circ}{X}(t) = X(t) - m_X, \tag{1}$$

and cross

$$R_{XY}(\tau) = M\left\{\overset{\circ}{X}(t)\,\overset{\circ}{Y}(t+\tau)\right\} \tag{2}$$

correlation functions (CF), received through them normalized CF (NCF)

$$\rho_{XX}(\tau) = R_{XX}(\tau)/R_{XX}(0), \rho_{XY}(\tau) = R_{XY}(\tau)\Big/\sqrt{R_{XX}(0)R_{YY}(0)}, \tag{3}$$

and as well as spectral power densities (SPD) $S_{XX}(\upsilon) = \mathrm{F}\left\{R_{XX}(\tau);\varphi(j\upsilon\tau)\right\}$ and $S_{XY}(\upsilon) = \mathrm{F}\left\{R_{XY}(\tau);\varphi(j\upsilon\tau)\right\}$, where $\mathrm{F}(\,\cdot\,)$ – continuous (then $\upsilon = \omega \in (-\infty,\infty)$) or discrete (then $\upsilon \in (-\pi/2,\pi/2)$) Fourier transform on base $\varphi(j\upsilon\tau)$, more often exponential, when $\varphi(j\upsilon\tau) = \exp\{j\upsilon\tau\}$, $j = \sqrt{-1}$.

As soon as CF $R(\tau)$ is correlation moment and NCF $\rho(\tau)$ – correlation coefficient, for their small suitability as connected characteristics of direction presence and closeness on average for non-linearly connected $X(t)$, $X(t+\tau)$ or $X(t)$, $Y(t+\tau)$ at different τ, all stated earlier is valid closeness. The elementary examples, confirming it for the monotonous one-to-one transformations $Y(t) = f[X(t)]$, are presented in [2] and in the picture.

As SPD $S(\upsilon)$ derive from by linear transformations from CF, it is followed from stated above that traditional spectral analysis can be not effective, even to mislead researchers, for non-linear RP. Therefore in the report various variants of the description of correlation-spectral analysis analogs of considered type non-linear random processes are given. These are expansion method for Voltaire rows (series), dispersition, concorrelation, wavelet analyses, the methods, connected with selection and separate research of the linear and non-linear components of systems, etc.

Explicitly the method concorrelation and conspectral analyses [1, 2] is described in detail: its bases, a singularity, merits and demerits. Examples of correlation and concorrelation functions for many typical bidimensional distributions are resulted and their merits and demerits for the decision of application-oriented tasks are presented.

As an example in the adduced table of bivariate distributions dependences formulas of CF and concorrelational functions (CCF) from connection parameters of distributions are given.

Concorrelational functions (CCF) $K_{XY}(\tau)$ is defined as [1, 2]:

$$K_{XY}(\tau) = \mathrm{M}\left\{\left(F_X\left[X(t)\right] - \mathrm{M}\left\{F_X\left[X(t)\right]\right\}\right)\left(F_Y\left[Y(t+\tau)\right] - \mathrm{M}\left\{F_Y\left[Y(t+\tau)\right]\right\}\right)\right\}, \tag{4}$$

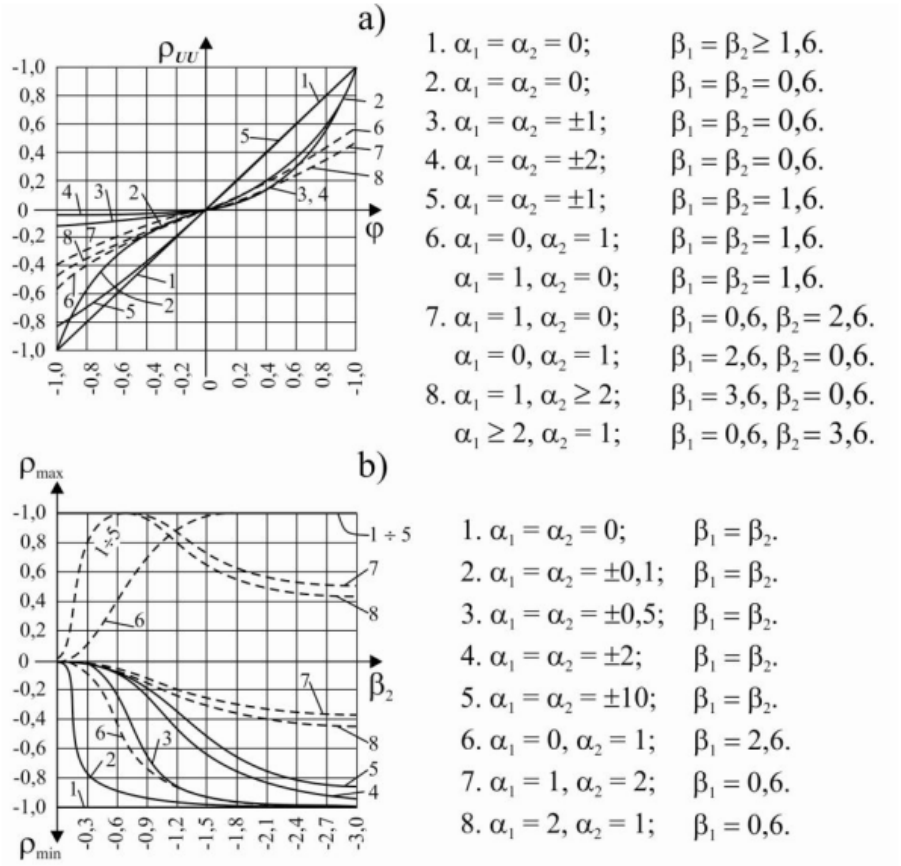
half-concorrelation (HCCF) $H_{F_XY}(\tau)$ and $H_{XF_Y}(\tau)$

$$H_{XF_Y}(\tau) = \mathrm{M}\left\{\left(X(t) - \mathrm{M}\left\{X(t)\right\}\right)\left(F_Y\left[Y(t+\tau)\right] - \mathrm{M}\left\{F_Y\left[Y(t)\right]\right\}\right)\right\}, \tag{5}$$

$$H_{F_XY}(\tau) = \mathrm{M}\left\{\left(F_X\left[X(t)\right] - \mathrm{M}\left\{F_X\left[X(t)\right]\right\}\right)\left(Y(t+\tau) - \mathrm{M}\left\{Y(t)\right\}\right)\right\}. \tag{6}$$

Normalized CCF (NCCF) are analoques of NCF $\rho_{XY}(\tau)$. For example,

$$\chi_{XY}(\tau) = K_{XY}(\tau)\Big/\sqrt{K_{XX}(0)K_{YY}(0)}. \tag{7}$$



$$\rho_{UU} = \frac{\nu_1\nu_2[\mathrm{ch}(\delta_1+\delta_2)(\nu_{12}^{\varphi}-1) - \mathrm{ch}(\nu_1-\nu_2)(\nu_{12}^{-\varphi}-1)]}{\sqrt{(\nu_1^2-1)(\nu_2^2-1)[\nu_1^2\,\mathrm{ch}(2\delta_1)+1]\,[\nu_2^2\,\mathrm{ch}(2\delta_2)+1]}},$$

$$\rho_{UL} = \frac{2\sqrt{2}\,\mathrm{ch}(\varphi/2\beta_1\beta_2)\,\mathrm{ch}(\varphi/2\beta_1\beta_2-\delta_2)}{\sqrt{(\nu_1^2-1)(\nu_2^2-1)[\nu_2\,\mathrm{ch}(2\delta_2)+1]}}, \qquad \delta_i = \frac{\alpha_i}{\delta_i}$$

Form parameters	α	0			|0,5| ÷ |1,0|				|2| ÷ |4|			
Form parameters	β	0,6	1,0	2,0	0,6	0,8	1,0	2,0	0,6	0,8	1,0	3,0
Skewnes coefficient	$\|\gamma_1\|$	0	0	0	6,4÷ 7,0	1,0 1,2	0,37 0,38	0,05 0,09	7,0 7,0	1,3 1,3	0,61 0,62	0,09 0,09
Excess coefficient	γ_2	3371	3,3	0,15	6785 7514	56,1 73,8	6,04 9,04	0,18 0,26	7598 7600	79,3 79,7	10,8 11,1	0,17 0,17

Figure 1: Permissible values ranges characteristic curves for correlation coefficient ρ of bidimensional S_L and S_U-Johnson distributions: a) – dependence of ρ from connection parameter ψ; b) – dependence of $\rho_{\max} = \rho(\psi = 1)$ and $\rho_{\min} = \rho(\psi = -1)$ form parameters β_1, β_2 of values (variables) X_1 and X_2 distributions

Direct continuous or discrete exponential Fourier transform F{·}of CCF and HCCF are conspectral (CCD) $C_{XY}(\upsilon)$ or half-conspectral "capacity" density (HCCD) $\Pi_{F_XY}(\upsilon)$ and $\Pi_{XF_Y}(\upsilon)$.

Let's give an example wonderful two properties of CCF and, hence, CCD. The first: CCF and CCD are for all random processes and series exist while CF and SCD are exist only for those ones, which have the finite moments of 1st and 2nd orders.

The second relate invariancy of CCF and CCD to noninertional transformations of signals. If $U(t) = f[X(t)]$ and $V(t) = g[Y(t)]$, where$f(\cdot)$and$g(\cdot)$ are determined monotone one-to-one functions,

$$K_{UV}(\tau) = \varepsilon_f \varepsilon_g K_{XY}(\tau), \tag{8}$$

$$C_{UV}(\upsilon) = \varepsilon_f \varepsilon_g C_{XY}(\upsilon), \tag{9}$$

$$\varepsilon_f = sign\left[df(x)/dx\right]. \tag{10}$$

5 Practical examples

Examples of practical tasks decision using various methods of non-linear random process representation are resulted in the report. These are examples on identification of non-linear systems, in particular Hammerschtein-Wiener (linear dynamic – non-linear static – linear dynamic links), estimations of correlation parameters of real processes dynamic, etc.

6 Conclusion

Initial approaches to representation and application of random signal non-linear models are stated. In summary priority problems, which should be resolved for wide application of non-linear probability models, are formulated.

References

[1] Gubarev, V.V. Algorithms of spectral analysis of random signals. Novosibirsk: NSTU, 2005. – 660 p.

[2] Gubarev, V.V. Probability models. – Novosibirsk: NETI, 1992. – 421 p.

[3] Sinitzin, I.N., Sinitzin, V.I., Korepanov, E.R. and athers. Automatica and telemehanika, 2011. – No. 2. – P. 183–194.

[4] Deutsch, Ralph. Nonlinear transformations of random processes. N.J.: Prentice-Hall, Inc., 1962. – 189 p.

Part VI
Robust Methods of Statistical Analysis

Robust Estimation of Qualitative Response Regression Models

Alexander A. Kalinin, Daniil V. Lisitsin
Novosibirsk State Technical University
Novosibirsk, Russian Federation
e-mail: kalinin.a.letters@gmail.com

Abstract

Qualitative response regression models such as logistic regression are typically estimated by the maximum likelihood method. To improve its robustness, two special cases of the M-estimation based approach for quantitative continuous random variables were extended to the variant of qualitative and mixed variables modeling. Expressions of the score functions for polytomous regression models were derived. In according to results of the research some conclusions and practical recommendations were given.

Keywords: qualitative response, Bayesian dot contamination, polytomous regression, robust estimation, 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. The general robust theory is developed in Huber [7] and Hampel, Ronchetti, Rousseeuw, and Stahel [6]. Recent work describing robust statistics in detail is Maronna et al. [10]. Generally robustness theory has been developed for the quantitative continuous random variables modeling. Qualitative and mixed variables modeling are paid much less attention. Several authors have studied the logistic regression model in terms of the robustness properties of the maximum likelihood estimation (MLE) and it's modifications. The maximum likelihood estimator attains the minimum asymptotic variance under the model and then it is optimal, but it is very sensitive to atypical data. Observations with extreme covariates, in particular, have a large influence on the estimator, and if they are accompanied by misclassified responses, the resulting estimates can be seriously biased. Pregibon (see [11]) made the earliest systematic attempts to fix this problem; he proposed methods to unmask influential observations and robust estimators for the logistic model. Later robust proposals in this area include Carroll and Pederson [2], Bianco and Yohai [1], Croux and Haesbroeck [4], and Gervini [5]. Typically, in these works binary regression models are considered. Many approaches for binary choice estimators development were introduced as alternatives to the maximum likelihood estimators, but they often are of semi-heuristic nature. Also note, that the numeric character of a binary variable is assumed in many papers. Recent examples include Victoria-Feser [13],

Čśamplež̆ek [3], and Kotlyarova and Zinde-Walsh [8]. All these estimators differ greatly in terms of outlier resistance and efficiency under the model.

The one of the most perspective approaches was suggested by Shurygin in [7] (see also [9]). Shurygin's approach based on Bayesian dot contamination of model distribution allows to get the estimators possessing a high robustness and efficiency. Originally the estimators within Shurygin's approach were formed only for continuous random variables models. However the theory developed in [12, 9] can be easily extended to the cases of scalar qualitative or count and vector mixed response models, where the latter consists of qualitative polytomous and quantitative responses. Qualitative polytomous (multinomial) response can be nominal or ordinal. In the latter case, one uses cumulative link model, continuation ratio model, stereotype model, and others. So, the purpose of this study is to develop a general theory of robust estimation for regression models with polytomous response and its application to the case of the nominal response.

1 Model Specification

Assume that discrete random variable Z has a fixed number of acceptable values $\{1, 2, ..., J\}$. Distribution of Z_t under observation t is set of model probabilities

$$P\left\{Z_t = j | x_t, \alpha\right\} = \pi_j\left(x_t, \alpha\right), t = 1, ..., N,$$

where x_t is a vector of covariates, α is a vector of parameters.

M-estimation $\hat{\alpha}$ of vector of parameters α is obtained by solving equations system

$$\sum_{t=1}^{N} \Psi(Z_t, x_t, \hat{\alpha}) = 0, \tag{1}$$

where $\Psi(Z_t, x_t, \hat{\alpha})$ is a vector score function satisfying further condition for all t

$$\sum_{j=1}^{J} \pi_j(x_t, \alpha)\Psi(j, x_t, \alpha) = 0. \tag{2}$$

2 Robust Estimation

One of the major indicators of estimator's robustness is an influence function which in the case under some regularity conditions takes the form

$$IF(Z, x, \alpha) = M^{-1}\Psi(Z, x, \alpha), \tag{3}$$

where

$$M = \sum_{t=1}^{N}\sum_{j=1}^{J} \Psi(j, x_t, \alpha)\frac{\partial}{\partial \alpha^T}\pi_j(x_t, \alpha).$$

In the Bayesian dot contamination model the distribution of Z_t is defined by the set of probabilities

$$P\left\{Z_t = j|x_t, \alpha, Z_t^*\right\} = (1-\varepsilon)\pi_j\left(x_t, \alpha\right) + \varepsilon\delta_{jZ_t^*},$$

where Z_t^* is discrete random variable with fixed number of acceptable values $\{1, 2, ..., J\}$ and distribution $P\left\{Z_t^* = j|x_t, \alpha, Z_t^*\right\} = s_j(x_t, \alpha)$, ε is contamination level ($0 < \varepsilon < 0.5$), δ is Kronecker delta.

Indicator of estimation badness in Bayesian dot contamination model can be written as functional

$$U_t(\Psi) = \sum_{t=1}^{N}\sum_{j=1}^{J} IF(j, x_t, \alpha)IF^T(j, x_t, \alpha)s_j(x_t, \alpha). \tag{4}$$

Corresponding optimum score function in Bayesian dot contamination model is of the form represented

$$\Psi(Z, x, \alpha) = C\left[\frac{\partial}{\partial\alpha}\ln\pi_Z(x, \alpha) + \beta\right]\frac{\pi_Z(x, \alpha)}{s_Z(x, \alpha)} =$$

$$= C\sum_{j=1}^{J}\frac{\partial}{\partial\alpha}\ln\pi_j(x, \alpha)\left[\delta_{jZ} - \frac{\pi_j^2(x, \alpha)/s_j(x, \alpha)}{\sum_{l=1}^{J}\pi_l^2(x, \alpha)/s_l(x, \alpha)}\right]\frac{\pi_Z(x, \alpha)}{s_Z(x, \alpha)}, \tag{5}$$

where C is nonsingular matrix, vector $\beta = \beta(x, \alpha)$ provides fulfillment of the condition (2).

2.1 Generalized Radical Estimator

Generalized radical estimation (GRE) corresponds to the case:

$$s_j(x, \alpha) = \left[\pi_j(x, \alpha)\right]^{1-\lambda}/\Delta(x, \alpha, \lambda),$$

where λ is estimator parameter ($\lambda \geq 0$), value of $\Delta(x_t, \alpha, \lambda)$ either equals $\sum\limits_{l=1}^{J}\left[\pi_l(x_t, \alpha)\right]^{1-\lambda}$ (used for satisfying probabilities normalizing condition) or is identity, if that condition is not used. Note that the case of $\lambda = 0$ matches maximum likelihood estimation.

For modeling dependence of nominal response from covariates polytomous logistic regression is often used. Corresponding probabilities are of the form

$$\pi_j(x_t, \alpha) = exp\left[\Phi(x_t)\alpha_j\right]\left\{1 + \sum_{k=1}^{J-1} exp\left[\Phi(x_t)\alpha_k\right]\right\}^{-1}, \tag{6}$$

where $\Phi(x_t)$ 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.

Generalized radical estimation of subvector α_j in polytomous logistic regression model is defined by the score function

$$\Psi_j(Z_t, x_t, \alpha) = \left\{ \delta_{jZ_t} - \frac{[\pi_j(x_t, \alpha)]^{1+\lambda}}{\sum\limits_{l=1}^{J} [\pi_j(x_t, \alpha)]^{1+\lambda}} \right\} [\pi_{Z_t}(x_t, \alpha)]^{\lambda} \Delta(x_t, \alpha, \lambda) \Phi^T(x_t). \tag{7}$$

2.2 Conditionally Optimal Estimator

In the set of robust estimators also can be used estimation with the optimum score function in Bayesian dot contamination model given by

$$\Psi(Z, x, \alpha) = C \left[\frac{\partial}{\partial \alpha} \ln \pi_Z(x, \alpha) + \beta \right] \frac{1}{1 + \dfrac{k^2}{\pi_Z(x, \alpha)}},$$

where k^2 is estimator parameter and C, β are the same as in (5).

To obtain conditionally optimal estimator, assume that distribution of Z_t^* is given by

$$s_j(x_t, \alpha) = \pi_j(x_t, \alpha) + k^2.$$

Hence, taking into account (6) the score function for conditional optimal estimation in polytomous logistic regression model is of the folowing form

$$\Psi_j(Z_t, x_t, \alpha) = \left\{ \delta_{jZ_t} - \frac{\dfrac{\pi_j^2(x_t, \alpha)}{\pi_j(x_t, \alpha) + k^2}}{\sum\limits_{l=1}^{J} \dfrac{\pi_l^2(x_t, \alpha)}{\pi_l(x_t, \alpha) + k^2}} \right\} \frac{1}{1 + \dfrac{k^2}{\pi_{Z_t}(x_t, \alpha)}} \Phi^T(x_t). \tag{8}$$

3 Experimental Research

In practice, there may be several solutions of equation system (1). Thus some methods of selection solutions are should to be used. Also it is necessary during solving to distinguish between consistent and inconsistent solutions and leave the latter out.

As a check of working capacity of proposed approaches was performed experimental research of generalized radical estimation of polytomous logistic regression model with nominal response having three levels. Maximum likelihood and robust estimators were compared under following values of estimators' parameters and model's parameters α. Vector of regressors is of the form $[1, x, x^2]$. True values of model's parameters are $\alpha_1 = [-8, 2, 1]$ and $\alpha_2 = [-5, 4, 1]$. The number of observations is 1000, the values of x are uniformly distributed on $[-10, 10]$. The response has contaminated distribution with level $\varepsilon = 0.05$. Contamination

also has uniform distribution. And the parameter of the generalized radical estimator has the value $\lambda = 1$ (this case is equivalent to conditionally optimal estimator with parameter $k^2 = \infty$) The results of MLE are $\hat{\alpha}_1 = [-3.45433123701182, 0.149008965849041, 0.214074248534943]$ and $\hat{\alpha}_2 = [-1.3719459314693, 0.907693561128312, 0.149007714343864]$. And corresponding results of GRE are $\hat{\alpha}_1 = [-11.8998482654318, -1.09199066567912, 0.432142103840774]$ and $\hat{\alpha}_2 = [-7.95057570361495, 7.29031267939047, 1.15240358346135]$.

Figure 1 provide us MLE-estimated probabilities dependence on the covariates. True probabilities are presented by black lines and estimated probabilities by grey. Solid lines correspond to the value of the response $j = 1$, dashed lines to the value $j = 3$ and dash-dot lines to the value $j = 2$.

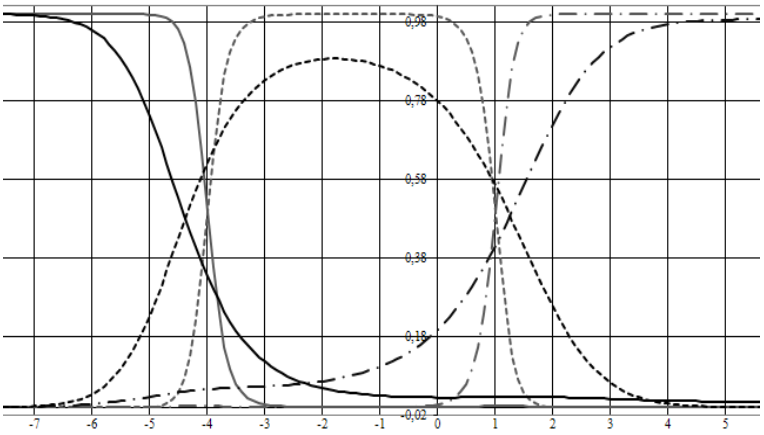


Figure 1: Probabilities estimated by MLE

Figure 2 provide us GRE-estimated probabilities dependence on the covariates. Designations for this figure are the same as for Figure 1.

As the results of the study, robust estimate is less affected by contamination than the MLE estimate. Although robust estimation of parameters quite substantially differ from the true values, dependences of the estimated probabilities are close enough to the true. Hence it is obvious that generalized radical estimator shows more accurate results of probabilities estimation than maximum likelihood estimator.

Conclusions

Due to the results of the research we conclude that:

- the proposed robust method is effective when level of contamination is not too high;
- it is often necessary to use more robust estimator for obtaining good results;

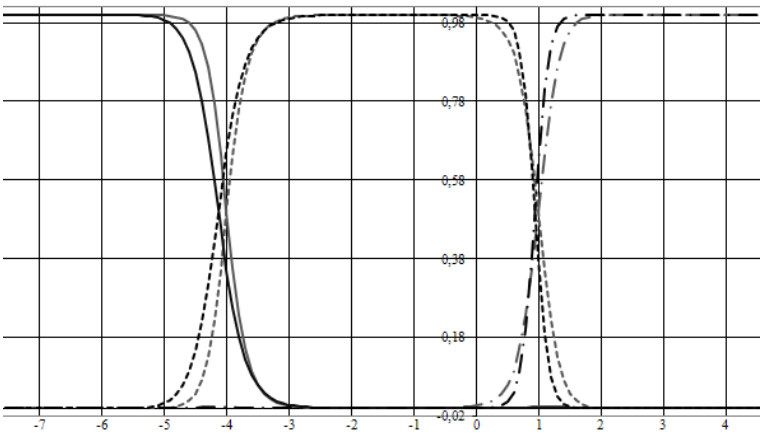


Figure 2: Probabilities estimated by GRE

- high quality of estimation requires a great number of observations;
- whereas methods of estimation are sensitive to initial point it is essential to develop special techniques for obtaining good initial approximation.

Acknowledgements

Research work has been done with a partial support of President of Russian Federation grant (№ МД-2690.2008.9).

References

[1] Bianco A.M., Yohai V.J. (1996). Robust estimation in the logistic regression model. *Robust Statistics, Data Analysis, and Computer Intensive Methods, Lecture Notes in Statistics*, Vol. **109**, pp. 17-34.

[2] Carroll R.J., Pederson S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society*, Vol. **B 55**, pp. 693-706.

[3] Čśampležek P. (2006). Trimmed likelihood-based estimation in binary regression models, *Austrian Journal of Statistics*, Vol. **35**, pp. 223-232

[4] Croux C., Haesbroeck G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, Vol. **44**, pp. 273-295.

[5] Gervini D. (2005). Robust adaptive estimators for binary regression models. *Journal of Statistical Planning and Inference*, Vol. **131**, pp. 297-311.

[6] Hampel F.R., Rouchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust Statistics: The approach based on influence functions*. Wiley, New York.

[7] Huber P.J. (1981). *Robust Statistics*. Wiley, New York.

[8] Kotlyarova Y., Zinde-Walsh V. (2010). Robust estimation in binary choice models. *Communications in Statistics - Theory and Methods*, Vol. **39:2**, pp. 266-279.

[9] Lisitsin D.V. (2009).On estimation of model parameters in presence of Bayesian dot contamination. *Reports of Russian higher education academy of sciences*. № **1(12)**, pp. 41-55 (in Russian).

[10] Maronna R.A., Martin R.D., Yohai V.J. (2006). *Robust Statistics: Theory and Methods*. Wiley, England.

[11] Pregibon D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, Vol. **38**, pp. 485-498.

[12] Shurygin A.M. (2009). *Mathematical methods of prediction*. Goryachaya Liniya - Telecom, Moscow (in Russian).

[13] Victoria-Feser M.-P. (2002). Robust inference with binary data. *Psychometrika*, Vol. **67**, pp. 21-32.

Statistical Forecasting for Censored Autoregressive Time Sefies

Yu.S. Kharin, I.A. Badziahin
Belarusian State University
Minsk, Belarus
e-mail: kharin@bsu.by

Abstract

Problems of optimal statistical forecasting are considered for autoregressive time series observed under distortions generated by interval censoring. If the model parameters are unknown, then the maximum likelihood estimators are found and the "plug-in" forecasting statistic can be constructed. Numerical results are given.

Keywords: Autoregression, censoring, log-likelihood function, mean-square risk.

Introduction

Autoregressive model of order p (AR(p)) is widely used to describe stochastic processes in many fields, such as economy, finance, meteorology, medicine [1]. The case of "complete data" for this model, where all observations are exactly known, is well studied. In practice, however, time series are usually observed under different distortions [2].

In this paper we consider distortions generated by interval censoring, that are often in engineering, economics, business, etc [3]. Censoring means that exact values of some observations are unknown and it is only known that they belong to certain given intervals. Interval censoring appears in real data because of detection limits of measuring devices, high costs of measurement, disorders of equipment, etc[3].

In statistical literature there is a lot of publications devoted to analysis of censored random samples, especially in the reliability theory and medicine studies. In papers of Li and Zhang[4] and Gomez *et al.*[5] regression data model is considered under censored response [4] or explanatory [5] variables; ML- and M-estimators for the regression parameters are proposed. There are significantly less results devoted to censored time series. In paper of Park *et al.*[3] the censored autoregressive time series is considered and imputation method for estimation of model parameters is proposed.

In this paper we construct the optimal forecasting statistic and propose an approach to estimate model parameters based on a special approximation of the log-likelihood function [7, 8].

1 Mathematical model of censored AR(p)

Consider the AR(p) time series model [1]

$$x_t = \sum_{i=1}^{p} \theta_i x_{t-i} + u_t, t \in \mathbf{Z}, \tag{1}$$

where $\theta_1, \ldots, \theta_p$ are unknown coefficients of the autoregression; all roots of the characteristic polynomial $z^p - \sum_{i=1}^{p} \theta_i z^{p-i}$ are inside the unit circle; $\{u_t\}$ are i.i.d. normal random variables, $\mathbf{E}\{u_t\} = 0,\ \mathbf{D}\{u_t\} = \sigma^2 < +\infty$.

Instead of the true values $x_1, \ldots, x_T$ we observe only random events:

$$A_t^* = \{x_t \in A_t\},\ t \in \{1, \ldots, T\}, \tag{2}$$

where $\{A_i\}$ are some known Borel sets, T is the length of the observation process. In this paper we consider two possible cases: 1) $A_t = \{x_t\}$ is a singleton, then the value of the t-th observation x_t is known; 2) $A_t = [a_t, b_t)$ is an interval, where a_t and b_t $(a_t < b_t)$ are known functions, then the observation x_t is censored.

A forecasting statistic for the future value $\hat{x}_{T+1}$ is a number function of the observed events:

$$\hat{x}_{T+1} = f\left(A_1^*,\ A_2^*, \ldots, A_T^*\right). \tag{3}$$

The conditional risk

$$r^* = \mathbf{E}\left\{\left(\hat{x}_{T+1} - x_{T+1}\right)^2 | A_1^*,\ A_2^*, \ldots, A_T^*\right\} \tag{4}$$

is the conditional mean-square error of forecasting under events $\{A_i^*\}_{i=1}^T$.

The censored time series (1), (2) can be represented as a sequence of fragments with fully observed data and fragments with fully censored data. Let τ_i be the length of the i-th fragment of censored data and t_i^* be the initial time moment of this fragment, $i \in \{1, \ldots, M\}$, M be the number of the censored fragments.

2 Optimal forecasting under known model parameters

At first consider the situation where all model parameters are known. Construct optimal forecasting statistic (3) which has the minimal conditional risk (4).

Theorem 1. *The optimal one-step forecasting statistic minimizing the conditional risk (4) for the time series (1), (2) is*

$$\hat{x}_{T+1} = \mathbf{E}\left\{x_{T+1} | A_T^*, \ldots A_1^*\right\}; \tag{5}$$

the minimal risk of forecasting

$$r^* = \mathbf{D}\left\{x_{T+1} | A_T^*, \ldots A_1^*\right\}.$$

Let us analyze more difficult case where the last q $(1 \le q \le T-p)$ observations be censored and other observations be exactly known. Introduce the notation:

$$P_{t_1}^{t_2} = [a_{t_1}, b_{t_1}) \times [a_{t_1+1}, b_{t_1+1}) \times \ldots \times [a_{t_2}, b_{t_2}) \in \mathbf{R}^{t_2-t_1+1},\ t_1 \le t_2;$$

$$\varphi(x) = \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) = \int\limits_{-\infty}^{x} \varphi(t)dt -$$

the standard normal probability density function and the standard normal distribution function;

$$\Psi(x, y, m, s, u, v) = \tfrac{u\varphi((x-m)/s) - v\varphi((y-m)/s)}{\Phi((y-m)/s) - \Phi((x-m)/s)}.$$

Theorem 2. *Let values $x_1, \ldots, x_{T-q}$ and events $A^*_{T-q+1}, \ldots A^*_T$ be observed. Then the optimal forecasting statistic is:*

$$\hat{x}_{T+1} = \frac{\int\limits_{P^T_{T-q+1}} \sum\limits_{i=1}^{p} \theta_i x_{T-i+1} p\left(x_T, \ldots x_{T-q+1} \,|x_{T-q}, \ldots x_1\right) dx_{T-q+1} \ldots dx_T}{\int\limits_{P^T_{T-q+1}} p\left(x_T, \ldots x_{T-q+1} \,|x_{T-q}, \ldots x_1\right) dx_{T-q+1} \ldots dx_T}. \tag{6}$$

Corollary 1. *If $a_T \to b_T, \ldots, a_{T-q+1} \to b_{T-q+1}$, then the optimal forecasting statistic and its risk are*

$$\hat{x}_{T+1} = \sum_{i=1}^{p} \theta_i x_{T-i+1}, \quad r^* = \sigma^2.$$

In Corollary 1 we consider the asymptotic case of "complete data": $x_T = a_T = b_T, \ \ldots, \ x_{T-q+1} = a_{T-q+1} = b_{T-q+1}$. The results indicated in Corollary 1 coincide with the well known results for this case [1].

Let us consider the AR(1) time series. In this case it was proved that the optimal forecasting statistic depends only on the last observed value of the time series and all random events preceding this value. If all T observations are censored, the following result is proved.

Theorem 3. *For the AR(1) model (1) if the events $A^*_1, \ldots, A^*_T$ are observed $(q = T)$, then the optimal forecasting statistic is:*

$$\hat{x}_{T+1} = \theta_1 \frac{\int\limits_{P_1^T} x_T p(x_T, \ldots x_1) dx_1 \ldots dx_T}{\int\limits_{P_1^T} p(x_T, \ldots x_1) dx_1 \ldots dx_T}.$$

Corollary 2. *For the AR(1) model (1) if $a_T \to -\infty, \ \ldots, \ a_{T-q+1} \to -\infty, b_T \to +\infty, \ldots, b_{T-q+1} \to +\infty$, then the optimal forecasting statistic and its risk are*

$$\hat{x}_{T+1} = \theta_1^{q+1} x_{T-q}, \quad r^* = \sigma^2 \sum_{i=0}^{q} \theta^{2i}.$$

Conditions of Corollary 2 mean that at the time moments $T, \dots, T-q+1$ the observations $x_T, \dots, x_{T-q+1}$ "become" missing. The results indicated in Corollary 2 coincide with the known results for this case [2].

Consider the case $q = 1$. In this case the optimal forecasting statistic can be calculated in explicit form.

Theorem 4. *For the AR(1) model (1) if the value x_{T-1} and the random event A_T^* are observed, then the optimal forecasting statistic and its conditional risk are*

$$\hat{x}_{T+1} = \theta_1^2 x_{T-1} + \theta\sigma\Psi(a_T, b_T, \theta x_{T-1}, \sigma, 1, 1), \tag{7}$$

$$r^* = (1+\theta^2)\,\sigma^2 - \left(\theta\sigma\Psi(a_T, b_T, \theta x_{T-1}, \sigma, 1, 1)\right)^2 +$$

$$+\theta^2\sigma\Psi\left(a_T, b_T, \theta x_{T-1}, \sigma, a_T - \theta x_{T-1}, b_T - \theta x_{T-1}\right).$$

Corollary 3. *Let the assumptions of Theorem 4 take place and $\tau_T = b_T - a_T \to 0$. Then the asymptotic expansion for the conditional risk is*

$$r^* = \sigma^2 + \theta_1^2\frac{\tau_T^2}{12} - \theta_1^2\tau_T^4\frac{3a_T^2 - 6a_T\theta_1 x_{T-1} + 3\theta_1^2 x_{T-1}^2 + 2\sigma^2}{720\sigma^4} + o(\tau_T^4).$$

It is known [1] that for the case of "complete data" the risk of the optimal one-step forecasting statistic is $r_0 = \sigma^2$. To evaluate the sensitivity of the risk of one-step forecasting to the length $\tau_T = b_T - a_T$ of the censoring interval $[a_T, b_T)$ we will use the risk sensitivity coefficient [2]:

$$\chi = (r^* - r_0)/r_0. \tag{8}$$

Corollary 4. *For the forecasting statistic (7) the risk sensitivity coefficient (8) has the following approximation:*

$$\chi \approx \frac{\theta_1^2}{12\sigma^2}\tau_T^2.$$

3 ML-estimators of the model by censored data

Usually in practice the model parameters are unknown. In this case at first we will estimate parameters of the model (1), and then we will construct the "plug-in" forecasting statistic [2] using optimal forecasting statistics constructed in section 2. The "plug-in" forecasting statistic can be constructed from optimal forecasting statistic by replacing exact values of parameters by their ML-estimators.

To simplify our results we consider here only the case of AR(1) model. However, all our results of this section can be generalized for the case of $p > 1$.

Introduce the notation: $\delta_{i,j}$ is the Kronecker symbol;

$$\mathrm{I}_\nu(a, b, m, s) = \int_a^b (t-m)^\nu \frac{1}{s}\varphi\left(\frac{t-m}{s}\right) dt,\ \ \nu \in \mathbf{N} \cup \{0\},\ \ y, m, s \in \mathbf{R}; \tag{9}$$

$$\nu(l;k,i_1,\ldots,i_k)=\sum_{j=1}^{k}\left(\delta_{l,i_j}+\delta_{l,i_j+1}\right),\ l,k,i_1,\ldots,i_k\in\mathbf{N};$$

$$d(l;A,\theta,\sigma)=\begin{cases}\sigma, & l\in A;\\ \frac{\sigma}{\sqrt{1+\theta^2}}, & l\notin A,\end{cases}\quad l\in\mathbf{N},\ A\subset\mathbf{N},\ \theta,\sigma\in\mathbf{R};$$

$$l^*(t_1,t_2|\theta,\sigma)=\sum_{t=t_1+1}^{t_2}\ln\frac{1}{\sigma}\varphi\left(\frac{x_t-\theta x_{t-1}}{\sigma}\right),\ 0<t_1<t_2\le T;$$

$$F_1(t^*,\tau|\theta,\sigma)=\frac{1}{\sigma}\varphi\left(\frac{x_{t^*+\tau}-\theta^{\tau+1}x_{t^*-1}}{\sigma\sqrt{\frac{1-\theta^{2\tau+2}}{1-\theta^2}}}\right)(1+\theta^2)^{-\frac{\tau}{2}}\left(\sum_{k=1}^{+\infty}\frac{\theta^k}{k!}\times\right.$$

$$\times\sum_{i_1,\ldots,i_k=1}^{\tau-1}\prod_{l=1}^{\tau}\mathrm{I}_{\nu(l;k,i_1,\ldots,i_k)}\left(a_{t^*+l-1},b_{t^*+l-1},\bar{\mu}_l,d(l;\emptyset,\theta,\sigma)\right)+$$

$$\left.+\prod_{l=1}^{\tau}\mathrm{I}_0\left(a_{t^*+l-1},b_{t^*+l-1},\bar{\mu}_l,d(l;\emptyset,\theta,\sigma)\right)\right),\qquad(10)$$

$$\bar{\mu}_l=\theta^l x_{t^*-1}+\theta^{\tau-l}\frac{1-\theta^{2l}}{1-\theta^{2\tau+2}}(x_{t^*+\tau}-\theta^{\tau+1}x_{t^*-1});$$

$$F_2(t^*,\tau|\theta,\sigma)=\frac{\sqrt{1-\theta^2}}{\sigma}\varphi\left(\frac{x_{t^*+\tau}\sqrt{1-\theta^2}}{\sigma}\right)(1+\theta^2)^{\frac{1-\tau}{2}}\times\left(\sum_{k=1}^{+\infty}\frac{\theta^k}{k!}\times\right.$$

$$\times\sum_{i_1,\ldots,i_k=1}^{\tau-1}\prod_{l=1}^{\tau}\mathrm{I}_{\nu(l;k,i_1,\ldots,i_k)}\left(a_{t^*+l-1},b_{t^*+l-1},\theta^{\tau-l+1}x_{t^*+\tau},d(l;\{1\},\theta,\sigma)\right)$$

$$\left.+\prod_{l=1}^{\tau}\mathrm{I}_0\left(a_{t^*+l-1},b_{t^*+l-1},\theta^{\tau-l+1}x_{t^*+\tau},d(l;\{1\},\theta,\sigma)\right)\right);\qquad(11)$$

$$F_3(t^*,\tau|\theta,\sigma)=(1+\theta^2)^{\frac{1-\tau}{2}}\times$$

$$\times\left(\prod_{l=1}^{\tau-1}\mathrm{I}_0\left(a_{t^*+l-1},b_{t^*+l-1},\theta^l x_{t^*-1},d(l;\{\tau\},\theta,\sigma)\right)+\right.\qquad(12)$$

$$\left.+\sum_{k=1}^{+\infty}\frac{\theta^k}{k!}\sum_{i_1,\ldots,i_k=1}^{\tau-1}\prod_{l=1}^{\tau}\mathrm{I}_{\nu(l;k,i_1,\ldots,i_k)}\left(a_{t^*+l-1},b_{t^*+l-1},\theta^l x_{t^*-1},d(l;\{\tau\},\theta,\sigma)\right)\right).$$

Note that (9) are calculated [6]. Also note that the functions (10) – (12) cannot be explicitly calculated in practice, because they are represented as infinite functional series; we use a finite number (k_{max}) of terms to approximate these functions.

Theorem 5. *For the interval censored AR(1) model (1), (2) with $M > 1$ censored fragments the log-likelihood function is*

$$l(\theta_1,\sigma) = \delta_{t_1^*,1}\ln F_2(t_1^*,\tau_1|\theta_1,\sigma) + (1-\delta_{t_1^*,1})\left(\ln\varphi\left(\frac{x_1\sqrt{1-\theta_1^2}}{\sigma}\right)+\right.$$

$$\left.+\ln\frac{\sqrt{1-\theta_1^2}}{\sigma} + l^*(1,t_1^*-1|\theta_1,\sigma) + \ln F_1(t_1^*,\tau_1|\theta_1,\sigma)\right)+$$

$$+\sum_{i=2}^{M-1}\left(l^*(t_{i-1}^*+\tau_{i-1},t_i^*-1|\theta_1,\sigma)+\ln F_1(t_i^*,\tau_i|\theta_1,\sigma)\right)+ \tag{13}$$

$$+l^*(t_{M-1}^*+\tau_{M-1},t_M^*-1|\theta_1,\sigma)+\delta_{t_M^*+\tau_M,T}\ln F_3(t_M^*,\tau_M|\theta_1,\sigma)+$$

$$+(1-\delta_{t_M^*+\tau_M,T})\left(\ln F_1(t_M^*,\tau_M|\theta_1,\sigma)+l^*(t_M^*+\tau_M,T|\theta_1,\sigma)\right).$$

Maximum likelihood estimator (MLE) of (θ_1,σ) is the solution of the maximization problem:

$$(\hat\theta_1,\hat\sigma) = \arg\max_{\theta_1,\sigma} l(\theta_1,\sigma). \tag{14}$$

4 Numerical results

At first, computer experiments are performed for the case of the AR(1) model and $q = 1$ to compare the optimal forecasting statistic and some statistics used in practice:

$$\hat x'_{T+1} = \theta_1\mathbf{E}\left\{x_{T+1}\,|A_T^*\right\}, \tag{15}$$

$$\hat x''_{T+1} = \theta_1\frac{a_T+b_T}{2}. \tag{16}$$

For simulations the following values of parameters are used: $\theta = 0.8$, $\sigma^2 = 1$, $T = 100$, $\tau_T = b_T - a_T \in \{0, 0.5, \ldots, 15\}$. The last observation of the time series is replaced by the random censoring interval $[a_T, b_T)$: the length of the interval $[a_T, x_T)$ is $\alpha\tau$ and the length of the interval $[x_T, b_T)$ is $(1-\alpha)\tau$, where α is the standard uniformly distributed random variable. For every fixed τ_T the Monte-Carlo experiments with 10000 simulations of time series are used to evaluate the experimental values of the conditional risk.

The results of these experiments are given in the Figure 1. It is seen that the optimal forecasting statistic is more preferable even for short intervals of censoring.

To estimate the model parameters the following approach is widely used [3]: each censored observation x_t is replaced by the lower bound a_t, and then classical estimators, for example, the least square estimator (LSE), are computed. Computer experiments are performed to compare MLE and LSE described above for right censored times series ($b_t \to +\infty$, $t \in [1, T]$). The experiment consists of the following: 1) autoregressive time series of the length $T = 300$ is

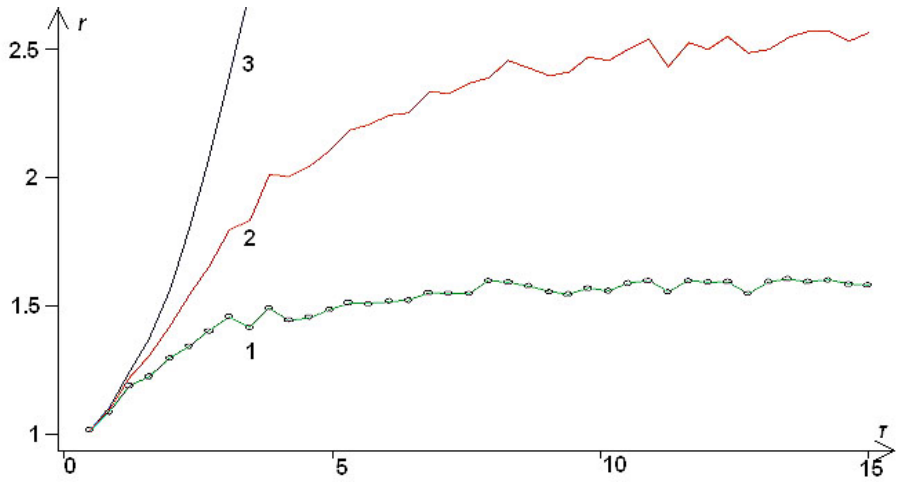


Figure 1: Comparison of forecasting statistics: 1 - risk of the optimal forecasting statistic (7), 2 - risk of the forecasting statistic (15), 3 - risk of the forecasting statistic (16)

generated with parameters $p = 1$, $\theta_1 = -0.3$, $\sigma = 1$; 2) if $x_t \geq a_t$, $a_t \equiv 0$, then the random event $A_t^* = \{x_t \geq a_t\}$ is observed; 3) the MLE and the LSE of the parameter θ_1 are calculated under the assumption that the parameter σ is known. The MLE of the θ_1 can be approximately found by the tabulation of the likelihood function (13) with certain accuracy. To calculate (13) $k_{max} = 6$ or $k_{max} = 10$ summands in (10) – (12) are used.

Empirical means of the estimators and of the errors based on 100 experiments are given in the Table 1. It is seen that the MLE gives a significant gain in accuracy.

Table 1: Comparison of MLE and LSE

LS		ML			
		$k_{max} = 6$		$k_{max} = 10$	
LSE	Error	MLE	Error	MLE	Error
0.180923	0.480923	-0.3088	0.041029	-0.30081	0.038993

Conclusion

For autoregressive time series observed under distortions generated by interval censoring the following results are presented in this paper.

1) The optimal forecasting statistic minimizing the conditional risk is found.

2) The log-likelihood function and the maximum likelihood estimators are constructed.

3) The comparison between the optimal forecasting statistic and two used in practice statistics is made. Numerical results confirm that the optimal forecasting statistic is more preferable even for short intervals of censoring.

4) The comparison of the maximum likelihood estimator and the least square estimator of parameters for the censored time series is made. Numerical results illustrate a significant gain in the accuracy for the MLE w.r.t. the LSE.

References

[1] Box G.E.P., Jenkins G.M. (1970) *Time Series Analysis: Forecasting and Control*. Holden-Day, San-Francisco.

[2] Kharin Yu.S. (2008) *Optimality and robustness in statistical forecasting*. BSU, Minsk. (in Russian)

[3] Park J.W., Genton M.G., Ghosh S.K. (2007). Censored Time Series Analysis with Autoregressive Moving Average Models. *The Canadian Journal of Statistics*. Vol. **35**, pp. 151-168.

[4] Li G., Zhang C.-H. (1998). Linear regression with interval censored data. *The Annals of Statistics*. Vol **26**, pp. 1306-1327.

[5] Gomez G., Espinal A., Lagakos.W. (2003) Inference for a Linear Model with an Interval-censored Covariate. *Statistics in Medicine*. Vol. **22**, pp. 409-425.

[6] Ryzhik I.M., Gradshteyn I.S. (2007). *Table of Integrals, Series, and Products, seventh edition*. Academic Press.

[7] Badziahin I., Kharin Yu. (2009). On robustness of forecasting for AR(p) time series under interval censoring. *Book of abstracts of International Conference on Robust Statistic 2009*. p. 9.

[8] Badziahin I., Kharin Yu. (2010). On estimation of parameters for autoregressive time series under left censoring. *Book of abstracts of International Conference on Robust Statistic 2010*. pp. 9-10.

Robust Estimation of Count Response Regression Models

Svetlana Yu. Dovgal, Daniil V. Lisitsin
Novosibirsk State Technical University
Novosibirsk, Russian Federation
e-mail: carcinoma.nsk@gmail.com

Abstract

This paper is concerned with models of event counts, particularly with the Poisson regression model examination. Robust methods of M-estimation parameters were researched. Expressions of the score function for Poisson model were given. Maximum likelihood estimation and M-estimation for model's parameters were compared by simulation.

Keywords: count data, Poisson regression, robustness, M-estimation, influence function.

Introduction

While doing practical researches we often have to operate with data being the result of counting. Count data can be defined as the quantity of appearing of some event during the specified time period. For instance, the amount of earthquakes occurred during a year, the number of car accidents, the number of university graduates and etc. Such counting has place in many fields of activity such as economics, sociology, insurance and others. The regression model in this case connects the response - the amount of events happened - with factors, which characterize the accompanying conditions [1]. The Poisson model and its generalization - negative binomial model - have become wide spread for modeling of suchlike variables.

Real data may not correspond to the theoretical distribution, there one can meet gross errors - "outliers". In this case we should use estimation methods which are not sensible to suchlike change of data. This feature is not usually a characteristic of the classical statistical estimation procedures. Particularly, the estimation of Poisson regression parameters, by using the maximum likelihood method, is not stable. The estimation having stable characteristics is called robust. At present there exist several classes of such estimation. We will use robust M-estimation of the regression model parameters. The one of the approaches to get robust M-estimation was suggested by Shurygin [2] (see also [3]). It based on the Bayesian dot contamination of model distribution. Also the Shurygin's approach was only used for modeling of continuous random variables, but it can be easily applied while working with discrete variables.

1 Model Specification

Let z be the Poisson random variable with density

$$\mathbf{P}(z = j|\lambda) = P(j, \lambda) = \frac{e^{-\lambda}\lambda^j}{j!},\ \lambda > 0.$$

In Poisson regression the following parameterization of parameter λ is used:

$$\lambda = \lambda(x, \theta) = \exp[f(x)\theta],$$

where $f(x)$ is a vector of explanatory variables, θ is a vector of regression parameters. As a result the Poisson regression can be presented like

$$\mathbf{P}(z = j|x, \theta) = P(j, x, \theta) = \frac{e^{-f(x)\theta}[f(x)\theta]^j}{j!}.$$

Define the sample of observations as $(x_i, z_i), i = 1, 2, ..., N$. We get M-estimation $\hat{\theta}$ of parameter θ by solving the equations system

$$\sum_{i=1}^{N} \psi(z_i, x_i, \hat{\theta}) = 0,$$

where ψ is a vector score function satisfying the following condition

$$\sum_{k=0}^{\infty} P\left(k, x_i, \theta\right) \psi(k, x_i, \theta) = 0. \tag{1}$$

2 Robust estimation

Bayesian dot contamination distribution model is organized as a mixture

$$\mathbf{P}\left(z_i = j|x_i, \theta, z_i^*\right) = (1 - \varepsilon)\mathbf{P}\left(z_i = j|x_i, \theta\right) + \varepsilon\delta_{jz_i^*},$$

where z_i^* is discrete random variable with distribution $\mathbf{P}\left(z_i^* = j|x_i, \theta\right) = S\left(j, x_i, \theta\right)$, ε - the level of contamination($0 < \varepsilon < 0.5$), δ - Kronecker delta.

Indicator of estimation badness in Bayesian dot contamination model can be written as functional

$$U_s(\psi) = \sum_{i=1}^{N} \sum_{k=0}^{\infty} IF\left(k, x_i, \theta\right) IF^T\left(k, x_i, \theta\right) S\left(k, x_i, \theta\right),$$

where $IF\left(j, x, \varphi\right)$ is the Hampel's influence function, under some conditions having the form $IF(j, x, \theta) = M^{-1}\psi(j, x, \theta)$, $M = \sum_{i=1}^{N} \sum_{k=0}^{\infty} \psi(k, x_i, \theta) \frac{\partial}{\partial \theta^T} P\left(k, x_i, \theta\right)$.

The functional has its minimum value in the estimation function [3]

$$\psi(z, x, \theta) = C(\theta) \left[\frac{\partial}{\partial \theta} \ln P(z, x, \theta) + \beta\right] \frac{P(z,x,\theta)}{S(z,x,\theta)},$$

where $C(\theta)$ is a nonsingular matrix, vector $\beta = \beta(x, \theta)$ provides the fulfillment of the condition (1).

For the Poisson regression it has the form

$$\psi(z, x, \theta) = C(\theta)\psi_\lambda\left(z, \mathrm{e}^{f(x)\theta}\right) \exp[f(x)\theta] f^T(x),$$

where $\psi_\lambda(z,\lambda) = c\left(\frac{z}{\lambda} - 1 + \beta_\lambda\right)\frac{P(z,\lambda)}{S(z,\lambda)}$ is an optimal score function for the Poisson distribution parameter, c is a nonzero constant, β_λ provides the fulfillment of the condition (1), $S(j,\lambda)$, $j = 0,1,2,...$ is a distribution of contamination value.

We can get generalized radical estimates after fixing the distribution of contamination value in the form

$$S(z,\lambda) = P^{1-\alpha}(z,\lambda)/\Delta(\lambda,\alpha),$$

where α - parameter of estimation ($\alpha \geq 0$), value $\Delta(\lambda,\alpha)$ is either $\sum\limits_{k=0}^{\infty} P^{1-\alpha}(k,\lambda)$ (it is used to provide the condition of the normalization of number of probabilities, that is possible only when $\alpha < 1$), or equal one, if the given condition is not used. Note that the case when $\alpha = 0$ corresponds the maximum likelihood estimation.

The score function of generalized radical estimation of parameter λ has the form

$$\psi_\lambda(z,\lambda) = \left(\tfrac{z}{\lambda} - 1 + \beta_\lambda\right) P^\alpha(z,\lambda)\Delta(\lambda,\alpha),$$

where $\beta_\lambda = 1 - \frac{1}{\lambda}\sum\limits_{k=0}^{\infty} k\lambda^{(1+\alpha)k}\big/[k!]^{1+\alpha}\Big/\sum\limits_{k=0}^{\infty}\lambda^{(1+\alpha)k}\big/[k!]^{1+\alpha}$.

As a result we got score function for estimation of parameter θ :

$$\psi(z,x,\theta) = \left\{ z - \frac{\sum\limits_{k=0}^{\infty} k\left[\mathrm{e}^{f(x)\theta}\right]^{(1+\alpha)k}\Big/[k!]^{1+\alpha}}{\sum\limits_{k=0}^{\infty}\left[\mathrm{e}^{f(x)\theta}\right]^{(1+\alpha)k}\Big/[k!]^{1+\alpha}} \right\} P^\alpha(z,\mathrm{e}^{f(x)\theta})\Delta(\mathrm{e}^{f(x)\theta},\alpha) f^T(x).$$

3 Experimental research

To check the working capacity of the suggested approaches the experimental research was carried out. Maximum likelihood estimation and M-estimation for model's parameters were compared. The model with scalar parameter was chosen: $f(x) = x$ - explanatory variable, $\lambda = \exp[\theta x]$ - parameter of Poisson distribution. True value of parameter θ is $\theta_{true} = 1$. The values of x are uniformly distributed on the interval $[-1,1]$. Distribution of observations is the mixture of model and contamination distributions. Contamination simulated like a Poisson distribution with parameter $\tilde{\lambda}$ is equal 3λ and with different contamination levels $\varepsilon \in \{0.05, 0.15, 0.3\}$. Size of the sample is 1000. Parameter of estimator is $\alpha = 0.5$, function $\Delta(\lambda,\alpha) = \sum\limits_{k=0}^{\infty} P^{1-\alpha}(k,\lambda)$.

Figure 1 represents a diagram with dependence of the standard deviation

$$\sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(\hat{\theta}_i - \theta_{true}\right)^2}$$

on the contamination level ε for M-estimation and maximum likelihood estimation (MLE), where $K = 1000$ is the number of simulations, $\hat{\theta}_i$ is a estimation value of the i-th simulation.

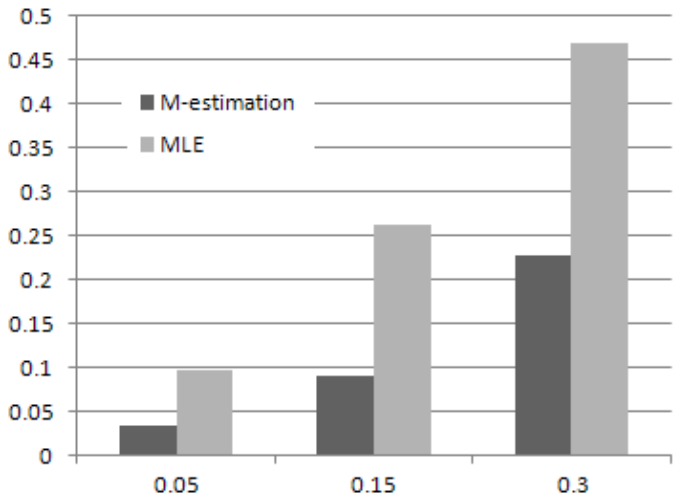


Figure 1: Standard deviation of estimation

Conclusions

Having done this research work we can make the following conclusions:

- the M-estimations that we got have robustness properties for examined contamination levels;
- the estimations of the maximum likelihood method turn out to be much worse and they are not robust even in the conditions of little contamination.

References

[1] Winkelmann R. (2008) Econometric Analysis of Count Data. *Springer Verlag, Heidelberg.*

[2] Shurygin A.M. (2000). Applied Stochastics: Robustness, estimation, prediction. *Finances and Statistics*, Moscow (in Russian).

[3] Lisitsin D.V. (2009). *On estimation of model parameters in presence of Bayesian dot contamination.* Reports of Russian higher education academy of sciences. № 1(12), pp. 41-55 (in Russian).

Part VII
Statistical Simulation of Natural Processes

Modeling of Nonstationary Processes with Periodic Properties on Basis of Markov Chains

Nina A. Kargapolova, Lev Ya. Saveliev, Vasily A. Ogorodnikov
Institut of computational Mathematics and Mathematical Geophysics SB RAS,
Sobolev's Institut of Mathematics SB RAS,
Novosibirsk State University,
Novosibirsk, Russia
e-mail: ova@osmf.sscc.ru

Abstract

In this paper heterogeneous Markov model with two states and periodic transition probability matrix is considered. Expressions for limiting probabilities and distributions of long-term identical value runs are obtained. On basis of real data, model is applied to investigation of air temperature's long-term overshoots.

Keywords: heterogeneous Markov chain, limiting probabilities, air temperature.

Introduction

During investigation of real time-series, for example meteorological (cloud amount, amount of precipitation, etc.), appears necessity of random sequences with finite number of states modeling. Such models may be used for modeling of every meteorological series, on condition that not all possible values of real process, but only some their gradations are considered. Several approaches to modeling of discrete sequences with given probabilistic properties are possible. Methods, based on different order Markov chains, are used quite often [1,2,3]. Another widespread method of discrete time-series modeling is based on threshold transformation of specially selected Gaussian process [2]. In this case dependencies of time-series elements are determined by correlation function. Meteorological parameters, such as air temperature, wind speed, humidity, etc. possess daily and seasonal variation. Therefore, it is necessary to use methods, which take into account periodical properties of statistical characteristics of considered processes. In connection with this in [1] different types of periodically correlated random processes are investigated. In the same book some ranges of their application to description of real oceanological processes are also considered.

In this paper model, based on scalar heterogeneous Markov chain with two states and periodic transition probability matrix is considered. Such models may be used for modeling of precipitation indicators, air temperature overshoots with taking into consideration of daily variation. For model construction long-term meteorological data, obtained in Astrakhan, are used.

1 Heterogeneous Markov chain with periodic transition probability matrix

1.1 Definition of process $\xi(k)$

Let's consider binary heterogeneous Markov sequence ξ of random variables $\xi_k,\ k \geq 0$ with set of states $C = \{1, 0\}$, initial probability vector A and transition probability matrixes Q, R:

$$A = (a_1,\ a_0) = (a,\ 1-a),$$

$$Q = \begin{pmatrix} q_{11} & q_{10} \\ q_{01} & q_{00} \end{pmatrix} = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix},$$

$$R = \begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix} = \begin{pmatrix} r & 1-r \\ 1-s & s \end{pmatrix},$$

$$(a = a_1,\ \ p = q_{11},\ \ q = q_{00},\ \ \ r = r_{11},\ \ \ s = r_{00}),$$

$$\begin{gathered} \Pr[\xi_0 = \alpha] = a_\alpha, \\ \Pr[\xi_{2i+1} = \beta \,|\xi_{2i} = \alpha] = q_{\alpha\beta}, \\ \Pr[\xi_{2i+2} = \beta \,|\xi_{2i+1} = \alpha] = r_{\alpha\beta}, \\ i \geq 0,\ \ \alpha = 0, 1,\ \ \beta = 0, 1. \end{gathered}$$

Transition matrixes Q, R are used in turn, starting from Q. Matrixes Q and R are stochastic matrixes, so they are defined with four independent parameters $p,\ q,\ r,\ s \in [0, 1]$. These parameters express probabilities of conservation of value during transition $\xi_k \to \xi_{k+1}$ depending on evenness of k.

General form of matrix of transition probability

$$p_{\alpha\beta}[k] = \Pr[\xi_{k+1} = \beta \,|\xi_k = \alpha]$$

is

$$P[k] = (1 - \theta[k])Q + \theta[k]R = \begin{pmatrix} p_{11}[k] & p_{10}[k] \\ p_{01}[k] & p_{00}[k] \end{pmatrix},$$

where function $\theta[2i-1] = 0, \quad \theta[2i] = 1, \quad i \geq 1$ describes evenness of k. So

$$p_{\alpha\beta}[k] = (1 - \theta[k])q_{\alpha\beta} + \theta[k]r_{\alpha\beta}.$$

Particularly,

$$P[1] = (1 - \theta[1])Q + \theta[1]R = Q, \quad P[2] = (1 - \theta[2])Q + \theta[2]R = R.$$

It will be convenient to use form of general transition probability matrix without indexes

$$P[k] = \begin{pmatrix} p[k] & 1-p[k] \\ 1-q[k] & q[k] \end{pmatrix}$$

where

$$p[k] = p_{11}[k] = \Pr[\xi[k+1] = 1 \,|\xi[k] = 1],$$

$$q[k] = p_{00}[k] = \Pr[\xi[k+1] = 0 \,|\xi[k] = 0].$$

1.2 Distribution of ξ_k

It is obvious that

$$\prod_{k=1}^{2m} P[k] = (QR)^m, \quad \prod_{k=1}^{2m+1} P[k] = (QR)^m Q, \quad m \geq 1 \ .$$

Let's define

$$S = QR = \begin{pmatrix} s_{11} & s_{10} \\ s_{01} & s_{00} \end{pmatrix} = \begin{pmatrix} q_{10}r_{01} + q_{11}r_{11} & q_{10}r_{00} + q_{11}r_{10} \\ q_{00}r_{01} + q_{01}r_{11} & q_{00}r_{00} + q_{01}r_{10} \end{pmatrix}$$

and

$$\begin{gathered} t = detR = r + s - 1, \\ u = detQ = p + q - 1, \\ d = detS = detQdetR = (p + q - 1)(r + s - 1), \\ b = (r - qt)/(1 - d), \ \ d \neq 1. \end{gathered}$$

Then

$$S = QR = \begin{pmatrix} 1-s+pt & s-pt \\ r-qt & 1-r+qt \end{pmatrix} = \begin{pmatrix} 1-s & s \\ r & 1-r \end{pmatrix} + t \begin{pmatrix} p & -p \\ -q & q \end{pmatrix}.$$

It is possible to prove by induction that next equality for $m-$th power of S takes place

$$S^m = \begin{pmatrix} b + (1-b)d^m & 1 - b - (1-b)d^m \\ b - bd^m & 1 - b + bd^m \end{pmatrix}.$$

If $m \geq 1$, then distribution

$$P[2m] = AS^m = (p_{2m}, 1 - p_{2m})$$

is equal to

$$P[2m] = (b + (a-b)d^m,\ 1 - b - (a-b)d^m),$$
$$p_{2m} = \Pr[\xi_{2m} = 1] = b + (a-b)d^m.$$

And distribution

$$P[2m+1] = AS^mQ = (p_{2m+1}, 1 - p_{2m+1})$$

is equal to

$$P[2m+1] = (1 - q + bu + (a-b)ud^m,\ q - bu - (a-b)ud^m),$$
$$p_{2m+1} = \Pr[\xi_{2m+1} = 1] = 1 - q + bu + (a-b)ud^m.$$

1.3 Limits of $P[2m]$, $P[2m+1]$.

If $|d| < 1$, then even and odd limiting probabilities and their difference are equal to

$$f_\infty = \lim_{m\to\infty} \Pr[\xi_{2m} = 1] = b,$$
$$g_\infty = \lim_{m\to\infty} \Pr[\xi_{2m+1} = 1] = 1 - q + bu,$$
$$g_\infty - f_\infty = (1-q)(1-b) - b(1-p).$$

Equality $g_\infty = f_\infty$ is equivalent to $p = q = 1$ or $b = (1-q)/(1-q)$, when $p + q \neq 2$. Previous inequality is equivalent to $p \neq 1$ or $q \neq 1$, and also $u \neq 1$.

Let $d = ut \neq 1,\ u \neq 1$ and $b = (r - qt)/(1 - ut)$. Then,

$$\frac{r - qt}{1 - ut} = \frac{1 - q}{1 - u},$$

$$r = \frac{(1-q) - (u-q)t}{1-u} = \frac{(1-q) + (1-p)(r+s-1)}{1-u}.$$

So

$$(1-u)r = (1-q) + (1-p)(r+s-1),$$
$$((1-u) - (1-p))r = (1-q)r = (1-q) - (1-p)(1-s),$$
$$(1-q)(1-r) = (1-p)(1-a)$$

and if $s \neq 1,\ q \neq 1$, proportion

$$\frac{1-r}{1-s} = \frac{1-p}{1-q},$$

which connects elements of Q, R, takes place.

1.4 Distribution of 1-runs duration

In this section only stationary chains are considered. It follows from this that

$$P(\xi_{2t-1}=0)=1-g_\infty=1-(1-q+bu)=q-bu,$$

$$P(\xi_{2t}=0)=1-f_\infty=1-b,\quad t\geq 1.$$

If 1-run begins at element with even number, distribution of 1-runs duration can be defined as

$$P(L_1=k)=\frac{P(\xi_{2t-1}=0,\xi_{2t}=1,\ldots,\xi_{2t+k-1}=1,\xi_{2t+k}=0)}{P(\xi_{2t-1}=0,\xi_{2t}=1)},$$

$$k=1,\ 2,\ \ldots\ .$$

If $k=2m,\ \ m=1,2,\ldots$, then

$$P(\xi_{2t-1}=0,\xi_{2t}=1,\ldots,\xi_{2t+2m-1}=1,\xi_{2t+2m}=0)=$$
$$=(q-bu)(1-s)p^m r^{m-1}(1-r),$$

$$P(\xi_{2t-1}=0,\xi_{2t}=1)=(q-bu)(1-s).$$

So

$$P(L_1=2m)=\frac{(q-bu)(1-s)p^m r^{m-1}(1-r)}{(q-bu)(1-s)}=p^m r^{m-1}(1-r).$$

If $k=2m-1,\ \ m=1,2,\ldots$

$$P(L_1=2m-1)=\frac{(q-bu)(1-s)p^{m-1} r^{m-1}(1-p)}{(q-bu)(1-s)}=p^{m-1} r^{m-1}(1-p).$$

When 1-run begins at element with odd number

$$P(L_1=k)=\frac{P(\xi_{2t}=0,\xi_{2t+1}=1,\ldots,\xi_{2t+k}=1,\xi_{2t+k+1}=0)}{P(\xi_{2t}=0,\xi_{2t+1}=1)},\ k=1,\ 2,\ \ldots$$

$$P(\xi_{2t}=0,\xi_{2t+1}=1)=(1-b)(1-q).$$

If $k=2m,\ \ m=1,2,\ldots$

$$P(\xi_{2t}=0,\xi_{2t+1}=1,\ldots,\xi_{2t+2m}=1,\xi_{2t+2m+1}=0)=$$

$$=(1-b)(1-q)p^{m-1}r^m(1-p),$$

$$P(L_1=2m)=\frac{(1-b)(1-q)p^{m-1}r^m(1-p)}{(1-b)(1-q)}=p^{m-1}r^m(1-p).$$

Similarly, for $k = 2m-1,\ \ m = 1, 2, \ldots$

$$P(L_1 = 2m-1) = \frac{(1-b)(1-q)p^{m-1}r^{m-1}(1-r)}{(1-b)(1-q)} = p^{m-1}r^{m-1}(1-r).$$

Average value of 1-runs duration, in case when runs begin at chain's element with even number, equals

$$ML_1 = \sum_{k=1}^{\infty} kP\left(L_1 = k\right) =$$
$$= \sum_{m=1}^{\infty} (2m)\, p^m r^{m-1}\left(1-r\right) + \sum_{m=1}^{\infty} (2m-1)\, p^{m-1}r^{m-1}\left(1-p\right) = \tfrac{1+p}{(1-pr)}.$$

(here we suppose that $pr \neq 1$). Average value of 1-runs duration, in case when runs begin at chain's element with odd number, equals

$$ML_1 = \sum_{m=1}^{\infty} (2m)\, p^{m-1}r^{m}\left(1-p\right) + \sum_{m=1}^{\infty} (2m-1)\, p^{m-1}r^{m-1}\left(1-r\right) = \frac{1+r}{(1-pr)}.$$

Variance of 1-runs duration, in case when runs begin at chain's element with odd number, is

$$ML_1^2 = \sum_{m=1}^{\infty} (2m)^2\, p^m r^{m-1}\left(1-r\right) + \sum_{m=1}^{\infty} (2m-1)^2\, p^{m-1}r^{m-1}\left(1-p\right) = \tfrac{4p(1+r)}{(1-pr)^2} + \tfrac{(1-p)}{(1-pr)},$$
$$DL_1 = ML_1^2 - (ML_1)^2 = \tfrac{-p^2+2p+4pr-1}{(1-pr)^2} + \tfrac{(1-p)}{(1-pr)}.$$

In case when 1-runs begin at chain's element with even number, variance equals

$$DL_1 = \frac{-r^2+2r+4pr-1}{(1-pr)^2} + \frac{(1-r)}{(1-pr)}.$$

2 Duration distribution of exit of air temperature for the given level

When numerical stochastic model of air temperature overshoots indicators with given one-dimensional distributions and transition probability matrix on base of heterogeneous Markov chain is being constructed, the essential thing is agreement between $f_\infty,\ g_\infty$ and $p,\ q,\ r,\ s$. Since these parameters are connected with each other in the following way

$$f_\infty = \lim_{m\to\infty} \Pr[\xi_{2m} = 1] = b,\ \ g_\infty = \lim_{m\to\infty} \Pr[\xi_{2m+1} = 1] = 1-q+bu,$$

where $b = (r-qt)/(1-d)$, $d \neq 1$, $t = r+s-1$, $d = (p+q-1)(r+s-1)$, $u = p+q-1$,

so it is necessary to choose probabilities $f_\infty,\ g_\infty,\ p,\ q,\ r,\ s$ in special way: they must be, in some sense, similar to real estimation. In this case considered Markov modelmay be used for estimations of different characterictics, which depend on $f_\infty,\ g_\infty$, for example, correlation function.

But if it is necessary to estimate distribution of runs duration, using real data and Markov model, considered in this paper, it is enough to use only estimations of transition probability matrixes, because distribution of runs duration does not depend on $f_\infty,\ g_\infty$. In this case, limiting probabilities may be used for model's verification. For verification of the model probabilities of rather short runs can be also used.

Let's give results of numerical experiments. Described above Markov model was constructed for estimations of probabilities of long-term overshoots above given level c on base of air temperature time-series in December in Astrakhan. Distributions of 1-runs duration were estimated on basis of real data and constructed model. As follows from Fig.1, model reproduces character of distribution quite well. Increasing order of Markov chains it is possible to achieve more accurate reproduction of real probabilities. As an example of such more accurate reproduction, distribution of 1-runs duration, obtained from numerical Markov model of second order with periodic transition matrix, is given in Fig.1 (curve 3).

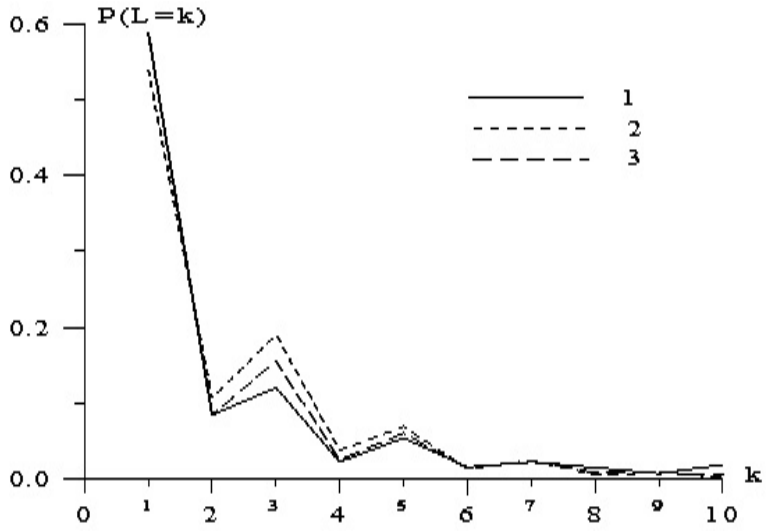


Figure 1: Distributions of 1-runs duration, obtained from: 1 - real data, 2 - Markov model of first order, 3 - Markov model of second order.

It should be noted, that considered model can be successfully used not for all levels c. For some levels, for example, levels close to average monthly temperature $c = -5^oC$, model does not give acceptable results. But for levels, relative remote from average monthly value (for example, $c = 1^oC$), model gives quite good results. Probabilities $f_\infty,\ g_\infty$, obtained with the help of model and corresponding probabilities $P(\xi = 1)$ calculated by real data are given in Table. 1. At the last column a estimations with the help modeling data are presented.

Table 1: Real probabilities $P(\xi = 1)$ and limiting probabilities obtained with the help of model

lim.prob	real data	model	model data
f_∞	0.17954	0.18208	0.18207
g_∞	0.37190	0.37821	0.37786

Conclusions

It is easy to generalize obtained formulas in case of transition probability matrix with arbitrary finite period. Such generalization is necessary for more realistic models construction of meteorological processes statistical characteristics daily periodicity. From this point of view, it is interesting to construct high-order Markov models, which take account of multivariate joint distribution. Such models can be carried out numerically with the aid of methods of statistical modelling.

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

References

[1] Dragan Ya.P., Rozhkov V.A., Yavorskiy I.N. (1987). *Methods of probabilistic analysis of oceanological processes rhythmic*. Gidrometeoizdat, Leningrad. (in Russian)

[2] Ogorodnikov V.A., Prigarin S.M. (1996). *Numerical Modelling of Random Processes and Fields: Algorithms and Applications*.-VSP, Utrecht, the Netherlands.

[3] Savelyev L.Ya., Balakin S.V. (2004). The joint distribution of the number of 1-s and the number of 1-runs in binary Markov sequences. *Discrete Mathematics and Applications*. Vol. **16**,No. **3**,pp. 43-62. (in Russian)

Effective Coefficients of Maxwell's Equations with Multiscale Isotropic Random Conductivity and Permittivity

O.N. Soboleva[1], E.P Kurochkina[2]
[1]*Novosibirsk State Technical University , Institute of Computational Mathematics and Mathematical Geophysics*
[2]*Institute of Thermophysics*
Novosibirsk, Russia
e-mail: olga@nmsf.sscc.ru

Abstract

The effective coefficients in Maxwell's equations are calculated for a multiscale isotropic medium by using a subgrid modeling approach. The correlated fields of conductivity and permittivity are mathematically represented by a Kolmogorov multiplicative continuous cascade with a lognormal probability distribution. The scale of solution domain is assumed to be large as compared with the scale of heterogeneities of the medium.

Keywords: Maxwell's equations, effective coefficients; subgrid modeling; multiscale random conductivity and permittivity.

Introduction

Most of natural media are extremely heterogeneous [1]. The physical processes in such media are described by mathematical models. The large-scale medium heterogeneities, for example, layers, intercalations, are taken into account in these models with the help of some boundary conditions. The spatial distributions of small-scale heterogeneities may not be exactly known. It is customary to assume that these parameters are random fields characterized by the joint probability distribution functions. As a rule, the small-scale fluctuations of parameters are taken into account by some effective coefficients, i.e. some simplified models with computationally resolvable scales are sought. The solution to governing equations in these models must be approximate, for example, the ensemble-averaged solutions to the initial governing equations. This is major subject of physical and engineering science that is encountered under various names, e.g. homogenization, coarse graining and subgrid modeling.

The above-mentioned methods can be applied to a "scale regular" medium. It has been experimentally shown that the irregularity of the electric conductivity, permeability, porosity and density increases as the scale of measurement decreases [1]–[4]. However, many natural media are considered to be "scale regular" in the sense that they can be described by fractals and multiplicative cascades [2]–[4]. In the present paper, the electric conductivity and permittivity are approximated by a multiplicative continuous cascade. Electromagnetic logging is an effective tool for studying a medium structure. The aim of this method is to estimate the medium conductivity

as precisely as possible. For the long probes the quasi-steady condition ($1 \ll \frac{\sigma(\mathbf{x},l)}{\omega\varepsilon(\mathbf{x},l)}$), the week influence of permittivity) is satisfied with a high accuracy, when a medium has typical values of resistivities. When resistivities of the medium are high, dependence of measured signal on dielectric permittivity occurs at high frequencies. In this case we have ($\frac{\sigma(\mathbf{x},l)}{\omega\varepsilon(\mathbf{x},l)} \ll 1$. Under this condition we obtain the effective electric conductivity and permittivity to estimate the mean-value of electric or magnetic field strengths and the current density in Maxwell's equations.

1 Governing equations and the electric conductivity and permittivity models

According to [5], Maxwell's equations for monochromatic fields $\widetilde{\mathbf{E}}(\mathbf{x},t) = Re\left(\mathbf{E}(\mathbf{x})\, e^{-i\omega t}\right)$, $\widetilde{\mathbf{H}}(\mathbf{x},t) = Re\left(\mathbf{H}(\mathbf{x})\, e^{-i\omega t}\right)$ in the absence of extraneous currents can be written as

$$\begin{aligned} rot\mathbf{H}(\mathbf{x}) &= (-i\omega\varepsilon(\mathbf{x}) + \sigma(\mathbf{x}))\,\mathbf{E}(\mathbf{x}), \\ rot\mathbf{E} &= i\omega\mu\mathbf{H}, \end{aligned} \tag{1}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the vectors of electric and magnetic field strengths, respectively; μ is the magnetic permeability; $\sigma(\mathbf{x})$ is the electric conductivity; $\varepsilon(\mathbf{x})$ is the permittivity; ω is the cyclic frequency; and $\mathbf{x}$ is the vector of spatial coordinates. The magnetic permeability is assumed to be equal to the magnetic permeability of vacuum. We also assume that the electric conductivity and the permittivity are constant outside a finite volume V with a smooth surface S. At the surface S, the tangent components of electric and magnetic field strengths are continuous. The electric conductivity, permittivity and cyclic frequency satisfy the inequality

$$\frac{\sigma(\mathbf{x},l)}{\omega\varepsilon(\mathbf{x},l)} \ll 1. \tag{2}$$

Let, for example, the field of electrical conductivity be known. This means that the field is measured on a small scale l_0 at each point $\mathbf{x}$, $\sigma(\mathbf{x})_{l_0} = \sigma(\mathbf{x})$. To pass to a coarser scale grid, it is not sufficient to smooth the field $\sigma(\mathbf{x})_{l_0}$ on a scale l, $l > l_0$. The field thus smoothed is not a physical parameter that can describe physical process, the governed by equations (1), on the scales (l, L), where L is the maximum scale of heterogeneities. This is due to the fact that the fluctuations of electric conductivity on the scale interval (l_0, l) correlate with the fluctuations of electric field strength $\mathbf{E}$ induced by the electric conductivity. To find an electric conductivity that can describe the physical process on the scales (l, L) system (1) will be used in this paper.

Following Kolmogorov [6], [7], consider a dimensionless field ψ, which is equal to the ratio of two fields obtained by smoothing the field $\sigma(\mathbf{x})_{l_0}$ on two different scales l', l. Let $\sigma(\mathbf{x})_l$ denote the parameter $\sigma(\mathbf{x})_{l_0}$ smoothed on the scale l. Then $\psi(\mathbf{x}, l, l') = \sigma(\mathbf{x})_{l'}/\sigma(\mathbf{x})_l, \quad l' < l$. We obtain expanding the field ψ into a power series in $l - l'$, and retaining the first order terms of the series, at $l' \to l$ the following equation:

$$\frac{\partial \ln \sigma(\mathbf{x})_l}{\partial \ln l} = \varphi(\mathbf{x}, l), \tag{3}$$

where $\varphi(\mathbf{x}, l') = (\partial\psi(\mathbf{x}, l', l'y)/\partial y) \mid_{y=1}$. The solution of equation (3) is

$$\sigma_{l_0}(\mathbf{x}) = \sigma_0 \exp\left(-\int_{l_0}^{L} \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right), \tag{4}$$

where σ_0 is a constant. The field φ determines the statistical properties of the electric conductivity. This approach is described in detail in [8]. According to the limit theorem for sums of independent random variables [9] if the variance of $\varphi(\mathbf{x}, l)$ is finite, the integral in (4) tends to a field with a normal distribution as the ratio L/l_0 increases. If the variance of $\varphi(\mathbf{x}, l)$ is infinite and there exists a nondegenerate limit of the integral in (4), the integral tends to a field with a stable distribution. In the present paper it is assumed that the field $\varphi(\mathbf{x}, l)$ is isotropic with a normal distribution and a statistically homogeneous correlation function

$$\begin{aligned} &< \varphi(\mathbf{x}, l)\ \varphi(\mathbf{y}, l') > - < \varphi(\mathbf{x}, l) >< \varphi(\mathbf{y}, l') > \\ &\qquad = \Phi^{\varphi\varphi}(|\mathbf{x} - \mathbf{y}|, l, l')\delta\left(\ln l - \ln l'\right). \end{aligned} \tag{5}$$

Here the angle brackets denote ensemble averaging. It follows from (5) that the fluctuations of $\varphi(\mathbf{x}, l)$ on different scales do not correlate. This assumption is standard in the scaling models [6]. This is due to the fact that the statistical dependence is small if the scales of fluctuations are different. To derive subgrid formulas to calculate effective coefficients, this assumption may be ignored. However, this assumption is important for the numerical simulation of the field σ.

For a scale-invariant medium the following relation holds for any positive K

$$\Phi^{\varphi\varphi}(|\mathbf{x} - \mathbf{y}|, l, l') = \Phi^{\varphi\varphi}(K|\mathbf{x} - \mathbf{y}|, Kl, Kl')$$

In a scale invariant medium, the correlation function does not depend on the scale at $\mathbf{x} = \mathbf{y}$ and the following estimation is obtained in [8]:

$$\begin{aligned} l_0 < l_\varepsilon < r < L \\ < \sigma_{l_0}(\mathbf{x})\sigma_{l_0}(\mathbf{x} + \mathbf{r}) > \sim C\left(\frac{r}{L}\right)^{-\Phi_0}, \end{aligned} \tag{6}$$

where $C = \sigma_0^2 e^{-\Phi_0\gamma/2}$, γ is the Euler constant. For $r \boldsymbol{g} L$, we have

$$< \sigma_{l_0}(\mathbf{x})\sigma(\mathbf{x} + \mathbf{r}, l_0) > \rightarrow \sigma_0^2. \tag{7}$$

If for any l the equality $< \sigma_l(\mathbf{x}) >= \sigma_0$ is valid, then it follows from (4), (5) that

$$\Phi_0^{\varphi\varphi}(l) = 2 < \varphi >, \tag{8}$$

where $\Phi_0^{\varphi\varphi}(l) = \Phi^{\varphi\varphi}(0, l)$. As the minimum scale l_0 tends to zero the electric conductivity field described by (4) becomes a multifractal. We obtain an irregular field on a Cantor-type set to be nonzero.

The permittivity coefficient $\varepsilon(\mathbf{x})$ is constructed by analogy to the conductivity coefficient:

$$\varepsilon_{l_0}(\mathbf{x}) = \varepsilon_0 \exp\left(-\int_{l_0}^{L} \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right). \tag{9}$$

The function $\chi(\mathbf{x}, l)$ is assumed to have the normal distribution and to be delta-correlated in the logarithm of the scale. We can write

$$\Phi^{\chi\chi}(\mathbf{x}, \mathbf{x}, l, l') = \langle\chi(\mathbf{x}, l)\chi(\mathbf{x}, l')\rangle - \langle\chi(\mathbf{x}, l)\rangle\langle\chi(\mathbf{x}, l')\rangle = \Phi_0^{\chi\chi}\delta(\ln l - \ln l'). \tag{10}$$

The permittivity field satisfies the equality $< \chi_l(\mathbf{x}) >= \chi_0$ for any l. Then it follows from (9), (10) that

$$\Phi_0^{\chi\chi}(l) = 2 < \chi > . \tag{11}$$

The correlation between the permittivity and conductivity fields is determined by the correlation of the fields $\chi(\mathbf{x}, l')$ and $(\varphi(\mathbf{x}, l')$:

$$\Phi^{\varphi\chi}(\mathbf{x}, \mathbf{y}, l, l') = \langle\varphi(\mathbf{x}, l)\chi(\mathbf{y}, l')\rangle - \langle\varphi(\mathbf{x}, l)\rangle\langle\chi(\mathbf{y}, l')\rangle = \Phi^{\varphi\chi}(|\mathbf{x} - \mathbf{y}|, l, l')\delta(\ln l - \ln l'). \tag{12}$$

2 Subgrid model

The electric conductivity and permittivity functions $\sigma(\mathbf{x}) = \sigma(\mathbf{x})_{l_0}$, $\varepsilon(\mathbf{x}) = \varepsilon(\mathbf{x})_{l_0}$ are divided into two components with respect to the scale l. The large-scale (ongrid) components $\sigma(\mathbf{x}, l)$, $\varepsilon(\mathbf{x}, l)$ are respectively obtained by statistical averaging over all $\varphi(\mathbf{x}, l_1)$ and $\chi(\mathbf{x}, l_1)$ with $, l_0 < l_1 < l$, $l - l_0 = dl$, where dl is small. The small-scale (subgrid) components are equal to $\sigma'(\mathbf{x}) = \sigma(\mathbf{x}) - \sigma(\mathbf{x}, l)$, $\varepsilon'(\mathbf{x}) = \varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}, l)$:

$$\begin{aligned}
\varepsilon(\mathbf{x}, l) &= \varepsilon_0 \exp\left[-\int_l^L \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\left\langle\exp\left[-\int_{l_0}^{l} \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\right\rangle \\
\varepsilon'(\mathbf{x}) &= \varepsilon(\mathbf{x}, l)\left[\frac{\exp\left[-\int\limits_{l_0}^{l} \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]}{\left\langle\exp\left[-\int\limits_{l_0}^{l} \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\right\rangle} - 1\right], \ \langle\varepsilon'(\mathbf{x})\rangle = 0, \\
\sigma(\mathbf{x}, l) &= \sigma_0 \exp\left[-\int_l^L \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\left\langle\exp\left[-\int_{l_0}^{l} \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\right\rangle \\
\sigma'(\mathbf{x}) &= \sigma(\mathbf{x}, l)\left[\frac{\exp\left[-\int\limits_{l_0}^{l} \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]}{\left\langle\exp\left[-\int\limits_{l_0}^{l} \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right]\right\rangle} - 1\right], \ \langle\sigma'(\mathbf{x})\rangle = 0.
\end{aligned} \tag{13}$$

Hence

$$\varepsilon(\mathbf{x},l) \simeq \left[1 - \langle\chi\rangle \frac{dl}{l} + \frac{1}{2}\Phi_0^{\chi\chi}(l) \frac{dl}{l}\right] \varepsilon_l(\mathbf{x}),$$
$$\sigma(\mathbf{x},l) \simeq \left[1 - \langle\varphi\rangle \frac{dl}{l} + \frac{1}{2}\Phi_0(l) \frac{dl}{l}\right] \sigma_l(\mathbf{x}). \tag{14}$$

The large-scale (ongrid) components of electric and magnetic field strengths $\mathbf{E}(\mathbf{x},l), \mathbf{H}(\mathbf{x},l)$ are obtained by averaging the solutions to system (1), in which the large-scale component of conductivity $\sigma(\mathbf{x},l)$ is fixed and the small component $\sigma'(\mathbf{x})$ is a random variable. The subgrid components of the electric and magnetic field strengths are equal to $\mathbf{H}'(\mathbf{x}) = \mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{x},l)$, $\mathbf{E}'(\mathbf{x}) = \mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{x},l)$. Substituting the relations for $\mathbf{E}(\mathbf{x}), \mathbf{H}(\mathbf{x})$ and $\sigma(\mathbf{x})$ into system (1) and averaging over small-scale components, we have

$$\begin{aligned} rot\mathbf{H}(\mathbf{x},l) &= (-i\omega\varepsilon(\mathbf{x},l) + \sigma(\mathbf{x},l))\mathbf{E}(\mathbf{x},l) + \langle(-i\omega\varepsilon' + \sigma')\mathbf{E}'\rangle, \\ rot\mathbf{E}(\mathbf{x},l) &= \mu i\omega\mathbf{H}(\mathbf{x},l). \end{aligned} \tag{15}$$

The subgrid term $\langle(-i\omega\varepsilon' + \sigma')\mathbf{E}'\rangle$ in system (15) is unknown. This term cannot be neglected without some preliminary estimation, since the correlation between the electric conductivity and the electric field strength may be significant. The form of this term in (15) determines the subgrid model. The subgrid term is estimated using perturbation theory. Subtracting system (15) from system (1) and taking into account only the first order terms, we obtain the subgrid equations:

$$\begin{aligned} rot\mathbf{H}' &= (-i\omega\varepsilon(\mathbf{x},l) + \sigma(\mathbf{x},l))\mathbf{E}' + (-i\omega\varepsilon'(\mathbf{x}) + \sigma'(\mathbf{x}))\mathbf{E}(\mathbf{x},l), \\ rot\mathbf{E}' &= \mu i\omega\mathbf{H}'. \end{aligned} \tag{16}$$

The variable $\mathbf{E}(\mathbf{x},l)$ on the right-hand side of (16) is assumed to be known. Solving system (16) for the components of the electric field strength we have [10]

$$\begin{aligned} E'_\alpha = \frac{1}{4\pi} i\omega\mu \int\limits_V \frac{1}{r} e^{ik} (-i\omega\varepsilon'(\mathbf{x}') + \sigma'(\mathbf{x}')) E_\alpha(\mathbf{x}',l)\, d\mathbf{x}' \\ + \frac{1}{4\pi(-i\omega\varepsilon(\mathbf{x},l)+\sigma(\mathbf{x},l))} \frac{\partial}{\partial x_\alpha} \frac{\partial}{\partial x_\beta} \int\limits_V \frac{1}{r} e^{ik} (-i\omega\varepsilon'(\mathbf{x}') + \sigma'(\mathbf{x}')) E_\beta(\mathbf{x}',l)\, d\mathbf{x}', \end{aligned} \tag{17}$$

where $r = |\mathbf{x} - \mathbf{x}'|$, $k^2 = (\mu\omega^2\varepsilon(\mathbf{x},l) + i\omega\mu\sigma(\mathbf{x},l))$. We take the square root such that $Re\, k > 0$, $Im\, k > 0$. Using (17) the subgrid term can be written as

$$\begin{aligned} &\langle(-i\omega\varepsilon'(\mathbf{x}) + \sigma'(\mathbf{x}))E'_\alpha(\mathbf{x})\rangle = \\ &= \frac{1}{4\pi} i\omega\mu \int\limits_V \frac{1}{r} e^{ik} \langle(-i\omega\varepsilon'(\mathbf{x}) + \sigma'(\mathbf{x}))(-i\omega\varepsilon'(\mathbf{x}') + \sigma'(\mathbf{x}'))\rangle\, d\mathbf{x}' E_\alpha(\mathbf{x},l) + \\ &+ \left\langle \frac{(-i\omega\varepsilon'(\mathbf{x}) + \sigma'(\mathbf{x}))}{4\pi(-i\omega\varepsilon(\mathbf{x},l) + \sigma(\mathbf{x},l))} \frac{\partial}{\partial x_\alpha} \frac{\partial}{\partial x_\beta} \int\limits_V \frac{1}{r} e^{ik} (-i\omega\varepsilon'(\mathbf{x}') + \sigma'(\mathbf{x}')) \right\rangle d\mathbf{x}' E_\beta(\mathbf{x},l). \end{aligned} \tag{18}$$

Since a small change in the scale of σ produces considerable fluctuations in the field (which is typical of fractal fields), the field $\sigma(\mathbf{x},l)$ and its derivatives are believed to change slower than σ' and its derivatives. Similar assumptions are made for $\mathbf{E}(\mathbf{x},l)$ and $\mathbf{H}(\mathbf{x},l)$. Therefore $\mathbf{E}(\mathbf{x},l)$, $\sigma(\mathbf{x},l)$ and their derivatives can be factored outside the integral sign in (18). Integrating by parts (18) we have

$$\begin{aligned}
&\langle(-i\omega\varepsilon'(\mathbf{x})+\sigma'(\mathbf{x}))E'_\alpha(\mathbf{x})\rangle = \\
&=\frac{1}{4\pi}i\omega\mu\int\limits_V\frac{1}{r}e^{ikr}\left(-\omega^2\langle\varepsilon'(\mathbf{x})\varepsilon'(\mathbf{x}')\rangle-2i\omega\langle\varepsilon'(\mathbf{x}')\sigma'(\mathbf{x})\rangle+\langle\sigma'(\mathbf{x})\sigma'(\mathbf{x}')\rangle\right)d\mathbf{x}'E_\alpha(\mathbf{x},l) \\
&+\frac{1}{4\pi(-i\omega\varepsilon(\mathbf{x},l)+\sigma(\mathbf{x},l))}\times \qquad (19)\\
&\times\int\limits_V\frac{\partial}{\partial x'_\alpha}\frac{\partial}{\partial x'_\beta}\frac{1}{r}e^{ikr}\left(-\omega^2\langle\varepsilon'(\mathbf{x})\varepsilon'(\mathbf{x}')\rangle-2i\omega\langle\varepsilon'(\mathbf{x}')\sigma'(\mathbf{x})\rangle+\langle\sigma'(\mathbf{x})\sigma'(\mathbf{x}')\rangle\right)d\mathbf{x}'E_\beta(\mathbf{x},l).
\end{aligned}$$

Here the summation of repeated indices is implied. As follows from formulas (2),(13) for a log-normal probability distribution of σ and ε at small dl we have $\langle\varepsilon'(\mathbf{x})\varepsilon'(\mathbf{x})\rangle \approx \varepsilon^2(\mathbf{x},l)\Phi_0^{\chi\chi}\frac{dl_1}{l_1}$, $\langle\sigma'(\mathbf{x})\sigma'(\mathbf{x}')\rangle = \Phi^{\varphi\varphi}(r,l)\sigma^2(\mathbf{x},l)\frac{dl}{l}$, $\langle\sigma'(\mathbf{x})\varepsilon'(\mathbf{x})\rangle \approx \varepsilon(\mathbf{x},l)\sigma(\mathbf{x},l)\Phi_0^{\chi\sigma}\frac{dl_1}{l_1}$, $\frac{1}{4\pi(-i\omega\varepsilon(\mathbf{x},l)+\sigma(\mathbf{x},l))} \approx -\frac{1}{4\pi i\omega\varepsilon(\mathbf{x},l)}\left(1-\frac{i\sigma(\mathbf{x},l)}{\omega\varepsilon(\mathbf{x},l)}\right)$. Substituting these formulas into (22) yields

$$\begin{aligned}
&\langle(-i\omega\varepsilon'(\mathbf{x})+\sigma'(\mathbf{x}))E'_\alpha(\mathbf{x})\rangle \approx \\
&\approx -\frac{1}{4\pi}i\omega^3\mu\varepsilon(\mathbf{x},l)\int\limits_V\frac{1}{r}e^{ikr}\Phi_0^{\chi\chi}(r)\,d\mathbf{x}'\frac{dl}{l}\varepsilon(\mathbf{x},l)E_i(\mathbf{x},l) \\
&+\frac{1}{2\pi}\omega^2\varepsilon(\mathbf{x},l)\mu\int\limits_V\frac{1}{r}e^{ikr}\Phi_0^{\chi\varphi}(r)\,d\mathbf{x}'\frac{dl}{l}\sigma(\mathbf{x},l)E_i(\mathbf{x},l) \\
&+\frac{1}{4\pi}i\omega\mu\sigma(\mathbf{x},l)\int\limits_V\frac{1}{r}e^{ikr}\Phi_0^{\varphi\varphi}(r)\,d\mathbf{x}'\frac{dl}{l}\sigma(\mathbf{x},l)E_i(\mathbf{x},l) \\
&-\frac{i\omega\varepsilon(\mathbf{x},l)}{4\pi}\int\limits_V\frac{\partial}{\partial x'_\alpha}\frac{\partial}{\partial x'_\beta}\frac{1}{r}e^{ikr}\left(1-\frac{i\sigma(\mathbf{x},l)}{\omega\varepsilon(\mathbf{x},l)}\right)\times \\
&\times\left(\Phi_0^{\chi\chi}(r)+2i\frac{\sigma(\mathbf{x},l)}{\omega\varepsilon(\mathbf{x},l)}\Phi_0^{\chi\varphi}(r)-\frac{\sigma^2(\mathbf{x},l)\Phi_0^{\varphi\varphi}(r)}{\omega^2\varepsilon^2(\mathbf{x},l)}\right)d\mathbf{x}'\frac{dl}{l}E_\beta(\mathbf{x},l). \qquad (20)
\end{aligned}$$

The integrals over V in (20) can be changed by an integral with infinite limits, since the correlation functions $\Phi^{\chi\chi}$, $\Phi^{\chi\varphi}$, $\Phi^{\varphi\varphi}$ are small if $|\mathbf{x}|>L$, $L\ll L_0$, where L_0 is minimum size of V. This change gives a sensible error only in a narrow region of the correlation radius size near the boundary. In formula (20), the Cartesian coordinates are changed for spherical coordinates. Integrating n_jn_m, where $n_m=x_m/r$, over the complete solid angle we arrive at the formula $\int n_jn_m d\vartheta=\frac{4\pi}{3}\delta_{jm}$. Using this formula, neglecting terms of second order of smallness of $\sigma(\mathbf{x},l)/\omega\varepsilon(\mathbf{x},l)$ and integrating

(20) by parts, we obtain

$$\begin{aligned}
&\langle(-i\omega\varepsilon'(\mathbf{x})+\sigma'(\mathbf{x}))\,E'_\alpha(\mathbf{x})\rangle\approx\\
&\approx-\frac{1}{3}\left(2\mu\omega^2\varepsilon(\mathbf{x},l)-i\omega\mu\sigma(\mathbf{x},l)\right)\int\limits_0^\infty re^{ikr}\Phi^{\chi\chi}(r)\,dr\frac{dl}{l}i\omega\varepsilon(\mathbf{x},l)E_\alpha(\mathbf{x},l)\\
&+\frac{2}{3}\left(2\mu\omega^2\varepsilon(\mathbf{x},l)-i\omega\mu\sigma(\mathbf{x},l)\right)\int\limits_0^\infty re^{ikr}\Phi^{\chi\sigma}(r)\,dr\frac{dl}{l}\sigma(\mathbf{x},l)E_\alpha(\mathbf{x},l)\\
&+i\omega\mu\sigma(\mathbf{x},l)\int\limits_0^\infty re^{ikr}\Phi^{\sigma\sigma}(r)\,dr\frac{dl}{l}\sigma(\mathbf{x},l)\,E_\alpha(\mathbf{x},l)+\\
&+\frac{1}{3}\Phi_0^{\chi\chi}\frac{dl}{l}i\omega\varepsilon(\mathbf{x},l)E_i(\mathbf{x},l)+\left(\frac{1}{3}\Phi_0^{\chi\chi}-\frac{2}{3}\Phi_{(0)}^{\chi\sigma}\right)\frac{dl}{l}\sigma(\mathbf{x},l)E_i(\mathbf{x},l).
\end{aligned}\tag{21}$$

If $\omega\mu L^2\left|(i\omega\varepsilon(\mathbf{x},l)+\sigma(\mathbf{x},l))\right|\ll 1$, the integrals in (21) are small [11]. This inequality is not restrictive for the problems of electromagnetic logging if $L\ll L_0$. Hence, the integrals in (21) can be neglected. We have:

$$\begin{aligned}
\langle-i\omega\varepsilon'(\mathbf{x})\,E'_i(\mathbf{x})\rangle+\langle\sigma'(\mathbf{x})\,E'_i(\mathbf{x})\rangle\approx-\tfrac{1}{3}\Phi^{\chi\chi}(0)\left(-i\omega\varepsilon(\mathbf{x},l)E_i(\mathbf{x},l)\right)\tfrac{dl}{l}-\\
-\left(\tfrac{2}{3}\Phi^{\chi\sigma}(0)-\tfrac{1}{3}\Phi^{\chi\chi}(0)\right)\tfrac{dl}{l}\sigma(\mathbf{x},l)E_i(\mathbf{x},l).
\end{aligned}\tag{22}$$

Substituting (22) into (15), we have

$$\begin{aligned}
rot\mathbf{H}(\mathbf{x},l)=-i\omega\varepsilon_{l0}\exp\left[-\textstyle\int_l^L\chi(\mathbf{x},l_1)\frac{dl_1}{l_1}\right]\mathbf{E}(\mathbf{x},l)+\sigma_{l0}\exp\left[-\textstyle\int_l^L\varphi(\mathbf{x},l_1)\frac{dl_1}{l_1}\right]\mathbf{E}(\mathbf{x},l),\\
rot\mathbf{E}(\mathbf{x},l)=i\omega\mu\mathbf{H}(\mathbf{x},l).\\
\varepsilon_{l0}=\left(1-\tfrac{\Phi_0^{\chi\chi}}{3}\tfrac{dl}{l}\right)\left[1+\left(\tfrac{\Phi_0^{\chi\chi}}{2}-\langle\chi\rangle\right)\tfrac{dl}{l}\right]\varepsilon_0.\\
\sigma_{l0}=\left(1-\left(\tfrac{2}{3}\Phi^{\chi\varphi}(0)-\tfrac{1}{3}\Phi^{\chi\chi}(0)\right)\tfrac{dl}{l}\right)\left[1+\left(\tfrac{\Phi_0^{\varphi\varphi}}{2}-\langle\varphi\rangle\right)\tfrac{dl}{l}\right]\sigma_0.
\end{aligned}\tag{23}$$

It follows from (23) that the new coefficients σ_{l0} and ε_{l0} are equal:

$$\begin{aligned}
\varepsilon_{l0}=\varepsilon_0+\left(\frac{\Phi_0^{\chi\chi}}{6}-\langle\chi\rangle\right)\varepsilon_0\frac{dl}{l}\\
\sigma_{l0}=\sigma_0+\left(-\frac{2}{3}\Phi_0^{\chi\varphi}+\frac{1}{3}\Phi_0^{\chi\chi}+\frac{1}{2}\Phi_0^{\varphi\varphi}-\langle\varphi\rangle\right)\sigma_0\frac{dl}{l},
\end{aligned}$$

with second order of accuracy. As $dl\to 0$ we obtain the equation

$$\begin{aligned}
\frac{d\ln\varepsilon_{0l}}{d\ln l}=\frac{1}{6}\Phi_0^{\chi\chi}(l)-\langle\chi\rangle,\\
\frac{d\ln\sigma_{0l}}{d\ln l}=-\frac{2}{3}\Phi_0^{\chi\varphi}+\frac{1}{3}\Phi_0^{\chi\chi}+\frac{1}{2}\Phi_0^{\varphi\varphi}-\langle\varphi\rangle.
\end{aligned}\tag{24}$$

For a scale-invariant medium, effective equations have the following simple form

$$\begin{aligned} rot\mathbf{H}(\mathbf{x},l) &= -i\omega\left(\frac{l}{L}\right)^{\langle\chi\rangle-\Phi_0^{\chi\chi}/6}\varepsilon_l(\mathbf{x})\,\mathbf{E}(\mathbf{x},l)+\left(\frac{l}{L}\right)^{\langle\varphi\rangle+\frac{2}{3}\Phi_0^{\chi\varphi}-\frac{1}{3}\Phi_0^{\chi\chi}-\frac{1}{2}\Phi_0^{\varphi\varphi}}\sigma_l(\mathbf{x})\,\mathbf{E}(\mathbf{x},l), \\ rot\mathbf{E}(\mathbf{x},l) &= i\omega\mu\mathbf{H}(\mathbf{x},l). \end{aligned} \tag{25}$$

Conclusions

Random media are considered in which the permittivity and conductivity are random functions of position. For them we have obtained effective coefficients depending on the scale of smoothing l when the condition $\frac{\sigma(\mathbf{x},l)}{\omega\varepsilon(\mathbf{x},l)} \ll 1$ is satisfied. To verify our theory Monte Carlo simulation is required. Such computations are being performed.

Acknowledgements

This work has been supported by the RFBR (Grant № 110100641)

References

[1] Sahimi M. (1993). Flow phenomena in rocks: from continuum models, to fractals, percolation, cellular automata, and simulated annealing. *Reviews of Modern Physics*. Vol. **65**, pp. 1393–1534.

[2] Biswal B, Manwart C. and Hilfer R. (1998) Three-dimensional local porosity analysis of porous media. *Physica A*. Vol. **255**, pp. 221–241.

[3] Krylov S.S. , Lyubchich V.A., The apparent resistivity scaling and fractal structure of an iron formation.(2002) *Izvestiya. Physics of the Solid Earth. 38*, pp. 1006–1012.

[4] Bekele A., Hudnall H. W., Daigle J.J., Prudente A. and Wolcott M.(2005) Scale dependent variability of soil electrical conductivity by indirect measures of soil properties, *Journal of Terramechanics*, Vol. **42**, pp. 339–351.

[5] Landau L. D., Lifshitz E. M. (1984) *Electrodynamics of continuous media*. Oxford-Elmsford, New York: Pergamon Press.

[6] Kolmogorov A. N. (1962) A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number. *J. Fluid Mech.*, Vol. **13**, pp. 82–85.

[7] Monin A. Yaglom A. (1975). *Statistical fluid mechanics*, II. Cambridge: MIT Press.

[8] Kuz'min G. A., Soboleva O. N. Subgrid modeling of filtration in porous self-similar media. (2002). *J. Appl. Mech. Tech. Phys.*, Vol. **43**, 583–592.

[9] Gnedenko B. V., Kolmogorov A. N. (1954). *Limit, distributions for sums of independent random variables*. Trans. and annotated by K. L. Chung. Cambridge: Addison-Wesley.

[10] Koshljakov N. S., Smirnov M. M., Gliner E. B. (1962) *Differential equations of mathematical physics*. Fizmatgiz, Moscow,(in Russian).

[11] Rytov S. M., Kravtsov Yu. A., Tatarskii V. I. (1987 *Principles of statistical radiophysics*, III, IV.Springer-Verlag, Berlin.).

Numerical Analysis of SDE on Supercomputers

S.S. Artemiev, V.D. Korneev
Institute of Computational Mathematics and Mathematical Geophysics,
Siberian Branch, Russian Academy of Sciences
Novosibirsk, Russia
e-mail: ssa@osmf.sscc.ru, korneev@ssd.sscc.ru.ru

Abstract

This paper deals with some problems of accuracy of algorithms for the numerical solutions of stochastic differential equations (SDEs) versus the size of the ensemble of trajectories simulated and on the mesh size of integrating the generalized Euler method. The problems of accuracy arise in estimating functionals of SDE solutions with increasing variance, highly asymmetric distributions, and an indefinite time of arrival of trajectories of solutions at the boundaries of given domains. Some ways of parallelization of statistical algorithms on a multiprocessor cluster are described. Results of numerical experiments performed on a supercomputer available at the Siberian Supercomputer Center are presented.

Keywords: stochastic differential equations, statistical algorithms, parallelization, supercomputer, cluster, van der Pol equation, phase trajectory, stochastic oscillators.

Introduction

To use Monte Carlo methods for finding solutions to boundary-value problems of mathematical physics in probabilistic terms, it is often necessary to solve numerically the resulting SDEs and calculate some integrals along the trajectories simulated.

This paper studies the accuracy of statistical algorithms for solving simple SDEs versus the size of the ensemble of trajectories simulated and on the mesh size of integrating the generalized Euler method. The results of numerical experiments on estimating some moments of SDE solutions with increasing variance and the mean time of first arrival of SDE trajectories with multiplicative noise at a boundary of a given domain are presented.

But the accuracy of estimates of functionals of SDE solutions depend not only on the size of the ensembles of simulated trajectories of solutions, but also on the size of the integration step used a numerical method for solving CDSs. Of particular difficulty is the statistical modeling of stochastic oscillators. Mathematical models in the form of CDS with oscillating solutions occur in many different fields of science [1-3]. Of particular interest is the analysis of the possible transitions from one type of oscillation to another, for example, in predicting failures and accidents caused by increasing amplitude of oscillations. In this connection there is the problem of estimating the stability of a given operation. As shown previously conducted experiments [4], the numerical solution of oscillatory SDEs with integration step $h < 10^4$ is often an instability of the numerical solution, ie, the strong growth of the oscillation amplitude and dispersion. In this regard, it is necessary to reduce the size of the integration step by several orders of magnitude. In addition,

small amounts of simulated trajectories of solutions of SDE give completely incorrect assessment of the moments of decision in the case of strong asymmetry in their distribution densities.

Ratings st functionals of solutions of SDE, the calculation of the integrals along the simulated trajectories, and the midget step size generalized Euler's method, and vast amounts of simulated trajectories of solutions of SDE, in both cases require the use of supercomputers with large number of processors to obtain satisfactory accuracy of the numerical analysis in a reasonable computation time. The first study authors for the numerical solution of SDE with increasing variance on supercomputers have been described in [5].

A description of parallel programs and estimates of the calculation time versus the number of processors used and the size of the ensemble of SDE trajectories simulated are presented. The numerical experiments were performed on a cluster, NKS-30T, available at the Siberian Supercomputer Center at the Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of the Russian Academy of Sciences.

1 Problems to be Solved

For solving the Cauchy problem for general SDEs in the sense of Ito,

$$dy = f(y)\,dt + \sigma(y)\,dw(t), \qquad y(0) = y_0, \quad 0 \le t \le T \tag{1}$$

the generalized Euler method

$$y_{n+1} = y_n + h\,f(y_n) + \sigma(y_n)\sqrt{h}\;\xi_{n+1}. \tag{2}$$

is commonly used. Here, y_{n+1} is a numerical solution on a grid $t_{n+1} = t_n + h$, and $\{\xi_{n+1}\}_0^{N-1}$ is a sequence of standard Gaussian random quantities that are mutually independent and also independent of y_n. To generate ξ_n, we use the standard formula $\xi = \sqrt{-2\ln\alpha_1}\,\sin 2\pi\alpha_2$, where α_1 and α_2 are uniformly distributed random quantities in (0,1] [6].

1) The Solution of the SDEs with Multiplicative Noise

a) The exact solution to the Cauchy problem for the simple linear SDE in the Ito sense with multiplicative noise

$$\begin{aligned} &dy = \alpha y\,dt + \sigma y\,dw(t), \qquad 0 \le t \le T, \\ &y(0) = y_0 \end{aligned} \tag{3}$$

SDEs (3) has an exact solution. The size of the ensemble of simulated trajectories required to obtain acceptable accuracy (2-3 mark) of the solution estimated.

b) It is often recommended to use Monte Carlo methods for finding solutions to boundary value problems for elliptic equations with the help of a probabilistic representation [7]. For instance, for the 1D Ditchless problem

$$\frac{1}{2}\,\sigma^2(y)\frac{d^2u}{dy^2} + f(y)\frac{du}{dy} + 1 = 0 \tag{4}$$

in an interval [a, b] with boundary conditions u(a) = u(b) = 0 the solution u(y0) at a point y0 can be represented in the probabilistic form

$$u(y_0) = E\tau(y_0), \tag{5}$$

where τ is the time of first exit from $[a, b]$ for the solution of SDE (1).

2) Stochastic Oscillators

a) Multiplicative noise is often associated with "noisy" the coefficients in SDEs. Consider the case of "noisy" the coefficients in the linear oscillatory circuit, which is given SDEs second-order form

$$\frac{d^2y}{dt^2} + (\lambda + \sigma_1 \frac{dw_1}{dt}) \frac{dy}{dt} + (\omega^2 + \sigma_2 \frac{dw_2}{dt}) y = 0 \tag{6}$$

with constant λ, ω, σ_1, σ_2.
We consider the special case of the SDE (3) of the form

$$\frac{d^2y}{dt^2} + (\omega^2 + \frac{dw}{dt}) y = 0, \quad y(0) \in N(1, 1), \quad \frac{dy}{dt}(0) = 0, \tag{7}$$

where the "noise" only the oscillation frequency solutions and there is no decrement.

b) A stochastic nonlinear Van der Pol oscillator with a "noisy" coefficient, written as a system of SDE in the sense of Ito type

$$\begin{aligned} dy_1 &= y_2\, dt, \\ dy_2 &= (a\, y_2\, (1 - b\, y_1^2) - \omega^2 y_1)\, dt + \sigma y_1\, dw(t), \end{aligned} \tag{8}$$

describes the oscillations of the nonlinear circuit. In (8) constants a, b, ω determine the rate of transition in the decision. For the expectation exact solution of SDEs (8) there is no explicit formula-representation, nor a closed system of ODE for its numerical calculation. The only constructive way to analyze the nonlinear SDE with a large noise is the Monte Carlo.

2 Description of Parallel Programs

Two ways of parallelization of algorithms on a multiprocessor cluster are described. The paralleling is performed in the MPI parallel programming system [8]. First, a method of parallelization is considered where the ME of the SDE solution is estimated over the entire integration interval and the simulation time of all ensemble trajectories is the same.

Since independent implementations of the SDE solution are simulated by Monte Carlo methods, an efficient organization of parallel implementation on a multiprocessor cluster is possible. In this case, a simple parallelization scheme is: different processors of the computational system provide fully independent solutions, that is, they calculate sequences obtained on the basis of different (in each processor) pseudorandom numbers.

Let K be the number of processors in a computational system that implements the algorithm and let M be the ensemble size of trajectories simulated. Let $M_k = M/K$ be the ensemble size on one processor. Then formula for estimating of the ME of the solution at a grid node t_n in the parallel implementation has the following form:

$$m_n = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M_k}\sum_{m=1}^{M_k} y_n^{(m,k)}. \tag{9}$$

Here, $y_n^{(m,k)}$ is the value of the mth implementation of the SDE solution at the nth grid node obtained on the kth processor.

In the parallel implementation of this algorithm, these times are minimal: here, the time spent on final averaging of independent results plays a minor role [9, 10].

In another parallelization method, the trajectories are simulated up to their first arrival at the boundary of a given domain, and the simulation times of different trajectories from the ensemble are different. The first arrival of the sequence $\{y_n\}$ that is simulated at the boundary of a given domain and the number of iteration steps n_m made prior to this arrival are fixed. Here, an iteration step means a calculation of the next value, y_{n+1}. Then the simulation process starts again from the point y_0.

As in the previous algorithm, here the realizations of SDE solutions are independent. Therefore, an efficient organization of the parallel implementation is possible. Let $M_k^{min} = M^{min}/K$. Then formula for estimating the mean time of first arrival of trajectories starting from y_0 in the parallel implementation can be written as

$$\hat{\tau}(y_0) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M_k^{min}}\sum_{m=1}^{M_k^{min}} n_m^{(k)}h. \tag{10}$$

Here, $n_m^{(k)}$ is the number of iteration steps prior to the arrival of the mth realization made on the kth processor. It should be noted that when the simulation terminates, the number of arrivals at the domain boundary on different processors can differ greatly. It is important that the total number of arrivals must be no less than M^{min}. It should be noted that, in contrast to the first algorithm, here the spacing h must be sufficiently small not to decrease the estimation accuracy when fixing the time of trajectory passage outside the domain.

In the second algorithm, the paralleling scheme differs from that in the first algorithm, because here after a given number of iteration steps the quantities M_k^{min} must be constantly scanned on all processors, their sum must be calculated, and it must be checked whether the common minimum M^{min} is reached. The number of iteration steps on each processor taken by a realization to pass outside the domain boundaries greatly depends on the parameters in (3), and the approximate mean number of these steps is unknown. In this case it is best to perform this testing (summation) synchronously, exchanging information between the processors after a given number of iteration steps, *Is* made on each processor.

The frequency of interactions between the processors, that is, the period *Is*, greatly affects the time of execution of the parallel algorithm. Summation of the current number of passages

on the processors and summation of the total number of steps on each processor in each interval Is is made by using a reduced MPI-function [8]. This operation takes much more time than the operations on one processor. Therefore, in this paper the following strategy of dynamics choosing the size of the interval Is is used:

1. At the initial calculation time, set a size of the interval, $Is = N_{Is}$.

2. If $LM_k^{min} = 0$, all processors double the size of the interval Is. Thus, next time all processors will interact after an interval that is twice as large as the previous one.

3. If $GSM_k^{min} \neq 0$ and $GSn_k \neq 0$, each processor assigns the quantity $C \cdot (GSn_k)/GSM_k^{min}$ to its variable Is.

Here, GSM_k^{min} is the current total (global) number of realization passages on all processors, and GSn_k is the current total (global) number of iteration steps on all processors for these passages. It should be specially noted that only those iteration steps that are associated with passages outside the domain boundaries are summed up. Notice also that, due to the reduced (MPI)-summation operation, the values of GSM_k^{min} and GSn_k counters will be on each processor.

The wide range of problems that can be solved by Monte Carlo methods, are ideal for vector calculations. The speed of the programs, due to the vectorization of calculations, increases by $40\% - 50\%$ (on each processor) in comparison to the same programs without vectorization of calculations.

In the problems being solved, the samples of basic random numbers are very large ($\approx 2 \cdot 10 \cdot 10^{13}$) and, therefore, it is reasonable to use "long-period" pseudorandom sequences. In these problems, a function from the MKL library is used [11]. This function implements: 1) type of statistical distributions-UNIFORM; 2) method-STD (standard); 3) generator of random numbers-MT2203. This is, in fact, a set of 1024 pseudorandom number generators designed for using in parallel Monte Carlo methods simulations. Each of them generates a sequence with a periodicity equal to 2 to the power 2203. The parameters of the generators provide the corresponding sequences of pseudorandom numbers that are mutually independent. One of these parameters is the processor identification logical number assigned by the MPI system of parallel programming [8].

3 Numerical Experiments

The calculations were performed on a cluster, NKS-30T, available at the Siberian Supercomputer Center at the Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of the Russian Academy of Sciences.

The following numerical experiments were performed to study the accuracy of Monte Carlo methods estimation versus the size of the ensemble of the SDE trajectories simulated.

Here, a processor means a processor kernel.

Test 1.

The solution of the SDE in the Ito sense

$$\begin{aligned} &dy = y\,dw(t), \quad 0 \le t \le 10, \\ &y(0) = 1 \end{aligned} \tag{11}$$

is simulated by the following exact recurrent formula: $y_n = y_{n-1}\exp\left(-\frac{1}{2} + \xi_n\right), \quad y_0 = 1,\ n = 1, 2, ..., 10$. The ME of the SDE solution is estimated by formula (9), and the second moment of the solution, by the formula

$$\alpha_{2n} = \frac{1}{M}\sum_{m=1}^{M} y_n^{(m)^2}, \quad n = 1, ..., 10.$$

For SDE (11), we have $Ey(t) \equiv 1,\ Ey^2(t) = e^t$. The variance $Dy(T) = \exp(T) - 1$ can be considerable. This means that the accuracy of estimating $Ey(T)$ and $Ey^2(T)$ by Monte Carlo methods with an ensemble of simulated SDE trajectories of size $10^2 \div 10^4$ would be very low, that means a possible loss of accuracy of $10\% \div 1000\%$ in comparison to its exact value [12, 13]. Calculations were performed for various sizes of the ensemble of trajectories simulated: $M = 10^3, 10^6, 10^9, 10^{12}$, and 10^{13}. The number of processors $K = 20$ for the first three sizes, and for the last two sizes $K = 50$. The accuracy of estimates m_n and α_{2n} at the end of the integration interval is high only with $M = 10^9 - 10^{13}$. The required accuracy of the estimate α_{2n} $(= 22943)$ is high only for $M = 10^{13}$. Recall that $Ey^2(10) = e^{10} = 22026$. The accuracy of estimates α_{2n} small sample sizes is low due to the fact that these samples do not have occasional greatly increasing trajectories of the SDE solution. Calculation time versus the number of processors is close to linear.

Test 2.

For solving SDE (11) simulated by the formula $y_n = y_{n-1}\exp\left(-\frac{h}{2} + \sqrt{h}\,\xi_n\right), \quad y_0 = 1$, we estimate the first and second moments of a random quantity τ, which is the time of first arrival of realizations at the boundary of the interval [0, 2]. Calculations were performed by formula (10) for $K = 64, h = 10^{-4}$ for the following ensemble sizes of the trajectories simulated: $M = 10^2, 10^4$, and 10^7. In this test, the exact values of $E\tau$ and $E\tau^2$ are unknown. Nevertheless, one can see a great difference in the estimates for $M = 10^2$ and $M = 10^4$, 10^7. This means that ensembles of maximal sizes should be used in problems on boundary reaching. The number of coinciding significant figures at various values of M can serve as a criterion of accuracy. The estimate of $\hat{\tau}^2$ for $M = 10^2$ is small due to the fact that the sample does not have occasional realizations that arrive at the interval boundary for a long time.

Test 3.

If we set $f(y) = 0$ and $\sigma(y) = \sigma$ in (4), the accompanying SDE is a SDE with additive noise,

$$dy = \sigma\,dw(t),\ y(0) = y_0 \tag{12}$$

For solve SDE (12) simulated by the formula $y_n = y_{n-1} + \sigma\xi_n\sqrt{h}$, we estimate the mean time and the second moment of τ, which is the time of first arrival of realizations at the boundary of the interval $[-1, 1]$. Calculations were performed for $K = 64$ and $h = 10^{-4}$ for ensemble

sizes $M = 10^2, 10^3, 10^6$, and 10^8. In the first case, $y_0 = 0$, the SDE trajectories start from the center of the interval $[-1, 1]$. In the second case, $y_0 = 0.9$, the trajectories start from a point near the right boundary of the interval, and the accuracy of estimation is strongly affected by occasional realizations that arrive at the left boundary of the interval. The results of calculations shows a great difference in the accuracy of estimation for various M, both for the case when y_0 is in the center of the interval $[-1, 1]$ and near the boundary for various σ. In all the tests, the second moment estimates for all $M = 10^2, 10^3$ have very low accuracy. The ensemble simulated for $K = 64$ has a large size, which requires too much calculation time, more than two days.

Test 4.

Evaluation of the expectation and second moment of solutions $y(t)$ linear SDEs (7). Expectation under the initial conditions $m(0) = 1$, $\frac{dm}{dt}(0) = 0$ given by $m(t) = cos(wt)$. If $\omega = 2\pi$ the function $m(t)$ on the interval $[0, 100]$ has 100 periods of oscillation with equal amplitude of 1.

CDE (6) can be rewritten as a linear system with $L = 2, \quad J = 2$:

$$\begin{aligned} dy_1 &= y_2\, dt, \\ dy_2 &= -(\omega^2 y_1 + \lambda\, y_2)\, dt - \sigma_1\, y_2\, dw_1(t) - \sigma_2\, y_1\, dw_2(t). \end{aligned} \tag{13}$$

Using the Euler method (2) to CDEs (7), written in the form of first-order system (13) with L = 2, J = 1, we obtain the following difference scheme:

$$\begin{aligned} y_{n+1}^{(1)} &= y_n^{(1)} + h y_n^{(2)}, \\ y_{n+1}^{(2)} &= y_n^{(2)} - h\omega^2 y_n^{(1)} + \sqrt{h}\, y_n^{(1)}\, \xi_{n+1}\,. \end{aligned} \tag{14}$$

Calculations of the phase trajectory of ($Ey_n^{(1)}$, $Ey_n^{(2)}$), The calculated using the difference scheme (14) with step $h = 10^{-3}$ and the size of ensemble of simulated trajectories of solutions $M = 8^7$. Calculations show that with this step size of integration has been a steady increase in amplitude estimation of the expectation. Stable phase trajectory in the form of an ellipse ($Cos\, 2\pi t$, $-2\pi\, sin\, \pi t$) can be obtained only when the integration step $h = 10^{-6}$ and less.

In evaluating the second moment of the generalized Euler method with step $h = 10^{-3}$, we have a maximum value of 205, indicating that the complete loss of estimation accuracy with the step. Satisfactory accuracy is obtained only if h is not more than 10^{-6}, when we estimate a maximum value equal to 4.00 close to the exact solution [9]. The number of simulated trajectories is chosen so large that in this calculation the size of the ensemble of simulated trajectories had no effect on the accuracy of the estimates.

Growth dispersion solutions CDSs (6) over time makes the problem an accurate assessment of the expectation of $Ey(t) = cos\, 2\pi t$ for a large number of oscillation periods and requires an increase in the size of the ensemble of simulated trajectories with increasing T, that naturally increases the computation time problem.

The computing time on 64 processors with step $h = 10^{-6}$, $M = 8^6$ was about 16 hours. Calculations with a smaller integration step requires a lot of daily calculating.

Test 5.

Evaluation of the expectation and second moment of the solution of the nonlinear Van der Pol (8) with parameters $a = 20$, $b = 1$, $\omega = 2\pi$, $\sigma = 1$.

Simulated the trajectory of a solution $y_1(t)$ of CDEs (7) with step $h = 10^{-8}$ If you choose the parameters $a = 20$ a stochastic equation of Van der Pol oscillator can be considered "tough" [9], we have fast transient plots and "shelves" near the values ± 3. Calculations show that the amplitude of trajectory solutions is not reduced. However, numerical calculations show that this can not be said about the behavior of the expectation of $Ey_1(t)$. Estimates were made of the expectation of $Ey_1(t)$, obtained by the generalized Euler method with step $h = 10^{-8}$ and the size of the ensemble $M = 8^5$. Calculations show that the amplitude of $Ey_1(t)$ decreases from period to period, which is in sharp contrast with the behavior of the expectation of a linear oscillator circuit (Test 4). We can say that the expectation of a nonlinear SDEs with time "loses information" about the behavior of each individual trajectory of the solution (8). Additional calculations showed that in the case of weak nonlinearity (with $a = 1$) the expectation of the solution (8) is also decreasing over time, the amplitude of oscillation. Such a dramatic difference in the behavior of the expectation of solutions of linear and nonlinear oscillators warns about the dangers of solutions of nonlinear SDEs with their linearization.

The estimation of second moment of solutions $Ey_1^2(t)$. Fast enough is stabilized dispersion solutions, which also differs from the behavior of the variance in the linear case. The computing time of this test on 128 processors with step $h = 10^{-8}$ and $M = 8^5$ accounted for about 73 hours. Note that the use of larger integration step leads to the instability of numerical solutions, which ultimately leads to the overflow bit arithmetic unit of the processor grid and the emergence of values Nan (nonexistent number).

Conclusions

The numerical calculations clearly show that supercomputers are necessary for solving SDEs numerically. In a subsequent paper, we plan to consider problems of numerical simulation of linear and nonlinear stochastic oscillators with the use of about 400 processors. Numerical experiments will be performed with the solutions to more complicated boundary value problems (where integrals are calculated along trajectories of SDE solutions).

Studies have shown that the numerical analysis stochastic oscillators Monte Carlo use large integration steps making the CDS, which is almost always done in the numerical experiments on personal computers, can lead to completely erroneous conclusions.

Also note that the conclusions drawn from the behavior of moments for solving the stochastic equation Van der Pol equation with large noise can not be obtained nor any other method, except for stochastic simulation. This applies to both methods approximation of Gauss and the spectral method [14].

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (project № 08-01-0034).

References

[1] Dimentberg, M.F. Nelineinye stokhasticheskie zadachi mekhanicheskikh kolebanii. -M.: Nauka, 1980.

[2] Babickii, V.I. Teoriya vibroudarnykh sistem. -M.: Nauka, 1978.

[3] Artemiev, S.S., Yakunin, M.A. Matematicheskoe i stokhasticheskoe modelirovanie v finansakh. Novosibirsk: Institute of Computational Mathematics and Mathematical Geophysics SB RAS, 2008.

[4] Artemiev, S.S., Chislennoe reshchenie zadachi Koshi dlya sistemy obyknovennykh i stokhasticheskikh differentsialnykh uravnenii (Numerical Solution of Cauchy Problems for Systems of Ordinary and Stochas tic Differential Equations), Novosibirsk: Institute of Computational Mathematics and Mathematical Geophysics SB RAS, 1993.

[5] Artemiev, S.S., Korneyev, V.D. Chislennoe reshenie stokhasticheskikh differencialnykh uravnenii na supercomputerakh //Sib. Zh. Vych. Mat./ SB RAS. - Novosibirsk, 2011. - vol. 8., no. 1. - pp. 1-15.

[6] Ermakov, S.M., Mikhailov, G.A. Stokhasticheskoe modelirovanie uravnenii. -M.: Nauka, 1982.

[7] Ermakov, S.M., Nekrutkin, V.V., and Sipin, A.S., Sluchainye protsessy dlya resheniya klassicheskikh uravnenii matematicheskoi fiziki (Random Processes for Solving Classical Equations of Mathematical Physics), Moscow: Nauka, 1984.

[8] Snir M., Otto S. W., Huss-Lederman S., Walker D., and Dongarra J.. MPI: The Complete Reference. MIT Press. Boston, 1996.

[9] Korneyev, V.D., Parallel'noe programmirovanie v MPI (Parallel Programming in MPI), Moscow: Institute of Computer Sciences, 2003.

[10] Korneyev, V.D., Parallel'noe programmirovanie klasterov. Uchebnoe posobie (Parallel Programming of Clusters: Tutorial), Novosibirsk: Novosibirsk State Tech. Univ., 2008.

[11] http://software.intel.com/ru-ru/intel-mkl/

[12] Averina, T.A. and Artemiev, S.S., Numerical Solution of Stochastic Differential Equations with Increasing Variance, Sib. Zh. Vych. Mat., 2005, vol. 8, no. 1, pp. 1-10.

[13] Averina, T.A. and Artemiev, S.S., Analysis of Accuracy of Monte Carlo Methods in Solving Boundary Value Problems by Probabilistic Representations, Sib. Zh. Vych. Mat., 2008, vol. 11, no. 3, pp. 239-250.

[14] Panteleyev, A.V., Rybakov, K.A., Sotskova, I.A. Spektralnyi metod analiza nelineynikh stokhasticheskikh sistem upravleniya. -M.: Vuzovskaya kniga, 2006.

Monte Carlo Modeling of the Radiation Transfer in Stochastic Scattering Media

Boris A. Kargin
Institute of computational Mathematics and Mathematical Geophysics SB RAS
Novosibirsk State University
Novosibirsk, Russia
e-mail: bkargin@osmf.sscc.ru

Abstract

The problems of statistical simulation of light propagation in stochastic scattering media as applied to the problems of optics of aerosol cloudy atmosphere are considered. A set of Monte Carlo algorithms, allowing the construction of numerical models for the field of multiply scattered optical radiation in the aerosol atmosphere and stochastic cloudiness has been provided for the purpose. A special attention has been paid to solving the problem of optimization of Monte Carlo algorithms. The optimization is based on the method of "dependent trials".

Keywords: stochastic media, transfer equation, Monte Carlo method.

Introduction

To stochastic problems of transfer theory we refer the problems, where spatial variations of optical parameters of the scattering media are of random nature. The transfer of optical radiation in a substance may be described by integral equation [1]

$$f(\vec{x}) = \int_X^{k(\vec{x}',\vec{x})f(\vec{x}')d\vec{x}'+\psi(\vec{x})}, \tag{1}$$

$$k(\vec{x}',\vec{x}) = \frac{\Sigma_s(\vec{r}')g(\mu,\vec{r}')\exp(-\tau(\vec{r}',\vec{r}))\Sigma(\vec{r})}{2\pi\left|\vec{r}-\vec{r}'\right|^2\Sigma(\vec{r}')}\delta(\omega-\frac{\vec{r}-\vec{r}'}{|\vec{r}-\vec{r}'|}),$$

where $f(\vec{x})$ is the collision density, $\vec{x}=(\vec{r},\vec{\omega})$ and $\vec{x}'=(\vec{r}',\vec{\omega}')$ are the points of the phase space

$$X=\left\{\vec{r}\in R\subset R^3,\quad \vec{\omega}=(a,b,c)\in\Omega=(a^2+b^2+c^2=1)\right\},$$

$$\mu=\left((\vec{\omega}',\vec{r}-\vec{r}')\,/\,|\vec{r}-\vec{r}'|\right)$$

is the cosine of the scattering angle; $g(\mu,\vec{r})$ is the scattering phase function (scattering indicatrix) such that $\int_{-1}^{1} g(\mu,\vec{r})d\mu=1$; $\tau(\vec{r}',\vec{r})=\int_0^l\sum(\vec{r}(s))ds$ is the optical length of the segment $[\vec{r}',\vec{r}]$, $\vec{r}(s)=\vec{r}'+s(\vec{r}-\vec{r}')/l$, $l=|\vec{r}'-\vec{r}|$; $\Sigma(\vec{r})=\Sigma_a(\vec{r})+\Sigma_s(\vec{r})$ is the extinction coefficient

of the flux, $\Sigma_a(\vec{r})$ is the absorption coefficient, $\Sigma_s(\vec{r})$ is the scattering coefficient; $\psi(\vec{x})$ is the distribution density of the source, $\int\limits_X \psi(\vec{x})d\vec{x} = 1$. Equation (1) determines the corresponding random Markov chain of collisions with the initial states $\psi(\vec{x})$ and the transition density $k(\vec{x}', \vec{x})$. The Monte Carlo method is in the simulation of this chain of trajectories on a computer and of the computation of statistical estimates for the sought for functionals. We consider a problem of computing the linear functionals $J_\phi = (f, \phi)$ from the solution of Eq.(1); here $\varphi(\vec{x}) \geq 0$. In the stochastic case, the kernel $k(\vec{x}', \vec{x})$ and the density of collisions $f(\vec{x})$ are dependent on a random field $\sigma(\vec{r}) = (\sigma_1(\vec{r}), \ldots, \sigma_s(\vec{r}))$, denoting the set of s optical parameters of the medium. The problem is solved on the basis of computation of some random values $\xi(\omega, \sigma)$ given on the trajectories ω of the simulated random process such that $E_{\omega(\sigma)}[\xi(\omega, \sigma)/\sigma] = J_\phi(\sigma)$. Here $E_{\omega(\sigma)}$ denotes mathematical expectation with respect to the distribution of random trajectories ω depending on σ. Sought for functional is determined by $I_\phi = \langle J_\phi(\sigma) \rangle$, where $\langle\rangle$ denotes mathematical expectation with respect to a random field σ. The solution of the formulated stochastic problem by the Monte Carlo method is based on the principle of "double randomization" resulting from the relation:

$\langle J_\phi(\sigma) \rangle = \left\langle E_{\omega(\sigma)}[\xi(\omega, \sigma)/\sigma \right\rangle = E_{(\omega,\sigma)}\xi(\omega, \sigma)$.This relation shows that for the estimate J_ϕ it is sufficient to construct one trajectory ω for any realization of the random field σ. Thus, the problem to compute the functionals I_ϕ includes:
the construction of realizations of the random field σ;
simulation for each realization of σ of m conditionally independent trajectories of the Markov chain ($m \geq 1$);
computation of corresponding random values $\xi(\omega, \sigma)$.

It is well known that for the majority of real transfer problems to increase the efficiency of the Monte Carlo calculations usually the weight algorithms are used [2]. In order to construct weight algorithms we use the Markov chain with some initial density $r_0(\vec{x})$ and transition density $r(\vec{x}', \vec{x}\,|\sigma)$ which should contain the generalized multiplier $\delta(\vec{\omega} - \frac{\vec{r}-\vec{r}'}{|\vec{r}-\vec{r}'|})$from Eq. (1). In this case one should calculate the auxiliary weights

$$Q_0(\vec{x}_0) = \frac{\psi(\vec{x}_0)}{r_0(\vec{x}_0)}, \quad Q_n(\sigma) = Q_{n-1}(\sigma) \frac{k(\vec{x}_{n-1}, \vec{x}_n|\sigma)}{r(\vec{x}_{n-1}, \vec{x}_n|\sigma)} \tag{2}$$

at every collision, n - number of the state of random Markov chain (collision number). For the functional $J_\phi(\sigma)$ the random estimate (so called "collision" estimate) $\xi(\sigma) = \sum\limits_{n=0}^{N} Q_n(\sigma)\phi(\vec{x}_n)$ is calculated, like that $J_\phi(\sigma) = E_{\omega(\sigma)}\xi(\sigma)$, N – random number of the last state of the random trajectory $\omega(\sigma)$ (last collision before escape from the scattering medium).

1 Optimization of the Monte Carlo algorithm

Let us consider a problem to compute the functional I_ϕ for optical radiation scatted by the random inhomogeneous layer, in general, with three-dimensional continuous stochasticity, i.e. when

$\Sigma(\vec{r})$ is a random field ($\vec{r} \in R = (-\infty, +\infty)^2 \times [0, H]$). It means that $\sigma(\vec{r}) = \Sigma(\vec{r})$and all the rest optical parameters are known and determined. In this case we have two main problems. The first one is to construct a most adequate mathematical model of a stochastic layer and its efficient numerical implementation. In very many cases the sample size of numerical realizations of random field $\Sigma(\vec{r})$ should be statistically representative and the number of realizations may be compared to the number of trajectories. The second difficulty arises in simulation of photon trajectories in a random inhomogeneous 3D – medium. These both procedures take much computer time. To avoid these complications we suggest the following algorithm based on the well-known "dependent trials" method [1]. The given method's main idea is that the estimates of the sought for functional I_ϕ for different values of $\Sigma(\vec{r})$ may be obtained from the same random photon trajectories, using weights (2) to remove the appearing bias. Specifically, the trajectories constructed for $\Sigma(\vec{r}) = \Sigma_0(\vec{r})$ may be used to estimate the sought for functional for other values of $\Sigma(\vec{r})$, if after every transition $\vec{x}' \to \vec{x}$ the auxiliary weight of the particle is multiplied by the value $k(\vec{x}', \vec{x}\,|\Sigma)/r(\vec{x}', \vec{x}\,|\Sigma_0)$. Let $\omega_n = \{(\vec{x}_0, \vec{x}_1, ..., \vec{x}_n); \quad \vec{x}_i = (\vec{r}_i, \vec{\omega}_i), \quad i = \overline{0, n}\}$ be an arbitrary n-link trajectory, constructed with the transition density $r(\vec{x}', \vec{x}\,|\sigma) = k(\vec{x}', \vec{x}\,|\Sigma_0)$. Then one can easily see, that the weight multiplier $Q_n(\Sigma)$ corresponding to the realization of $\Sigma(\vec{r})$ is calculated by the formula

$$Q_n(\Sigma) = \prod_{i=1}^{n-1} \left\{ \frac{\Sigma_s(\vec{r}_{i-1})\Sigma_0(\vec{r}_{i-1})}{\Sigma(\vec{r}_{i-1})\Sigma_{s,0}(\vec{r}_{i-1})} \right\} \cdot \frac{\Sigma(\vec{r}_n)}{\Sigma_0(\vec{r}_n)} e^{-\sum\limits_{i=1}^{n} [\tau(\vec{r}_{i-1}, \vec{r}_i) - \tau_0(\vec{r}_{i-1}, \vec{r}_i)]}, \tag{3}$$

where $\tau_0(\vec{r}_{i-1}, \vec{r}_i) = \int\limits_0^{|\vec{r}_{i-1} - \vec{r}_i|} \Sigma_0(\vec{r}_{i-1} + s\vec{\omega}_{i-1})ds, \quad \tau(\vec{r}_{i-1}, \vec{r}_i) = \int\limits_0^{|\vec{r}_{i-1} - \vec{r}_i|} \Sigma(\vec{r}_{i-1} + s\vec{\omega}_{i-1})ds$. Let $\xi(\Sigma)$ be the random estimate of the functional $J_\phi(\Sigma)$ for the given realization of the random field $\Sigma(\vec{r})$. Then from (3) we can obtain that, in order to calculate $\xi(\Sigma)$ from the trajectory constructed with the transition density $k(\vec{x}', \vec{x}\,|\Sigma_0)$, it isn't necessary to construct the random field $\Sigma(\vec{r})$ at every point of the considered space, but we have to know the values of the random field $\Sigma(\vec{r})$ only at the points $\vec{r}_1, ..., \vec{r}_N$, i.e. $\Sigma(\vec{r}_1), ..., \Sigma(\vec{r}_N)$, as well as the values $\tau(\vec{r}_{i-1}, \vec{r}_i)$, along the directions $\vec{\omega}_i = (\vec{r}_i - \vec{r}_{i-1})/|\vec{r}_i - \vec{r}_{i-1}|$ on the segments $(\vec{r}_{i-1}, \vec{r}_i), i = \overline{1, N}$. Thus, the problem of calculating the functional $\langle J_\phi(\Sigma) \rangle$, instead of averaging a random functional $J_\phi(\Sigma)$ over the realizations of the random field $\Sigma(\vec{r})$, is reduced to determining it by the realizations of random vectors $\{\Sigma(\vec{r}_i)\}_{i=\overline{1,N}}$ and $\{\tau(\vec{r}_{i-1}, \vec{r}_i)\}_{i=\overline{1,N}}$. In many cases this allows to considerably decrease the computing time. The effectiveness of such an approach is dependent on the model of the random field $\Sigma(\vec{r})$. To illustrate this approach, we shall limit ourselves to considering the following problem, related to the problem of optical radiation propagation through the continuous stratus-type stochastic cloudiness.

2 Numerical experiment

Suppose the scattering substance occupies the space $R = (-\infty, +\infty) \times (-\infty, +\infty) \times [0, H]$ and all optical characteristics of the medium are not dependent on the horizontal coordinates, $\Sigma(\vec{r}) =$

$\Sigma(z)$ be a random process with correlation function $K_\Sigma(s)$ and marginal distribution $p(\Sigma)$ with mean value $\bar{\Sigma}(z)$ and variance σ. One of the most popular spectral model of the random process $\Sigma(z)$ is an approximate model (see, for example, [3])

$$\Sigma(z) \approx \Sigma^{(k)}(z) = \bar{\Sigma}(z) + \sigma_\Sigma \sum_{j=1}^{k} a_j \sqrt{-2\ln\alpha_j}\cos(\lambda_j z + 2\pi\beta_j),$$

where α_j and β_j are independent random values uniformly distributed in $[0,1]$, $a_j^2 = \frac{1}{k}$, λ_j are distributed on $[0,\infty)$ with probability density $s(\lambda) = \frac{2}{\pi}\int\limits_0^\infty \cos(\lambda z)K(z)dz$.In this case $\tau[\vec{r}_{i-1},\vec{r}_i]$, $i = 1,\ldots,n$ in (3) are calculated as follows

$$\tau[\vec{r}_{i-1},\vec{r}_i] =$$

$$= \frac{1}{|(\vec{\omega}_{i-1},\vec{k})|}\{\bar{\Sigma}|z_{i-1}-z_i| + \sigma_\Sigma \sum_{j=1}^{k}\frac{a_j}{\lambda_j}\sqrt{-2\ln\alpha_j}\times[\sin(\lambda_j z_{i-1}+2\pi\beta_j)-\sin(\lambda_j z_i+2\pi\beta_j)]\}.$$

Here $\vec{k} = (0,0,1)$. In the capacity of the correlation function was used the following function $K_\Sigma(s) = \exp(-\tau s/H)$,where $\tau = \int\limits_0^H \bar{\Sigma}(z)dz$ denotes the optical thickness of the determinate layer. The marginal distribution $p(\Sigma)$ of the random field$\Sigma(z)$in accordance with [4] was chosen as $p(\Sigma) = (H/\tau)\exp(-H\Sigma/\tau)$. To illustrate the influence of stochasticity we compare functionals from the solution of the transfer equation in a determinate flat cloud layer $0 \le z \le H$ with average values of those functionals corresponding to stochastic cloud layer when $\Sigma(z)$ is mentioned above random process. In the Table 1 the calculation results of the probability of passage through the scattering layer of radiation incident perpendicularly to the layer's upper boundary $z = 0$ are given, obtained by the described algorithm. The phase function $g(\mu,\vec{r}) = (1-\mu_0)/2 + \mu_0\delta(\mu-1)$with an average cosine of the angle of scattering $\mu_0 = 0.9$ was considered in the capacity of the scattering indicatrix. The probability of survival of a photon in a collision (albedo of single scattering) Σ_s/Σ was set to be 0.7.

Note: S – stochastical layer, D – determinate layer, $P_S$andP_D – probabilities of passage respectively through stochastic and determinate layers, ε - relative statistical error of the estimate for passage.

Conclusions

Due to the space limitation of the current paper it is not possible to carry out a more detailed analysis of, for example, the influence of 3D- stochasticity of the scattering medium on the transfer of optical radiation. But even the abovementioned calculation results for a simple enough model show that taking stochasticity into account gives an increase of the probability of passage. The ratio of the probability of passage in a stochastical medium to the probability of passage in a

Table 1: Probabilities of radiation passage through the determinate and stochastic scattering layers

Layers	P_S, P_D	$\varepsilon, \%$	Ratio P_S/P_D	τ
S	0.082	1.2	2.1	10
D	0.039	1.8		
S	0.022	2.2	2.9	15
D	0.0075	3.1		
S	0.0059	3.3	3.9	20
D	0.0015	3.8		

determinate medium grows with the increase of optical thickness. The obtained numerical results demonstrate that, when interpreting physical changes of radiation fields in a cloudy medium in some cases the stochastical properties of the medium play a significant role. The weight algorithm considered in this paper allows avoiding the laborious procedure of modeling random photon trajectories in randomly-inhomogeneous scattering media and therefore reducing the calculation time. Realization of the algorithm is quite simple. The profit in terms of calculation time depends upon the choice of model for the random field.

Acknowledgements

The work has been done under the financial support of the RFFR (grant 09-01-00035), SB RAS (grants 1.3.2 and 2.2).

References

[1] G. I. Marchuk, G. A. Mikhailov, M. A. Nazaraliev, R. A. Darbinian, B. A. Kargin, B. S. Elepov. (1989). *Monte Carlo methods in atmospheric optics*. Springer-Verlag, Berlin, Heidelberg, New York.

[2] G. A. Mikhailov. (1991). *Optimization of weighted Monte Carlo Methods*. Springer-Verlag, Berlin, Heidelberg.

[3] V. A. Ogorodnikov, S. M. Prigarin. (1996). *Numerical Modelling of Random Processes and Fields: Algorithms and Applications*. VSP, The Netherlands.

[4] V. N. Bol'shakov. (1975). On probability density of the optical depth in the lower and middle cloudiness. *Izv. AN SSSR, FAO*. Vol. **11**,No. **6**,pp. 590-598.

Monte Carlo Modeling in Problems of Lidar Remote Sensing of Crystal Clouds from Satellites

Boris A. Kargin, Arsenii B. Kargin, M.V. Lavrov
Institute of computational Mathematics and Mathematical Geophysics SB RAS
Novosibirsk State University
Novosibirsk, Russia
e-mail: bkargin@osmf.sscc.ru

Abstract

Laser sensing is an effective way of studying optical properties of various atmospheric structures. If we consider strongly scattering media, like clouds, there arises the necessity of taking into account the effects of multiple scattering which changes the space and time characteristics of the light pulse. The Monte Carlo method is the most convenient one for obtaining practical results in such problems. In this paper two problems were solved. One is constructing an adequate optical model of crystal clouds taking into account optical anisotropy of the medium. The other is Monte Carlo modeling of laser radiation transfer in such a medium. The form and duration of light pulses reflected by clouds (lidar returns) are obtained by the Monte Carlo method in the case of single layer continuous crystal cloud and double layer continuous cloudiness (a crystal cloud of highest level is located above a drop cloud).

Keywords: Monte Carlo method, transfer equation, laser radiation, crystal clouds, optical anisotropy.

The object of the current paper is to construct an adequate optical model of crystal clouds and establish connections between the properties of a light impulse reflected from a cloud (lidar return) and certain parameters of the cloudy medium. Knowing these connections allows determining which parameters of cloudiness may be obtained from a reflected signal with a given degree of reliability. A no less important task is taking into account of thin clouds in remote sensing of the ocean with optical methods. To solve those problems, data must be obtained about the form and size of the time-base of the light impulse reflected from clouds when illuminated by an impulse source. The problem was solved in the conditions of single-layer and double-layer continuous cloudiness for various optical parameters of the clouds and various properties of the source and receiver. The calculations were carried out using the Monte Carlo method.

Unlike drop clouds, for which quite long ago a Mie solution was obtained as a result of applying Maxwell's electromagnetic field theory to the problem of scattering of light on a homogeneous spherical particle [1], the properties of crystal clouds are little-studied. It is due to the fact that crystal clouds are an anisotropic medium, i. e. its optical characteristics (scattering indicatrix, single scattering albedo, attenuation, scattering and absorption cross-sections) depend on the direction of the photon propagation before colliding with a crystalline particle.

For their calculation in accordance with the idea from [2, 3] it's convenient to define the position of a crystal in space through Euler angles α, β, γ (see Fig. 1). Those angles, describing

the orientation of a rigid body in 3-D Euclidean space allow defining any rotation of the system in the initial (global) coordinate system. Denote the initial coordinate system with (x, y, z), the transformed one with (X, Y, Z). The intersection of the coordinate planes xy and XY is called the line of nodes N. α is the angle between axis x and the line of nodes N, β is the angle between axes z and Z, γ is the angle between axis X and the line of nodes N: $\alpha \in [0, 2\pi)$, $\beta \in [0, \pi]$, $\gamma \in [0, 2\pi)$.

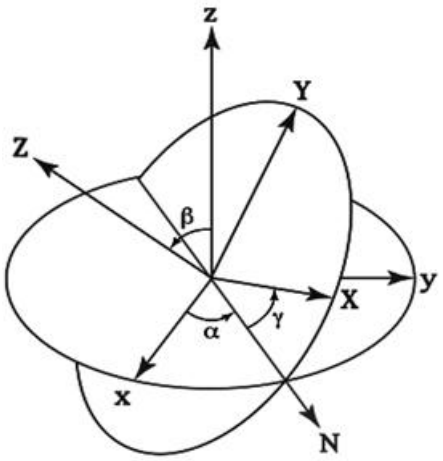


Figure 1: Defining Euler angles α, β, γ.

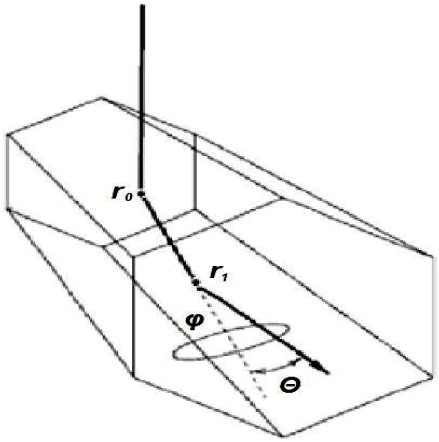


Figure 2: Modeling of photon trajectory in a crystal.

To clarify further calculations let's choose a regular 6-prism as an example. The coordinates of all inner and surface points of the crystal shall be set in the $X'Y'Z'$ system so that the origin

coincides with the geometrical center of the crystal, the axis $O'Z'$ is directed along the axis of symmetry of the crystal connecting the centers of its base faces (further on in this text we'll refer to it simply as the crystal's axis), the axis $O'Y'$ passes through the center of a lateral edge and the axis $O'X'$is perpendicular to one of the crystal's lateral faces and passes through its center. Note, that all numerical results (concerning crystal clouds) given in section 5 are obtained for the case when crystal clouds consist of crystals having the form of regular 6-prisms. According to [3,4], this form of crystal, among a great variety of form in dependence of temperature is present in from 10 to 60 percent cases. Any orientation of the rotation of the crystal's axis relative to the global coordinate system may be described by the triplet $[\alpha, \beta, \gamma]$. Let's make a discretization of the space in Euler angles $[n_\alpha, n_\beta, n_\gamma]$. Place the crystal into a regular cuboid and from random points on one fixed face let's emit rays inside the regular cuboid. Part of them will pass through the crystal and experience inside it refraction and scattering (see Fig. 2). Random photon trajectories inside crystals are modeled in a standard way used for modeling the transfer process in optically isotropic media. After refraction of an incoming photon trajectory at a certain random point r_0 on one of the faces of the crystalline particle (see Fig.2) a random value of the free pass length is generated $l = -(\ln \nu)/\sigma_c$, where σ_c – the attenuation cross-section inside the crystal, ν is a random value uniformly distributed on the segment (0,1). If the photon does not reach any of the particle's faces then at the point $r_1 = r_0 + \omega_0 \cdot l$ a new direction of the photon's movement ω_1 is generated. ω_0 is the direction of the refracted ray. The vector ω_1 is defined by the local inclination angle Θ and azimuth angle ϕ (Fig. 2). Here ϕ is a value distributed uniformly on the segment (0,2π) and the inclination angle Θ is calculated from the relation $\int\limits_0^{\Theta} g_c(\Theta') \sin(\Theta') d\Theta' = \nu \int\limits_0^{\pi} g_c(\Theta') \sin(\Theta') d\Theta'$ where g_c denotes the inner scattering function. Absorption is taken into account by multiplying the weight of the photon by the inner single scattering albedoq_c. This procedure is repeated while the photon is inside the crystal where it experiences scattering, reflection and refraction. The modeling process stops when the photon's weight becomes lower than a certain value or when the refracted ray escapes the crystal's boundaries. Using the method described above we can obtain the matrix $G(r, \omega, \omega')$ of order $n_\alpha \times n_\beta \times n_\gamma$, its' elements being the scattering functions of a cloudy medium $g_{ijk}(r, \omega, \omega'), \quad i = 1, ..., n_\alpha, j = 1, ..., n_\beta, k = 1, ..., n_\gamma$. In addition we must take into account diffraction on a projection of the particle of area S. For polyhedral particles the projection is a closed polygon for which the diffraction scattering function Σ_{ext}^D may be calculated analytically. Beam and diffraction properties are summed weighted with their own scattering cross-sections. By definition, the beam scattering cross-section is equal to the geometrical section of a particle $\Sigma_{ext}^s(\omega) = S(\omega)$and the beam scattering cross-section Σ_{sca}^s is the direction of the radiation incident to the crystal. The diffraction attenuation cross-section $\Sigma_{ext}^D(\omega) = S(\omega)$ is always equal to the diffraction scattering cross-section $\Sigma_{sca}^D(\omega)$. Therefore the full attenuation cross-section, albedo of a single scattering event and scattering functions are of the form [2]

$$\Sigma_{ext}(\omega) = \Sigma_{ext}^s(\omega) + \Sigma_{ext}^D(\omega) = 2S(\omega),$$

$$\Sigma_{sca}(\omega) = \Sigma_{sca}^s(\omega) + \Sigma_{sca}^D(\omega) = (1 + q_c(\omega)) S(\omega),$$

$$q(\omega) = \frac{\Sigma_{sca}(\omega)}{\Sigma_{ext}(\omega)} = \frac{1 + q_c(\omega)}{2},$$

$$g(r, \omega, \omega') = \frac{\Sigma^s_{sca}(\omega)\, g(r, \omega, \omega') + \Sigma^D_{sca}(\omega)\, g(r, \omega, \omega')}{\Sigma^s_{sca}(\omega) + \Sigma^D_{sca}(\omega)}.$$

To obtain the macroscopic attenuation, scattering, absorption cross-sections the corresponding microscopic cross-sections must be multiplied by the concentration of particles n_0 in a unit volume ($\sigma = \Sigma_{ext} n_0$, $\sigma_s = \Sigma_{sca} n_0$).

In the calculations it was stipulated that the scattering medium consists of prisms with regular hexagon base faces and rectangular lateral faces. Then the orientation of the crystal in space is explicitly described by the position of its axis. All in all, as an addition to the matrix $G(r, \omega, \omega')$ we have two tables of size $n_\alpha \times n_\beta \times n_\gamma$, in each cell of which we keep a triplet

$$\sigma(r, \omega\,|\alpha, \beta, \gamma), \quad q(r, \omega\,|\alpha, \beta, \gamma).$$

Consider a model for laser sensing with a monostatic LIDAR. It means that the source of the signal (emitter) and receiver are located at one and the same point. Suppose that a delta-impulse of unit energy (with a certain wavelength λ) is emitted from point $r*$ in space vertically into the atmosphere in a certain cone of directions with the semiaperture Θ_s. At the same point $r*$ a receiver with area $S*$ is located which registers radiation incoming vertically upwards in a spatial angle bounded with a cone with a plane angle $2\Theta_d$ (Θ_d - semiaperture of the detector). Suppose the optical characteristics of the atmosphere and underlying surface are known. The problem consists of calculating the echo-signal $J(t)$ which will be registered by the receiver (LIDAR return – time distribution of the laser radiation reflected by the underlying medium). The process of transfer of optical radiation (without taking into account of polarization) may be described by the following non-stationary integral equation

$$f(r, \omega, t) = \int\limits_{R^3} \int\limits_{\Omega} \int\limits_{0}^{\infty} \frac{q(r', \omega')\sigma(r, \omega') e^{-\tau(r', r)} g(r', \omega', \omega)}{2\pi |r - r'|^2} \times \delta\left(\omega - \frac{r - r'}{|r - r'|}\right) \times$$
$$\times \delta\left(t - \left(t' + \frac{|r - r'|}{v}\right)\right) \cdot f(r', \omega', t')\, dt' d\omega' dr' + \psi(r, \omega, t).$$

Here $f(r, \omega, t)$ is the particle (photon) collision density, $r = (x, y, z), r' = (x', y', z') \in R^3$; $\tau(r', r) = \int\limits_{r}^{r'} \sigma(\rho, \omega)\, d\rho$ is the optical length of the segment $[r', r]$; $\omega, \omega' \in \Omega = \{\omega \in R^3, |\omega| = 1\}$; δ is the Dirac delta function; $q(r', \omega')$- single scattering albedo; $\sigma(r', \omega')$ is the attenuation coefficient; $\psi(r, \omega, t)$ is the density of sources; v is the velocity of light propagating in the medium; $g(r', \omega', \omega)$ is the scattering indicatrix at the point r' satisfying the normalization condition $\int_\Omega g(r', \omega', \omega) = 1$.

With the Monte Carlo method the following functional was estimated

$$J(r^*,\omega^*,t)=J^*(t)=\int\limits_D\int\limits_\Omega \frac{f(r,\omega,t)}{\sigma(r,\omega)}dr\,d\omega=(f,\varphi),$$
$$\varphi(r,\omega,t)=\begin{cases}\frac{1}{\sigma(r,\omega)}, & r\in D_r,\ \ \omega\in\theta_d\\ 0, & overwise\end{cases}.$$

Within the bounds of the model under consideration the transfer process is a homogenous Markov chain of collisions of a photon with matter particles. An important distinction of an optically anisotropic medium for an isotropic one is the fact that the functions $q(r',\omega'),\ \ \sigma(r,\omega')$ depend on the previous direction ω'.

Let's describe one possible algorithm of constructing an estimate $J^*(t)$for the trajectory ensemble $W_i=\left\{(p_{n-1}^{(i)},\omega_{n-1}^{(i)},r_n^{(i)},t_n^{(i)}:n=0,\ldots,N(i)\right\}$.Here, apart from the photon's weight, the moment of time t_n in which the $n-$th collision occurs is also an element of trajectory, $t_0=0$. At first $p_{-1}=p_0=1$, afterwards the weight is multiplied by the scattering probability $p_n=q\cdot p_{n-1},n\geq 1$. For each trajectory $W=\left\{\left(p_{n-1},\omega_{n-1},r_n,t_n^{(i)}\right)\right\}$the so-called local estimate is constructed

$$L_n^i=p_{n-1}q\left(r_n,\omega_{n-1}\right)e^{-\tau(r_n,r^*)}g\left(r_n,\omega_{n-1},\omega^*\right)\frac{S^*}{2\pi\left|r_n-r^*\right|^2}\Delta\left(r_n\right)\Delta\left(t_n^{(i)}\right),$$

where $\Delta\left(r\right)$ is the indicator function of the receiver equal to 1 if the point r is located inside the receiver cone and equal to 0 otherwise, $\Delta\left(t_n^{(i)}\right)$- is the indicator function of the i-th interval in the time hystogram. Thus, only those collisions make a contribution to the echo-signal which take place in the field of view of the receiver. The local estimate L_n^i gives a contribution to the echo-signal at the moment of time $t_n^{(i)}=t_n+t^*$, where t^* is the time necessary for the photon to get in a straight line from the point r_n to the receiver r^*. On the local estimates obtained from an ensemble of M trajectories a frequency bar chart is constructed:

$$J^*(t)=\frac{1}{M\Delta t_k}\sum_{\hat{t}_n^{(i)}\in\Delta t_k}L_n^{(i)},\quad t\in\Delta t_k \tag{1}$$

Here $\Delta t_k$are the intervals into which the time axis is partitioned for the chart's construction. This chart is the estimate of the echo-signal$J^*\left(t\right)$.

In the case of an anisotropic medium standard calculation algorithms of the Monte Carlo method must be modified considerably. Suppose the crystal cloud is a homogeneous medium (single scattering albedo, attenuation coefficient and scattering indicatrix are not dependent on the space variable) consisting of identical crystals of 6-prism form. Let (a,b,c) be the direction of the photon's movement. Before the collision of a photon and a crystal its position in the cloud is modeled according to a given distribution for crystal's axes. For example, all axes and two faces of the crystals lie in a horizontal plane and Euler angles define a crystal's orientation in the global coordinate system $[\alpha',\beta',\gamma']$. Place the crystal in a regular cuboid in such a way that a beam is perpendicularly incident to one of the faces. Transforming to the crystal's local coordinate

system we can find Euler's angles of the crystal in the regular cuboid [5]: $\alpha = \alpha' - tg^{-1}\left(\frac{a}{b}\right), \beta = \beta' + \cos^{-1}(c), \gamma = \gamma'$.

In the table $[n_\alpha, n_\beta, n_\gamma]$ we find the optical properties

$$g(r, \omega\,|\alpha, \beta, \gamma), \sigma\,(r, \alpha, \beta, \gamma)\,, \sigma_s(r, \alpha, \beta, \gamma), q\,(r, \alpha, \beta, \gamma)$$

for this orientation.

For numerical examination of the dependence of the time-base of the echo-signal from orientations of crystals in a crystal cloud the double-layer model of cloudiness were chosen. In this model the crystal cloud consists of crystals having the form of regular 6-prisms with a height to radius ratio 200 μm/100 μm for three kinds of crystal's orientation in space described below. In the double-layer variant a drop cloud is located beneath the crystal cloud. The concentration of the crystals is 0.5 cm^{-3}. To calculate the sought for values of intensity of reflected radiation the local estimate (1) was used. The distance from the source to the upper boundary of cloudiness was set to Δh=200 km. A combined source-receiver system was considered. The source and receiver are disks of diameter D_s and D_r. The source emits and the receiver registers light inside cones with whole conical angles Θ_s and Θ_d accordingly. Their axes coincide and are directed parallel to the OZ axis. The impulse emitted by the source is considered to be a δ-function of time. The following variant was considered: D_s =19 cm, $\Theta_s = 2'$, D_r= 28 cm, $\Theta_d = 3'$. The density functions of the crystal's positions in the cloud were taken from the book [4]. For the drop cloud $h_1 = 0.2$km, it's thickness is 0.3 km. The volume scattering coefficient in the drop cloud was set to be constant at the same altitude and equal to $\sigma_w = 30$ km^{-1}. For the crystal cloud the altitude of lower boundary was set to $h_3 = 6$ km, and the cloud's thickness to 0.5 km. In the crystal cloud the crystals' orientation is isotropic. The scattering indicatrix for the drop cloud was taken from [1] for a C1 cloud model with wavelength $\lambda = 0.7\mu$m. In this case we can see two clear spikes (see Fig. 3). The first spike represents the echo-signal off the crystal cloud. After a certain time period the echo-signal has a second spike connected with the reflected radiation off the drop cloud.Relative error of the calculations did not exceed 2%. Using these values we model the photon's movement in a standard way.

In this paper the problem of statistical modeling of radiation transfer in crystal and drop clouds was considered. An anisotropic optical model of crystal clouds, taking into account diffraction on the crystals, was proposed and examined. This model is sufficiently versatile and allows modeling radiation transfer for various crystals' orientations inside the cloud. From the calculations' results a conclusion can be made that crystal clouds, despite their small optical thickness, have a considerable influence on the reflected impulse signal for LIDARs of aerospace basing.

The work has been done under the financial support of the RFFR (grant 09-01-00035), SB RAS (grants 1.3.2 and 2.2).

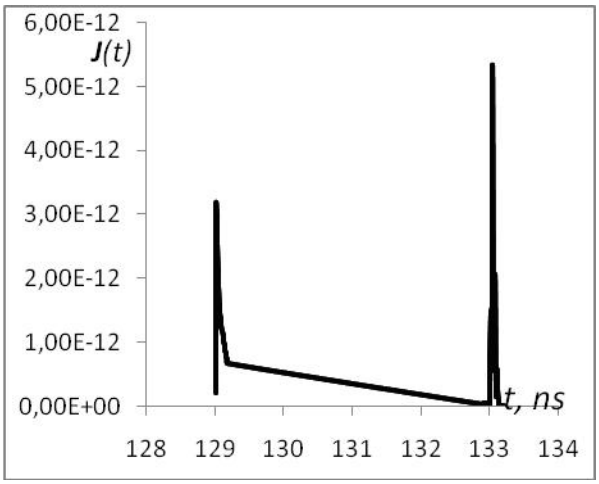


Figure 3: Time-base of the echo-signal (double-layer variant)

References

[1] Deirmendjian D. (1969). *Electromagnetic scattering on spherical polydispertions*. Elsevier, New-York.

[2] Mishchenko M., Hovenier J. W.,Travis I. D.. (2000).*Light Scattering by Nonspherical Particles. Theory, Measurements and Applications.*Academic Press, San Diego.

[3] Macke A., Mishchenko M., Cairns B. "The influence of inclusions on light scattering by large ice particles." J. of Geophysical research, 101(D18), 23311-23316 (1996)

[4] Liou K. N. (2002), *An introduction to atmospheric radiation*]. Elsevier, New-York

Goldstein H., Poole C., Safko J. (1980), *Classical Mechanics*. Addison Wesley, Reading.

Application of the Modified Method of the Maximum Section for Statistical Modeling of Systems with a Separated Time

Tatyana A. Averina
Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
Novosibirsk State University
Novosibirsk, Russia
e-mail: ata@osmf.sscc.ru

Abstract

The algorithm for statistical modeling of systems with a separated time, which can be described as a system with a distributed change of structure has been constructed. The offered algorithm is based on numerical methods of the solution to the stochastic differential equations and uses the modified maximum cross section method when the intensity of transition depends on a vector of state.

Keywords: Numerical methods, stochastic differential equations, systems with random structure, systems with a separated time, maximum cross section method.

Introduction

Many models of dynamic systems in various areas of science (including automatic control) are described by stochastic differential equations (SDEs). The term *dynamic systems with a sudden random changed structure* (or *random-structure system*) appeared in the 60s. These are dynamic systems, which on random time intervals are described by different SDEs [1]. Their characteristic feature is a sudden change of some system parameters.

There are many different probabilistic analysis problems of random-structure systems. For example, the estimate of the probability of the fact that the system is in any of its possible structures (the structure probability); the estimates of mean transition time from one structure to another and variance of this time; the estimate of distribution or the probabilistic characteristics of all or a part of phase coordinates. There problems can be solved by integrating the generalized Focker-Planck-Kolmogorov equation - a second-order partial differential equation for the distribution density of solution. Difficulties arising in solving problems by this method are associated with complicated computational procedure for solving partial differential equations. Other approximate methods of the probabilistic analysis of random-structure systems are based on two-moment parametric approximation of distribution density and statistical linearization. Thus, one obtains a system of ordinary differential equations for expansion coefficients, viz. probabilistic moments, cumulants or other time-dependent values. A serious disadvantage of these methods is the complexity of obtaining approximation estimations and the fact that changing the original model causes essential changes in the equations for probabilistic characteristics.

A statistical simulation algorithm for the probabilistic analysis of *systems with a distributed change of structure* was constructed in [2]. It was based on the numerical methods for stochastic differential equations [3] and maximum cross section method [4].

In this paper, we consider a new modification of this algorithm. We use the modified maximum cross section method [5, 6]. New algorithm is simple and allows us to estimate different probabilistic characteristics of the solution with high accuracy. The computational cost of modified algorithm is lower then that of algorithm from [2].

1 Systems with distributed change of structure

A random-structure system is defined by the state vector $\mathbf{Y}(t)$ and a structure number $L(t) = 1, ..., N_0$; where N_0 is a number of determinate structures. The structure number $L(t)$ is a discrete random scalar process with integer values.

The vector equation for the fixed l–th structure has the form of stochastic differential equations (SDEs) in the Stratonovich sense:

$$d\mathbf{Y}(t) = \mathbf{a}^{(l)}(\mathbf{Y}, t)dt + \sigma^{(l)}(\mathbf{Y}, t)d\mathbf{W}(t), \mathbf{Y}(t_0) = \mathbf{Y}_0, \quad l = 1, ..., N_0. \tag{1}$$

The state vector $\mathbf{Y}(t)$ of the system for each of the l–th structures is the n–dimensional continuous random process; $\mathbf{W}(t)$ is an m dimensional standard Wiener process;
$\mathbf{a}^{(l)}(\mathbf{Y}, t)$ is an n–dimensional vector function; $\sigma^{(l)}(\mathbf{Y}, t)$ is a matrix-valued function of $n * m$ dimension; $\mathbf{Y}_0$ is the initial state of the system.

The discrete random process $L(t)$ may be an arbitrary non-Markov process, Markov process or a conditional Markov process from the vector $\mathbf{Y}(t)$ [1].

We consider *systems with a distributed change of structure* when the process $L(t)$ is Markov process or a conditional Markov process, the dependence on the vector $\mathbf{Y}(t)$ statistically being given. The conditional transition probabilities from the l-th structure to the r-th one within a short time interval Δt are defined through the conditional transition intensity $\nu_{lr}(\mathbf{Y}, t)$ from the l-th structure to the r-th one and have the form [1]:

$$p_{lr}(r, t + \Delta t|l, t, \mathbf{Y}) = \nu_{lr}(\mathbf{Y}, t)\Delta t + o(\Delta t), \quad l \neq r;$$

$$p_{ll}(l, t + \Delta t|l, t, \mathbf{Y}) = 1 - \nu_{ll}(\mathbf{Y}, t)\Delta t + o(\Delta t), \quad \nu_{ll}(\mathbf{Y}, t) = \sum_{r=1 \neq l}^{N_0} \nu_{lr}(\mathbf{Y}, t),$$

where $o(\Delta t)$ is a small value of order no less than $(\Delta t)^2$, $\nu_{lr} \geq 0$.

The conditional distribution density $f_{\tau_{lr}}(t, \tau)$ of the transition time interval τ from the l-th structure to the r-th one is the exponential density and has the form

$$f_{\tau_{lr}}(t, \tau) = \nu_{lr}(\mathbf{Y}, t + \tau)exp(-\int_0^\tau \nu_{lr}(\mathbf{Y}, t + t_1)dt_1). \tag{2}$$

The lr-th elements of *absorption and reconstruction functions* have the forms [1]:

$$v^*_{lr}(\mathbf{Y},t) = \nu_{lr}(\mathbf{Y},t)p_1^{*(l)}(\mathbf{Y},t), \quad l,r = 1,...,N_0; l \neq r,$$
$$u^*_{lr}(\mathbf{Y},t) = \int_{-\infty}^{\infty} \nu_{lr}(\mathbf{Y}',t)p_1^{*(l)}(\mathbf{Y}',t)q^{lr}(\mathbf{Y},t|\mathbf{Y}',t)d\mathbf{Y}', \tag{3}$$
$$v^*_{ll}(\mathbf{Y},t) = u^*_{ll}(\mathbf{Y},t) = 0,$$

where $q^{lr}(\mathbf{Y},t|\mathbf{Y}',t)$ is a conditional distribution density of reconstruction of the r-th structure from the l-th, $p_1^{*(l)}(\mathbf{Y},t)$ is a conditional distribution density of $\mathbf{Y}$ on condition that $L(t) = l$ (* mean non-absorbed realizations, i.e., the function $p_1^{*(l)}(\mathbf{Y},t)$ is non-normalized).

The one-dimensional distribution density $p_1^{*(l)}(\mathbf{Y},t)$ for each of the l–th structures of the system is satisfied by the generalized Focker-Planck- Kolmogorov equation [1]. The differential equation for the l–th structure probability $P^{(l)}(t) = \int_{-\infty}^{\infty} p^{*(l)}(\mathbf{Y},t)d\mathbf{Y}$ has the form

$$\dot{P}^{(l)}(t) = -\sum_{r=1}^{N_0}\int_{-\infty}^{\infty} v^*_{lr}(\mathbf{Y},t)d\mathbf{Y} + \sum_{r=1}^{N_0}\int_{-\infty}^{\infty} u^*_{rl}(\mathbf{Y},t)d\mathbf{Y}.$$

The form of the function q^{lr} in (3) is defined by the physical sense of a problem. This function defines the initial conditions of reconstruction of the r-th structure from the l-th one. For example, if

$$q^{lr}(\mathbf{Y},t|\mathbf{Y}',t) = \delta(\mathbf{Y} - \mathbf{Y}'),$$

then the reconstruction is *exact*, i.e., the finite value of the process in the previous structure is equal to the initial value of the process in the new structure.

2 Algorithms for statistical modeling of an inhomogeneous Poisson process

The inhomogeneous Poisson process $\xi(t) = \xi([0,t])$, $t \geq 0$ can be considered as an inhomogeneous Poisson ensemble in the one-dimensional case. The algorithms proposed in [6] can be used for modeling inhomogeneous Poisson processes.

It is well known (see [7]) that the probability density of the time intervals between the adjacent points of the Poisson process is distributed exponentially. For that reason, Poisson processes are usually simulated using the exponential probability distribution.

The following algorithm is a direct consequence of Algorithm 1 from [6].

Maximum cross section method for modeling an inhomogeneous Poisson process under the assumption $\lambda(t) \leq \lambda_0, \quad t \geq 0$: *If $t_1, \ldots, t_{k-1}$ is an ordered sequence of the Poisson point process with the intensity $\lambda(t)$, then two sequences of independent sample values are constructed to model t_k. These are the sequence $\{\theta_i\}$ with the probability distribution density $\lambda_0 exp(-\lambda_0 t)$ $\{\alpha_i\}$ and the sequence $\{\alpha_i\}$ with the elements uniformly distributed in (0, 1). Set $\zeta_n = \sum_{i=1}^{n}\theta_i$. Let*

$$N = min\{n : \alpha_n \leq \lambda(t_{k-1} + \zeta_n)/\lambda_0\}.$$

Then, $t_k = t_{k-1} + \zeta_N$ *and* $\xi(t_k) = k$.

The validity of this algorithm was rigorously proved in [8] for the case of a constant majorant; for a variable majorant, it was heuristically justified in [9].

Using Theorem 1 from [6] and its corollaries, one can derive the following algorithm.

Modification of the maximum cross section method for modeling an inhomogeneous Poisson process under the assumption $\lambda(t) \leq \lambda_0, \quad t \geq 0$: *If* $t_1, \ldots, t_{k-1}$ *is an ordered sequence of the Poisson point process with the intensity* $\lambda(t)$*, then, in order to model* t_k*, a sequence* $\{\theta_i\}$ *of independent sample values is constructed with the probability density* $\lambda_0 exp(-\lambda_0 t)$*. Set* $\zeta_n = \sum_{i=1}^{n} \theta_i$*. Let*

$$N = min\{n : 1 - \alpha > \prod_{i=1}^{n} \left(1 - \frac{\lambda(t_{k-1}+\zeta_i)}{\lambda_0(t_{k-1}+\zeta_i)}\right)\},$$

where α *is a random variable uniformly distributed in (0, 1). Then,* $t_k = t_{k-1} + \zeta_N$ *and* $\xi(t_k) = k$.

This method is more efficient than the maximum cross section method. It produces a sequence of random variables by calling a random variable generator only once.

The results of the statistical modeling are presented in [6]. The results show that the computational cost of modified maximum cross section method is lower than that of maximum cross section method. It is seen that the estimates obtained by both algorithms are within the confidence intervals and the computation time of modified maximum cross section method is at least 10% less than the computation time of maximum cross section method.

3 Modified algorithm of statistical simulation of dynamic systems with distributed change of structure

The algorithm for statistical simulation of solutions of dynamic Markov systems with a distributed independent change of structure was described in [10]. This algorithm was generalized for statistical simulation of solutions of dynamic systems with a distributed conditional Markov change of structure using maximum cross section method [2]. Now we modify this algorithm using modification of the maximum cross section method.

Since each structure is described by the SDEs, then the numerical algorithm should include: 1) the solution of the SDE systems and 2) modeling of absorption and reconstruction conditions.

The transition from the l–th structure may be to each r-th structure, $r = 1, ..., N_0$, $r \neq l$. The transition time interval τ_{lr} has the conditional probability density (2). We propose the modified maximum cross section method for simulation of moments of a structure change, because the transition intensities depend on phase coordinates. The application of this method requires execution of the following conditions

$$\nu_{li}(\mathbf{Y}(t), t) \leq \nu_{li}^m = const, \quad i = 1, ..., N_0, \quad i \neq l$$

within the whole time interval $[0, T]$.

The algorithm for statistical simulation of the transition from the l-th structure for systems with a distributed (independent Markov or conditional Markov) change of structure:

1. Let the system at the moment t_k be in the l-th structure and the state vector be $\mathbf{Y}_k$.

2. We simulate additional random variable α_1 with an uniform density on the interval (0,1) and we establish the counter, set $z = 1$; .

3. We simulate a possible moment of leaving the l-th structure $t_{k+1} = t_k + \tau$, where τ is a random variable with distribution density $p(x) = \nu_l^m * exp(-\nu_l^m x)$, $\nu_l^m = \sum\limits_{i \neq l} \nu_{li}^m$ (by the formula $\tau = -ln\alpha/\nu_l^m$, α with an uniform density on the interval $(0, 1)$).

4. We simulate the number r (a possible number of a new structure) with probability $p_r(x) = \frac{\nu_{lr}^m}{\nu_l^m}$, $r \neq l$, $r = 1, ...N_0$;

5. we solve equations (1) for the l-th structure on the interval $[t_k, t_{k+1}]$ by the numerical method for the SDEs [2] with time step $h > 0$ and find $\mathbf{Y}_{k+1}$, viz. the state vector of the system at the moment t_{k+1} (the step must be consistent with transition intensity, for example, $h \leq 0.1/\nu_l^m$).

6. $t_k := t_{k+1}$, $\mathbf{Y}_k := \mathbf{Y}_{k+1}$; $z := z * (1 - \nu_{lr}(\mathbf{Y}_k, t_k)/\nu_{lr}^m)$.

7. We verify the structure change condition: if $1 - \alpha_1 > z$, then go to 8); else go to 3).

8. We change the structure number for r-th; we are modelling $\mathbf{Y}_k$ according to the required conditional reconstruction density q^{lr} .

Remark. Item 7) of the algorithm is absent for Markov systems with a distributed *independent* change of structure, because the verified condition is always valid.

The examples used for verifying the statistical simulation algorithm were taken from [1, 7].

Conclusions

The new modified algorithm is proposed for statistical modeling of system with a separated time. The computational cost of this algorithm is lower than computational cost of algorithm from [2], because we used a sequence of rejections with respect to the same random number.

In addition to decreasing the computational cost (which can be insignificant), the decrease in the number of the employed values of α decreases the constructive dimension of the algorithm related to the multidimensional uniformity of the pseudorandom numbers used in the algorithm (see [11]).

Acknowledgements

This work was supported by the Russian Foundation for basis Research, project nos. 09-01-00798 and 11-01-00282.

References

[1] Kazakov I.YE., Artemiev V.M., Bukhalev V.A. (1993). *Analysis of systems with random structure*. Moscow, Nauka.

[2] Averina T.A. (2004). Algorithm of statistical simulation of dynamic systems with distributed change of structure. *Monte Carlo Methods and Applications*. Vol. **10**, N 3-4, pp. 221–226.

[3] Artemiev S.S., Averina T.A. (1997). *Numerical Analysis of Systems of Ordinary and Stochastic Differential Equations*. VSP, Utrecht, The Netherlands.

[4] Ermakov S.M., Mikhailov G.A. (1982). *Statistical Modeling*. Moscow, Nauka.

[5] Mikhailov G.A., Averina T.A. (2009). The maximal section algorithm in the Monte Carlo method. *Doklady Akademii Nauk*. Vol. **428**, N. 2, pp. 163165.

[6] Averina T.A. (2010). New Algorithms for Statistical Modeling of Inhomogeneous Poisson Ensembles. *Computational Mathematics and Mathematical Physics*. Vol. **50**, N. 1, pp. 1218.

[7] Tikhonov V. I., Mironov M.A.. (1977). *Markov processes*. Moscow, Sov. Radio.

[8] Coleman W.A. (1968). Mathematical verification of a certain Monte Carlo sampling technique and applications of the techniques to radiation transport problems. *J. Nucl. Sci. and Eng.*. Vol. **32**, N 1, pp. 76-81.

[9] Mikhailov G.A. (1970). Method of modeling of length of free path of particles. *Atomic energy*. Vol. **28**, N 1, pp. 175.

[10] Averina T.A. (2001). Algorithm for statistical simulation of two types of random-structure systems. *Russ.J.Numer.Anal.Math.Modelling*. Vol. **16**, N 6, pp. 467-482.

[11] Sobol' I. M. (1973). *Numerical Monte Carlo Methods*. Moscow, Nauka.

Construction of "Modelled" Probabilistic Densities

Anton V. Voytishek[1], Irena E. Graifer[2]
Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia[1]
Zefat Academic College (Northern branch of Bar-Ilan University), Israel[2]
e-mail: vav@osmf.sscc.ru, irena.graifer@gmail.com

Abstract

In this study, we represent some recommendations for construction the probabilistic densities allowing efficient numerical realization of the sample values.

Keywords: Monte Carlo methods, numerical statistical simulation, numerical realization of sample value of random variable, probabilistic densities, method of inverse distribution function, numerical simulation of stochastic vectors, superposition method, majorant rejection method

Introduction

Considering the numerical models with random parameters (particularly, applying the Monte Carlo methods), emerges the necessity to choose the parameters distribution laws (based on experimental statistical data) and at the same time, the need for algorithms which would numerically realize the sample values of those chosen parameters with respect to the corresponding probability laws (see, eg, [1–4]). Thereby, it is worthwhile to engage in the detailed specification of the class of distributions which allows to design high-effective algorithms for numerical simulation.

This paper presents the opportunities for construction of the probabilistic distribu-tions, allowing effective use of following methods of sample value realization: the inverse distribution function algorithm, the method for simulation the two-dimensional random vector with dependent components, the integral and discrete superposition algorithms; the majorant rejection method (the relevant algorithms are described, e. g., [1]). In particular it lays the foundation for the "bank" of the "modelled" probabilistic distributions. Such bank may be used for constructing high-efficient algorithms for numerical statistical simulation.

1 Technology of recursive substitutions

Let's review *the standard algorithm* for numerical realization of the sample value ξ_0 of continuous random variable ξ which is distributed on the interval (a, b) with respect to the continuous distribution function $F(x)$, monotone increasing on (a, b) or *the method of inverse distribution function* (see, e. g., [1]). This algorithm is based on the relation

$$\xi_0 = F^{-1}(\alpha_0). \tag{1}$$

Here α_0 is the *standard random number* (i. e. sample value of the random variable α which is uniformly distributed on the interval $(0, 1)$). The computation of the sample values α_i of the random variable α are realized with appropriate subroutines (generators from *RAND* or *RANDOM* group).

Usage of algorithm (1), is complicated by the "programming" problem: how to express the inverse function $F^{-1}(v)$ through elementary functions. As this problem is considered for practically important cases of absolutely continuous distributions which are described by the piecewise continuous distribution densities $f(u)$ of the random variables ξ (see, e.g., [1]), arises the question of solvability of the equation

$$\int_a^{\xi_0} f(u)\,du = \alpha_0 \tag{2}$$

with respect to the upper limit of the integral in elementary functions. Difficulties related both to solving the integral on the left side of the relation (2), and to analytical solution of the equation may occur after integration with respect to elementary functions. When expression of the kind of (1) for solution of the equation (2) exists

$$\xi_0 = \psi(\alpha_0) \tag{3}$$

and it is relatively simple for programming, then the density $f(u)$ and formula (3) itself are called *elementary* [1]. As we need the elementary densities in other general algorithms of numerical realization of random variables and vectors (refers to the superposition and rejection methods and special techniques – see, e. g., [1] and further sections 2–4) as well as in numerous applications of the Monte Carlo method (see, e. g., [1–4]), the problem of expansion of the set of such densities arises.

The following ***technology of "recursive substitutions"*** seems to be virtually unlimited on constructing such elementary densities.

Technology 1 [1]. *Let $f_\eta(v)$ be the density of the random variable η, which has an elementary distribution on the interval (c, d): i. e. from the relation of type* (2)

$$\int_c^{\eta_0} f_\eta(v)\,dv = \alpha_0$$

for the corresponding sample value η_0 of the random variable η we can obtain a formula like (3)*: $\eta_0 = \psi_\eta(\alpha_0)$: here $\psi_\eta(w)$ is a simple composition of elementary functions. Let's consider the one-to-one transformation which is defined by a monotone increasing differentiable function $\varphi(x)$, which transforms the interval (a, b) to the interval (c, d); $\varphi(a) = c$, $\varphi(b) = d$. We also assume that the function $\varphi(x)$ and its inverse $\varphi^{-1}(y)$ can be represented as a simple composition of elementary functions. Let the random variable ξ has a distribution density*

$$f(u) = f_\eta(\varphi(u))\,\varphi'(u), \quad u \in (a, b). \tag{4}$$

Under these assumptions, we can assert that $f(u)$ is the density of the elementary distribution, i. e. relation (2) *is solvable with respect to ξ_0 in elementary functions and the formula $\xi_0 = \varphi^{-1}(\psi_\eta(\alpha_0))$ is valid.*

Really, writing equation (2) for the density (4), we obtain

$$\int_a^{\xi_0} f_\eta(\varphi(u))\,\varphi'(u)\,du = \alpha_0, \quad \text{then} \quad \int_{\varphi(a)}^{\varphi(\xi_0)} f_\eta(v)\,dv = \alpha_0,$$

$$\text{then} \quad \varphi(\xi_0) = \psi_\eta(\alpha_0), \quad \text{then} \quad \xi_0 = \varphi^{-1}(\psi_\eta(\alpha_0)). \tag{5}$$

The term "technology of recursive substitutions" for technology 1 relates to the fact that the resulting density (4) can be taken as the initial density of $f_\eta(v)$ and make another one-to-one transformation of type $\varphi(u)$. Using such recursive substitutions one can obtain an unlimited number of new elementary density distributions.

Example 1. Formula

$$\eta_0 = -\frac{\ln \alpha_0}{\lambda}, \tag{6}$$

which corresponds to *the exponential distribution* with the density

$$f_\eta(v) = \lambda\, e^{-\lambda v}, \quad v > 0, \quad \lambda > 0 \tag{7}$$

is widely applied in numerical methods of statistical simulation. Formula (6) is used for construction of Poisson streams, applied in queuing theory, in elementary models of radiative transfer theory, random fields simulation, etc. (see, e. g., [1]).

Let's also consider a random variable ξ, which has the distribution density

$$f(u) = \exp u \times \exp(-\exp u), \quad -\infty < u < +\infty. \tag{8}$$

That's *the extremal* (more precisely, *minimal*) *distribution* density (see, e.g., [5])), describing one of three possible asymptotic distributions of linear combinations of the type $a_n \min\{\eta^{(1)}, \ldots, \eta^{(n)}\} + b_n$ herewith $a_n \neq 0,\ n \to \infty$; here $a_n,\ b_n$ are the numerical sequences and $\{\eta^{(i)}\}$ are the independent identically distributed random variables. Applications of the distribution (8) is associated with the multiple comparisons in the complex decision-making procedures (inter alia, as ranking of mean values).

Function (8) can be obtained from the density (7) according to technology 1 via the transformation $\varphi(x) = \exp x$, which transfers interval $(a, b) = (-\infty, +\infty)$ to interval $(c, d) = (0, +\infty)$. According to (5), simulation formula for the distribution (8) looks like $\xi_0 = \ln(-\ln \alpha_0)$.

This example shows that the use of technology 1 allows to obtain the distribution densities and the corresponding simulation formulas for the various branches of the Probability theory and related applications.

2 The technology of weighed parameter

In numerous applications of numerical statistical simulation (mostly, for realization of Markov chain paths as well as applying the method of dual randomization – see, e. g., [1]) it's required to construct the "modelled" distribution densities $f(u,v)$ of two-dimensional random vectors (ξ,η) with dependent components. Two representations are valid here (see, e. g., [1]):

$$f(u,v)=f_\xi(u)f_\eta(v|u);\ f_\xi(u)=\int f(u,v)\,dv,\ f_\eta(v|u)=\frac{f(u,v)}{f_\xi(u)};\tag{9}$$

$$f(u,v)=f_\eta(v)f_\xi(u|v);\ f_\eta(v)=\int f(u,v)\,du,\ f_\xi(u|v)=\frac{f(u,v)}{f_\eta(v)}.\tag{10}$$

Representation (9) corresponds to the following vector (ξ,η) numerical simulation algorithm: initially, the sample value ξ_0 is realized with respect to the density $f(\xi_0,v)/f_\xi(\xi_0)$, and later the sample value η_0 is simulated with respect to the density $f(\xi_0,v)/f_\xi(\xi_0)$. Similarly, for the representation (10) initially the sample value η_0 with respect to the density of $f_\eta(v)$ is realized, and afterwards the sample value ξ_0 is simulated with respect to the density of $f(u,\eta_0)/f_\eta(\eta_0)$. The formulated algorithms may be far not equivalent in terms of their effective computer implementation.

Example 2. Let's assume that it is necessary to construct an efficient algorithm for simulation of two-dimensional random vector (ξ,η) with the distribution density

$$f(u,v)=\frac{1}{2}\,ve^{-uv},\quad u>0,\ \ 0<v<2.$$

Consider the representation (10):

$$f_\eta(v)=\int_0^{+\infty}\frac{1}{2}\,ve^{-uv}\,du=\frac{1}{2},\quad 0<v<2;\quad f_\xi(u|v)=\frac{f(u,v)}{f_\eta(v)}=ve^{-vu},\quad u>0.$$

Here $f_\eta(v)$ is the density of uniform distribution over the interval $(0,2)$; the corresponding simulation formula is: $\eta_0=2\alpha_1$. Function $f_\xi(u|\eta_0)$ is an exponential distribution density with parameter $\lambda=\eta_0$ (see the relation (7)) and, therefore, $\xi=-(\ln\alpha_2)/\eta_0$ (see formula (6)). Now let's consider the representation (9). Integrating by parts, we obtain

$$f_\xi(u)=\int_0^2\frac{1}{2}\,ve^{-uv}\,dv=\frac{1-(2u+1)e^{-2u}}{2u^2},\quad u>0.$$

The obtained function clearly is not an elementary distribution density and, therefore, for this example, the representation (9) is a certainly the worst (in terms of computer implementation), as compared with the representation (10).

The following ***technology of "weighed parameter"*** provides the examples of densities, for which at least one of the presentations (9) or (10) gives an effective algorithm for numerical realization of sample value (ξ_0,η_0).

Technology 2. *Consider an elementary distribution density $f_\xi(u;\lambda)$, $u \in (a,b)$, depending on parameter λ, allowed values of which belong to the interval (C,D). Elemen-tary of distribution means the existence of simple (elementary) formula $\xi_0 = \psi_\xi(\alpha_1;\lambda)$ for obtaining the sample value of the random variable ξ. Consider also another elemen-tary density $f_\eta(v)$ of random variable η, with values in the interval $(c,d) \subseteq (C,D)$; in this case the appropriate elementary simulation formula $\eta_0 = \psi_\eta(\alpha_2)$ exists. Now let's set the problem of construction an efficient algorithm for realization the sample value (ξ_0,η_0) of two-dimensional random vector (ξ,η) which takes values in the rectangle $G = \{(u,v) :\ a < u < b;\ c < v < d\}$ and has the distribution density*

$$f(u,v) = f_\eta(v) \times f_\xi(u;v), \quad (u,v) \in G. \tag{11}$$

This is a result of a formal product the densities $f_\eta(v)$ and $f_\xi(u;v)$ (here proceeds substitution of variable v instead of parameter λ). In the representation (10) *for the density* (11)*, we obtain $f_\xi(u|v) = f_\xi(u;v)$. For this representation an efficient algorithm exists:*

$$\eta_0 = \psi_\eta(\alpha_1), \quad \xi_0 = \psi_\xi(\alpha_2;\eta_0). \tag{12}$$

Per contra: attempt to construct an effective formulas, like (12)*, for the representation* (9) *of density* (11) *usually fails.*

In particular, example 2 describes application of technology 2. As the initial density with parameter function (7) was taken (here $(C,D) = (0,+\infty)$), and density of the uniform distribution was selected on the subset $(c,d) = (0,2) \subset (C;D)$.

3 Technology of forming the mix

Consider *the method of double randomization* (or, in other words – *integral superposition method* – see, e. g., [1]), in which during realization of random variable ξ an auxiliary random variable η and density

$$f(u) = \int f_\eta(v) f_\xi(u|v)\, dv \tag{13}$$

are introduced and the corresponding representation (10) gives an efficient simulation algorithm for couple (ξ_0,η_0) (see, e. g., [1]). Particular for such method is the case when auxiliary η is taken as a discrete integer value with distribution $\mathbf{P}(\eta = i) = p_i$; $i = 1,2,\ldots$. Here the density (13) has the following form

$$f(u) = \sum_i p_i f_i(u), \quad \text{where} \quad f_i(u) = f_\xi(u|\eta = i), \tag{14}$$

and simulation algorithm (*the discrete superposition method* – see, e. g., [1]) includes the choice of item $\eta_0 = m$ according to *the standard method of realization of a discrete random variable* or any of it's modifications (see, e. g., [1]) together with simulation of value ξ_0 with respect to the density $f_m(u)$. It is possible to obtain a wide set of examples when the discrete superposition

method may be realized effectively, when amount M of numbers i; $i = 1, \ldots, M$ is small enough (in particular, for $M = 2$). Let us describe the appropriate ***technology of "forming the mix"***.

Technology 3. *Let's take two densities of elementary distributions $f_1(u)$ and $f_2(u)$, defined on an interval (a, b), such that a linear combination with positive coefficients*

$$f(u) = p_1\, f_1(u) + p_2\, f_2(u),\ u \in (a, b),\ p_1 > 0,\ p_2 > 0,\ p_1 + p_2 = 1 \tag{15}$$

is not an elementary density. It is possible to obtain such densities $f_1(u)$ and $f_2(u)$, in particular, using the heterogeneous substitutions in technology 1. *For sample values $\xi_0^{(i)}$, realized with respect to the densities $f_i(u)$, simulation formulas $\xi_0^{(i)} = \psi_i(\alpha_0)$, $i = 1, 2$ can be written. For density* (15) *one can construct an efficient discrete superposition algorithm such as: if $\alpha_1 < p_1$, then η_0 – the sample value of the auxiliary integer random value η – is equal to unit, and the sample value ξ_0 of the random variable ξ is realized according to the formula $\xi_0 = \psi_1(\alpha_2)$; otherwise $\xi_0 = \psi_2(\alpha_2)$.*

Example 3 [1]. Suppose we want to construct an algorithm for numerical simulation of the random variable ξ, with distribution density

$$f(u) = \frac{3}{8}(1 + u^2), \quad -1 < u < 1. \tag{16}$$

Relation (16) represents the so-called *Rayleigh molecular scattering of photons in atmo-sphere*, used in radiative transfer theory. Function (16) is not a density of elementary distribution, since the equation $\int_{-1}^{\xi_0} f(u)\, du = \alpha_0$ is reduced to the relation $\xi_0^3 + 3\xi_0 - 8\alpha_0 - 4 = 0$, which prevents obtaining an elementary formula for realization the random variable ξ. The density (16) can be represented as a mix of the (15) type:

$$f(u) = \frac{3}{4} \times \frac{1}{2} + \frac{1}{4} \times \frac{3}{2}\, u^2, \quad -1 < u < 1,$$

i. e. $p_1 = 3/4$, $f_1(u) = 1/2$; $p_2 = 1/4$; $f_2(u) = 3u^2/2$. Function $f_1(u)$ is a density of uniform distribution on the interval $(-1, 1)$, while density $f_2(u)$ is elementary (power). The discrete superposition algorithm here looks like this: *if $\alpha_1 < 3/4$, then $\xi_0 = 2\alpha_2 - 1$; otherwise* $\xi_0 = \sqrt[3]{2\alpha_2 - 1}$.

A generalization of technology 3 may be associated with increased number of terms M in (15) (up to consideration of function series), as well as with the transition to simulation of multivariate random variables $\boldsymbol{\xi}$.

4 Technology of "spoiling" the "modelled" density

Use the special options of *the majorant rejection method*, leads to the further broadening of the set of probabilistic distributions, which admit the efficient numerical realization of sample values. The method consists in the following [1].

Let it be required to realize the numerical sample value $\boldsymbol{\xi}_0$ of the random vector (random variable) $\boldsymbol{\xi}$, which is distributed in $U \in R^d$ with respect to the density $f(\mathbf{u})$, which is proportional to a given nonnegative function $g(\mathbf{u})$, i. e.

$$f(\mathbf{u}) = \frac{g(\mathbf{u})}{\bar{G}}, \quad \bar{G} = \int_U g(\mathbf{u})\, d\mathbf{u}. \tag{17}$$

It is assumed that none of the known standard and special methods does provide an efficient algorithm for realization the value $\boldsymbol{\xi}_0$. The majorant $g^{(1)}(\mathbf{u})$ of function $g(\mathbf{u})$ such that $g(\mathbf{u}) \leq g^{(1)}(\mathbf{u})$ for $\mathbf{u} \in U$ is considered. The first requirement for majorant $g^{(1)}(\mathbf{u})$ is as that for density

$$f^{(1)}(\mathbf{u}) = \frac{g^{(1)}(\mathbf{u})}{\bar{G}^{(1)}}, \quad \bar{G}^{(1)} = \int_U g^{(1)}(\mathbf{u})\, d\mathbf{u}, \tag{18}$$

exists an efficient algorithm (formula) of the form $\boldsymbol{\xi}_0^{(1)} = \psi^{(1)}(\bar{\alpha}_1)$ for realization a sample value $\boldsymbol{\xi}_0^{(1)}$ of random vector $\boldsymbol{\xi}^{(1)}$ (here $\bar{\alpha}_1$ is the corresponding set of standard random numbers).

The majorant rejection method includes the realization of sample value $\boldsymbol{\xi}_0^{(1)}$ with respect to the density (18) and also the value $\eta_0 = \alpha_2 g^{(1)}(\boldsymbol{\xi}_0^{(1)})$. It is easy to get (see, e. g., [1]) that the pair $(\boldsymbol{\xi}_0^{(1)}, \eta_0)$ is uniformly distributed in the "subgraph" $G^{(1)} = \{\mathbf{u} \in U,\ 0 < v < g^{(1)}(\mathbf{u})\}$ of function $g^{(1)}(\mathbf{u})$. If

$$\eta_0 < g(\boldsymbol{\xi}_0^{(1)}), \tag{19}$$

then the realized point $(\boldsymbol{\xi}_0^{(1)}, \eta_0)$ gets into the "subgraph" $G = \{\mathbf{u} \in U,\ 0 < v < g(\mathbf{u})\}$ of function $g(\mathbf{u})$. Since in this case, the pair $(\boldsymbol{\xi}_0^{(1)}, \eta_0)$ is uniformly distributed in the domain G, then for the required sample value $\boldsymbol{\xi}_0$ of vector $\boldsymbol{\xi}$ we take $\boldsymbol{\xi}_0 = \boldsymbol{\xi}_0^{(1)}$. In case when inequality (19) is not true, we draw a pair of $(\boldsymbol{\xi}_0^{(1)}, \eta_0)$ again; then check the inequality (19) etc. It is easy to get that the point $\boldsymbol{\xi}_0$, realized in this way, will be distributed with respect to the density (17).

The average simulation time for the sample value $\boldsymbol{\xi}_0$ is proportional to mathematical expectation of number of those realizable pairs $(\boldsymbol{\xi}_0^{(1)}, \eta_0)$, which is equal to $s = \bar{G}_1/\bar{G}$ (see, e. g., [1]). When s is close to unit, then the majorant rejection method can be treated as effective one.

It is possible to construct examples for effective realization of the rejection method using the following ***technology of "spoiling" the "modelled" density***.

Technology 4. *First, let's construct the "modelled" density $f^{(1)}(\mathbf{u})$, $(\mathbf{u} \in U \subseteq R^d)$ of vector $\boldsymbol{\xi}^{(1)}$, for which exists an efficient algorithm (formula) of numerical realization: $\boldsymbol{\xi}_0^{(1)} = \psi^{(1)}(\bar{\alpha}_1)$ (this algorithm is afterwards used in the first step of the rejection algorithm). To construct the function $f^{(1)}(\mathbf{u})$ the entire set of possibilities (particularly technologies* 1–3*) may be used. Next, we transform the density $f^{(1)}(\mathbf{u})$ to turn it into function $g(\mathbf{u})$, which is proportional to the "non-modelled" density $f(\mathbf{u})$ (actually, we "¡spoil"¿ the "modelled" density $f^{(1)}(\mathbf{u})$). One of the simplest transformations is to multiply the density $f^{(1)}(\mathbf{u})$ by the low-varying function $Y(\mathbf{u})$:*

$$g(\mathbf{u}) = f^{(1)}(\mathbf{u}) \times Y(\mathbf{u}), \quad \mathbf{u} \in U; \quad \textit{where } 0 < A \leq Y(\mathbf{u}) \leq B \tag{20}$$

and $(B - A)$ *is close to zero positive value. Then we can take as a majorant* $g^{(1)}(\mathbf{u}) = B\, f^{(1)}(\mathbf{u})$. *The density which is proportional to such a function is obviously equal to* $f^{(1)}(\mathbf{u})$. *By integrating the non-negative functions* $g^{(1)}(\mathbf{u})$ *and* $g(\mathbf{u})$ *over the domain* U *according to the relation* $Af^{(1)}(\mathbf{u}) = Ag^{(1)}(\mathbf{u})/B \le g(\mathbf{u})$, *we obtain* $A\,\bar{G}^{(1)}/B \le \bar{G}$. *Then* $s \le B/A$, *i. e. for* $A \approx B$ *the value of* s *is small (close to unit), and the appropriate rejection algorithm can be regarded as efficient (economical).*

For the convenience of the calculations in (20), instead of density $f^{(1)}(\mathbf{u})$ we can consider the proportional function $\tilde{g}^{(1)}(\mathbf{u})$ (for example, by omitting the normalizing constant).

Example 4. Suppose it is required to construct an algorithm for simulation the random variable ξ, which has distribution density $f(u)$, proportional to the function

$$g(u) = \left(2 + \frac{\arcsin u}{5\pi}\right) u^3, \quad 0 < u < 1.$$

It is easy to verify that the density $f(u)$ is not elementary. Note that $g(u) = Y(u) \times \tilde{g}^{(1)}(u)$, moreover, due to monotonicity of the function $\arcsin u$ on $(0, 1)$, the inequality $2 < Y(u) < 2.1$ is valid. Thus, $g(u) < g^{(1)}(u) = 2.1\, u^3$. The density which is proportional to the majorant $g^{(1)}(u)$, is equal to $f^{(1)}(u) = 4u^3$, $0 < u < 1$; the appropriate simulation formula is: $\xi_0^{(1)} = \sqrt[4]{\alpha_0}$ [1]. Algorithm for the rejection method comprises the following steps.

1. Simulate the sample value $\xi_0^{(1)}$ *according to formula* $\xi_0^{(1)} = \sqrt[4]{\alpha_1}$, *and also the value* $\eta_0 = \alpha_2\, g^{(1)}(\xi_0^{(1)}) = 2.1\, \alpha_2\, (\xi_0^{(1)})^3$.

2. Check the inequality $\eta_0 < g(\xi_0^{(1)})$ *or*

$$10.5\, \pi\, \alpha_2 < 10\pi + \arcsin \sqrt[4]{\alpha_1}. \tag{21}$$

If the inequality is satisfied, we take $\xi_0 = \xi_0^{(1)}$ *as sample value* ξ_0 *of the random variable* ξ. *If the inequality* (21) *is not true, then repeat step* 1, *and so on, etc.*

The cost s (i. e. the average number of attempts to draw couples $(\xi_0^{(1)}, \eta_0)$ till fits the inequality (21)) has the following upper bound: $1 < s < 2.1/2 = 1.05$.

Conclusions

In this paper we have stated the technologies for constructing the distribution probabilis-tic densities, which admit the efficient numerical realization of sample values: the technology of recursive substitutions (to implement the method of inverse distribution function); the technology of weighed parameter (to implement the method of simulation a two-dimensional random vector with dependent components); the technology of forming the mix (to implement the discrete superposition method); the technology of "spoiling" the "modelled" density (to implement the majorant rejection method). This work was sponsored by the Russian Foundation for Basic Research (projects No. 10–01–00040, 09–01–00035), and partially supported by Israeli Research budget of the Zefat Academic College (Northern branch of Bar-Ilan University).

References

[1] Mikhailov G. A., Voytishek A. V. *Numerical Statistical Simulation. Monte Carlo Methods.* Moscow: "Academy", 2006 (in Russian).

[2] Ivanoff B. G., Merzbach E. What is a multi-parameter renewal process // *Stochastics.* 2006. V. 78, No. 6. P. 411–441.

[3] Rissanen J. *Lectures on Statistical Modeling Theory.* Helsinki Institute for Information Technology, 2007.

[4] Davies K., Braun W. J. Modelling point patterns with the Ivanoff-Merzbach renewal process // *Environmetrics.* 2010. V. 21. P. 85–101.

[5] David H. A. *Order Statistics*. New York: Wiley, 1983.

Sensitivity of a Diffusion Process to the Moving Boundary Parameters *

Sergey A. Gusev
Institute of Computational Mathematics and Mathematical Geophysics SB RAS
Novosibirsk, Russia
e-mail: sag@osmf.sscc.ru

Abstract

A 2D parabolic boundary problem with moving boundary is considered in the paper. The moving part of the boundary is approximated by a broken line. A statistical modeling method for estimation of the solution of this problem and its parametric derivatives is proposed. Desired estimates are obtained as results of numerical simulation of trajectories of the corresponding to the problem diffusion process and its derivatives with respect to parameters determining the boundary motion. We set a biunique correspondence at any time point between the moving boundary domain and a fixed domain which coincides with the initial state of the moving boundary domain. In calculations the numerical simulation of the diffusion process is performed in the fixed domain.

Keywords: moving boundary problem, stochastic differential equations, statistical modeling, Euler method .

Introduction

A possibility of applying statistical modeling diffusion processes to estimating solutions of elliptic and parabolic types problems is well known.

In this paper we consider a 2D moving boundary problem

$$\partial u/\partial t + Lu = 0 \ , \qquad (t,x,y) \in Q_T \equiv (0,T) \times G(t), \tag{1}$$

$$u(T,x,y) = \varphi(x,y,\theta), \tag{2}$$

$$u(t,x,y) = 0, \qquad (x,y) \in \Gamma(t,\theta), \tag{3}$$

where $L \equiv b_{11}(t,x,y)\frac{\partial^2}{\partial x^2} + 2b_{12}(t,x,y)\frac{\partial^2}{\partial x \partial y} + b_{22}(t,x,y)\frac{\partial^2}{\partial y^2} + a_1(t,x,y)\frac{\partial}{\partial x} + a_2(t,x,y)\frac{\partial}{\partial y}$; $B = (b_{ij})_{ij=1,2}$ is a symmetric positively definite matrix; $G(t)$ is a time dependent 2D bounded domain having a moving boundary $\Gamma(t)$.

We introduce a diffusion process $(X_\cdot, Y_\cdot)$, which is defined by the two following stochastic differential equations:

$$\begin{aligned} X_s &= x + \int_t^s a_1(v, X_v, Y_v)dv + \int_t^s \sum_{i=1}^{2} \sigma_{1i}(v, X_v, Y_v)dW_{iv} \ , \\ Y_s &= y + \int_t^s a_2(v, X_v, Y_v)dv + \int_t^s \sum_{i=1}^{2} \sigma_{2i}(v, X_v, Y_v)dW_{iv} \ , \end{aligned} \tag{4}$$

*The work was supported by RFBR, grant № 11-01-00252-a.

where σ is 2×2 matrix such that $\sigma\sigma^* = 2B$; $W_i.,\ \ i = 1, 2$ are independent Wiener processes.

Let us denote by $E_{t,x,y}$ the conditional expectation of a functional of $(X., Y.)$ on the condition that the process starts from a point (x, y) at a time point t; $\tau = \inf\{t : (t, X_t, Y_t) \notin Q_T\}$.

Then a solution of the moving boundary problem (1) — (3) at a point $(t, x, y) \in Q_T$ coincides with the mathematical expectation of a functional of the diffusion process (4) in the form (see, for example, [1])

$$u(t, x, y, \theta) = E_{t,x,y}\big[\varphi(X_T, Y_T, \theta)\chi_{\tau>T}\big]. \tag{5}$$

We propose a statistical modeling method of estimating solution of the problem (1) — (3) at some given point in Q_T and its sensitivities to parameters, which define the motion of the boundary.

In this work the moving boundary $\Gamma(t)$ is approximated by a broken line. The corresponding domain is denominated by $G_m(t)$, and we define a fixed domain as $G_c = G_m(t_0)$ that coincides with the start state of G_m. The motion of boundary vertexes of the broken line is determined by some set of parameters. We constract a triangle net in a neighborhood of the broken line, so that legs of the broken line are sides of the triangles.

Then a bijective mapping between $G_m(t)$ and G_c at any point in time is established on the base of triangle nets in $G_m(t)$ and G_c.

Further we define a stochastic process $(\bar{X}., \bar{Y}.)$ in G_c that one-to-one corresponds to the original diffusion process in $G_m(t)$. This construction allows us to realize modeling the trajectories of a random process in the fixed domain G_c instead of G_m.

The requierred estimates of the solution of the moving boundary problem and the sensitivities are obtained as mathematical expectations of the corresponding functional of $(\bar{X}., \bar{Y}.)$ and its parametric derivatives that are modeled in a fixed domain G_c.

1 A Biunique Correspondence between Triangles

Let us establish a biunique correspondence between triangles in G_m and G_c. Let T be a triangle with vertexes (x_1, y_1), (x_2, y_2), (x_3, y_3) in G_m, and $\bar{T}$ be a triangle in G_c with vertexes $(\bar{x}_1, \bar{y}_1)$, $(\bar{x}_2, \bar{y}_2)$, $(\bar{x}_3, \bar{y}_3)$.

We remind that coordinates of boundary vertexes of T depend on t and θ.

Let us define a plane in $\mathbf{R}^3$ passing through the following three points $X_1 = (\bar{x}_1, \bar{y}_1, x_1)$, $X_2 = (\bar{x}_2, \bar{y}_2, x_2)$, $X_3 = (\bar{x}_3, \bar{y}_3, x_3)$. We denote an equation of the plane as follows

$$A_x\bar{x} + B_x\bar{y} + C_x x + D_x = 0\ . \tag{6}$$

Values of the coefficients A_x, B_x, C_x can be obtained by vector product of two not collinear vectors lying in this plane, for example, $X_2 - X_1\ \ X_3 - X_1$. Value of D_x can be defined as result

of substitution X_1 in the equation (6). So we have

$$A_x = (\bar{y}_2 - \bar{y}_1)(x_3 - x_1) - (\bar{y}_3 - \bar{y}_1)(x_2 - x_1) \ , \tag{7}$$
$$B_x = (\bar{x}_2 - \bar{x}_1)(x_3 - x_1) - (\bar{x}_3 - \bar{x}_1)(x_2 - x_1) \ , \tag{8}$$
$$C_x = (\bar{x}_2 - \bar{x}_1)(\bar{y}_3 - \bar{y}_1) - (\bar{x}_3 - \bar{x}_1)(\bar{y}_2 - \bar{y}_1) \ , \tag{9}$$
$$D_x = -(A_x\bar{x}_1 + B_x\bar{y}_1 + C_x x_1). \tag{10}$$

Let us similarly define a plane

$$A_y\bar{x} + B_y\bar{y} + C_y y + D_y = 0 \ , \tag{11}$$

passing through three points $Y_1 = (\bar{x}_1, \bar{y}_1, y_1)$, $Y_2 = (\bar{x}_2, \bar{y}_2, y_2)$, $Y_3 = (\bar{x}_3, \bar{y}_3, y_3)$.

And coefficients of the plane (11) are written in the form

$$A_y = (\bar{y}_2 - \bar{y}_1)(y_3 - y_1) - (\bar{y}_3 - \bar{y}_1)(y_2 - y_1) \ , \tag{12}$$
$$B_y = (\bar{x}_3 - \bar{x}_1)(y_2 - y_1) - (\bar{x}_2 - \bar{x}_1)(y_3 - y_1) \ , \tag{13}$$
$$C_y = (\bar{x}_2 - \bar{x}_1)(\bar{y}_3 - \bar{y}_1) - (\bar{x}_3 - \bar{x}_1)(\bar{y}_2 - \bar{y}_1) \ , \tag{14}$$
$$D_y = -(A_y\bar{x}_1 + B_y\bar{y}_1 + C_y y_1). \tag{15}$$

Note that $C_x = C_y$, therefore we will write further this coefficient without any index. Let us derive x, y from the equations (6), (11)

$$x = -\frac{1}{C}(A_x(t,\theta)\bar{x} + B_x(t,\theta)\bar{y} + D_x(t,\theta)), \tag{16}$$
$$y = -\frac{1}{C}(A_y(t,\theta)\bar{x} + B_y(t,\theta)\bar{y} + D_y(t,\theta)). \tag{17}$$

Equations (16, (17) establish a biunique correspondence between points $(\bar{x}, \bar{y})$ of the fixed triangle $\bar{T}$ in G_c and points (x, y) of the moving triangle T in G_m under the condition

$$A_x B_y - A_y B_x \neq 0. \tag{18}$$

The condition (18) corresponds to noncollinearity of vectors directed along sides of T and $\bar{T}$.

In the following we will mark denominations of points and triangles in G_c by bar above.

Let us denote by f_x, f_y right-hand sides of the equations (16), (17) respectively. These functions establish an one-to-one correspondence between points of the fixed triangle net in G_c and points of moving triangle net in G_m. Therefore, under the condition (18), we can express from (16, (17) coordinates of points of a fixed triangle $\bar{T} \subseteq G_c$ through coordinates of the corresponding points of moving triangle $T \subseteq G_m$

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \frac{C}{\Delta} \begin{pmatrix} -B_y & B_x \\ A_y & -A_x \end{pmatrix} \begin{pmatrix} x + D_x/C \\ y + D_y/C \end{pmatrix}, \tag{19}$$

where $\Delta = A_x B_y - A_y B_x$. The equation (19) gives us inverse functions to f_x, f_y

$$f_x^{-1}(t,x,y) = \frac{C}{\Delta}\left[-B_y(x + D_x/C) + B_x(y + D_y/C)\right], \tag{20}$$

$$f_y^{-1}(t,x,y) = \frac{C}{\Delta}\left[A_y(x + D_x/C) - A_x(y + D_y/C)\right]. \tag{21}$$

Let k_T be a number of triangles in the net. We denote by T_i, $\bar{T}_i$, $i = 1, \ldots, k_T$ triangles of the moving net in G_m and the fixed net in G_c correspondingly. And plains coefficients corresponding to triangles T_i, $\bar{T}_i$, $i = 1, \ldots, k_T$ will be marked by upper index i: A_x^i, B_x^i, D_x^i, A_y^i, B_y^i, D_y^i, C^i. We assume that the condition (18) holds for all coefficients A_x^i, B_x^i, A_y^i, B_y^i.

We define a biunivocal mapping $F : G_c \to G_m$ by the equality

$$(x,y) = \begin{cases} (f_x, f_y), & (\bar{x},\bar{y}) \in \bigcup\limits_{i=1}^{k_T} \bar{T}_i, \\ (\bar{x},\bar{y}), & (\bar{x},\bar{y}) \notin \bigcup\limits_{i=1}^{k_T} \bar{T}_i. \end{cases} \tag{22}$$

2 Stochastic Process in the Fixed Domain

We define a stochastic process in G_c which corresponds to the diffusion process (4) in G_m via bijection (22).

Outside of the triangle net processes in G_m and G_c must be coinside. Now we define the corresponding process on each triangle in G_c by determining the corresponding grift and diffusion coefficients.

Let $(\bar{x},\bar{y}) \in \bar{T}_i$ be some point in G_c. The point $(\bar{x},\bar{y})$ corresponds to the following point in G_m at the mapping (22)

$$(x,y) = (f_x(t,\bar{x},\bar{y},\theta), f_y(t,\bar{x},\bar{y},\theta)). \tag{23}$$

Let us study a transformation of the equation (1) if it is considered as equation on G_c at substitution instead of x, y functions f_x, f_y respectively. For this purpose we define on triangles of G_c functions $\bar{u}$, $\bar{a}_i$, $\bar{b}_{ij}$, $\bar{\sigma}_{ij}$ by the following equations

$$\bar{u}(t,\bar{x},\bar{y},\theta) = u(t, f_x(t,\bar{x},\bar{y},\theta), f_y(t,\bar{x},\bar{y},\theta)), \tag{24}$$

$$\bar{a}_i(t,\bar{x},\bar{y},\theta) = a_i(t, f_x(t,\bar{x},\bar{y},\theta), f_y(t,\bar{x},\bar{y},\theta)), \tag{25}$$

$$\bar{b}_{ij}(t,\bar{x},\bar{y},\theta) = b_{ij}(t, f_x(t,\bar{x},\bar{y},\theta), f_y(t,\bar{x},\bar{y},\theta)), \tag{26}$$

$$\bar{\sigma}_{ij}(t,\bar{x},\bar{y},\theta) = \sigma_{ij}(t, f_x(t,\bar{x},\bar{y},\theta), f_y(t,\bar{x},\bar{y},\theta)). \tag{27}$$

We express all derivatives of u in the equation (1) via derivatives of function $\bar{u}$ with respect to t, $\bar{x}$ and $\bar{y}$. After substitution obtained expressions in (1) we derive a parabolic equation on G_c. Then we use this equation to defining drift and diffusion coefficients of the stochastic process in the domain G_c.

After differentiating (24) with respect to t we obtain

$$\frac{\partial u}{\partial t} = \frac{\partial \bar{u}}{\partial t} - \frac{\partial u}{\partial x}\frac{\partial f_x}{\partial t} - \frac{\partial u}{\partial y}\frac{\partial f_y}{\partial t}. \tag{28}$$

Let us differentiate (24) with respect to $\bar{x}$, $\bar{y}$ with provision (16), (17). Then we derive expressions of the first derivatives of $\bar{u}$ with respect to $\bar{x}$, $\bar{y}$ in the domain G_c trough the first derivatives of the function u with respect to x, y in G_m

$$\begin{pmatrix} \frac{\partial \bar{u}}{\partial \bar{x}} \\ \frac{\partial \bar{u}}{\partial \bar{y}} \end{pmatrix} = -\frac{1}{C^i}\begin{pmatrix} A_x^i & A_y^i \\ B_x^i & B_y^i \end{pmatrix}\begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix}. \tag{29}$$

By solving system (29) with respect to $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$ we derive

$$\begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix} = \frac{1}{\Delta^i}\begin{pmatrix} -B_y^i & A_y^i \\ B_x^i & -A_x^i \end{pmatrix}\begin{pmatrix} \frac{\partial \bar{u}}{\partial \bar{x}} \\ \frac{\partial \bar{u}}{\partial \bar{y}} \end{pmatrix}, \tag{30}$$

where $\Delta^i = (A_x^i B_y^i - A_y^i B_x^i)/C^i$.

Substituting expressions of derivatives of u (28), (30) in the equation (1), we receive a vector of coefficients at $\partial\bar{u}/\partial\bar{x}$, $\partial\bar{u}/\partial\bar{y}$ (drift coefficients) for the triangle $\bar{T}_i$

$$\begin{pmatrix} \bar{a}_1^i \\ \bar{a}_2^i \end{pmatrix} = \frac{1}{\Delta^i}\begin{pmatrix} -B_y^i & A_y^i \\ B_x^i & -A_x^i \end{pmatrix}\begin{pmatrix} a_1 - \frac{\partial f_x}{\partial t} \\ a_2 - \frac{\partial f_y}{\partial t} \end{pmatrix}. \tag{31}$$

Differentiating the equation (29) with respect to $\bar{x}$, $\bar{y}$ we derive expressions of the the second derivatives of $\bar{u}$ with respect to $\bar{x}$, $\bar{y}$ in G_c trough the second derivatives of u with respect to x, y in G_m

$$\begin{pmatrix} \frac{\partial^2 \bar{u}}{\partial \bar{x}^2} \\ \frac{\partial^2 \bar{u}}{\partial \bar{x}\partial \bar{y}} \\ \frac{\partial^2 \bar{u}}{\partial \bar{y}^2} \end{pmatrix} = \frac{1}{{C^i}^2}\begin{pmatrix} {A_x^i}^2 & 2A_x^iA_y^i & {A_y^i}^2 \\ A_x^iB_x^i & A_x^iB_y^i + A_y^iB_x^i & A_y^iB_y^i \\ {B_x^i}^2 & 2B_x^iB_y^i & {B_y^i}^2 \end{pmatrix}\begin{pmatrix} \frac{\partial^2 u}{\partial x^2} \\ \frac{\partial^2 u}{\partial x\partial y} \\ \frac{\partial^2 u}{\partial y^2} \end{pmatrix}. \tag{32}$$

Then we determine the second derivatives of u from the linear equations system (32)

$$\begin{pmatrix} \frac{\partial^2 u}{\partial x^2} \\ \frac{\partial^2 u}{\partial x\partial y} \\ \frac{\partial^2 u}{\partial y^2} \end{pmatrix} = \frac{1}{(\Delta^i)^2}\begin{pmatrix} {B_y^i}^2 & -2A_y^iB_y^i & {A_y^i}^2 \\ -B_x^iB_y^i & A_x^iB_y^i + A_y^iB_x^i & -A_x^iA_y^i \\ {B_x^i}^2 & -2A_x^iB_x^i & {A_x^i}^2 \end{pmatrix}\begin{pmatrix} \frac{\partial^2 \bar{u}}{\partial \bar{x}^2} \\ \frac{\partial^2 \bar{u}}{\partial \bar{x}\partial \bar{y}} \\ \frac{\partial^2 \bar{u}}{\partial \bar{y}^2} \end{pmatrix}. \tag{33}$$

Let us denote by E_{kl}^i $(k,l = 1,3)$ elements of the matrix of the system equations (33). Then on the base of (33) we obtain elements of the second derivatives matrix of $\bar{u}$ in the parabolic equation in the domain G_c that corresponds to the parabolic equation (1) in G_m

$$\begin{aligned} b_{11}^i &= \left(b_{11}E_{11}^i + 2b_{12}E_{21}^i + b_{22}E_{31}^i\right)/2 \ , \\ b_{12}^i &= b_{21}^i = \left(b_{11}E_{12}^i + 2b_{12}E_{22}^i + b_{22}E_{32}^i\right)/2 \ , \\ b_{22}^i &= \left(b_{11}E_{13}^i + 2b_{12}E_{23}^i + b_{22}E_{33}^i\right)/2 \ . \end{aligned} \tag{34}$$

Then it is necessary to verify positive definiteness of the matrix (b_{kl}^i) $(k,l = 1,2)$. The following theorem is valid.

Theorem 1. *Under the condition (18) the matrix* (b^i_{kl}) $(k,l=1,2)$ *is positive definite.*

The matrix (b^i_{kl}) defined in (34) can be represented in terms of matrix product

$$\begin{pmatrix} b^i_{11}b^i_{12} \\ b^i_{12}b^i_{22} \end{pmatrix} = \frac{1}{{\Delta^i}^2} \begin{pmatrix} -B^i_y & B^i_x \\ A^i_y & -A^i_x \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{pmatrix} \begin{pmatrix} -B^i_y & A^i_y \\ B^i_x & -A^i_x \end{pmatrix}. \tag{35}$$

We can obtain from (35) the cooresponding to the triangle $\bar{T}_i$ the diffusion matrix for SDE system in the domain G_c

$$\sigma^i = \frac{1}{\Delta^i} \begin{pmatrix} -B^i_y & B^i_x \\ A^i_y & -A^i_x \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Then from the equations (31), (36) we derive that the stochastic process in the triangle $\bar{T}_i \in G_c$ satisfies to the following vector equation

$$\begin{pmatrix} d\bar{X}^i \\ d\bar{Y}^i \end{pmatrix} = \frac{1}{\Delta^i} \begin{pmatrix} -B^i_y & A^i_y \\ B^i_x & -A^i_x \end{pmatrix} \left[\begin{pmatrix} a_1 - \frac{\partial f_x}{\partial t} \\ a_2 - \frac{\partial f_y}{\partial t} \end{pmatrix} dt + \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} dW_{1t} \\ dW_{2t} \end{pmatrix} \right]. \tag{36}$$

We obtain the following SDE system describing the part of a trajectory of the stochastic process belonging to the triangle $\bar{T}_i \in G_c$ by substitution in coefficients a_i, b_{ij} in (36) instead of $X.$, $Y.$ right-hand sides of the equations (16), (17).

$$\begin{pmatrix} d\bar{X}^i \\ d\bar{Y}^i \end{pmatrix} = \frac{1}{\Delta^i} \begin{pmatrix} -B^i_y & A^i_y \\ B^i_x & -A^i_x \end{pmatrix} \left[\begin{pmatrix} \bar{a}_1 - \frac{\partial f_x}{\partial t} \\ \bar{a}_2 - \frac{\partial f_y}{\partial t} \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}_{11} & \bar{\sigma}_{12} \\ \bar{\sigma}_{21} & \bar{\sigma}_{22} \end{pmatrix} \begin{pmatrix} dW_{1t} \\ dW_{2t} \end{pmatrix} \right]. \tag{37}$$

The equation (36) gives us a relation between processes $(\bar{X}^i, \bar{Y}^i)$ in G_c and $(X., Y.)$ in G_m on each pair corresponding to triangles $\bar{T}_i$ and T_i.

$$\begin{pmatrix} dX \\ dY \end{pmatrix} = -\frac{1}{C^i} \begin{pmatrix} A^i_x & B^i_x \\ A^i_y & B^i_y \end{pmatrix} \begin{pmatrix} d\bar{X}^i \\ d\bar{Y}^i \end{pmatrix} + \begin{pmatrix} \frac{\partial f_x(t,\bar{X}^i_t,\bar{Y}^i_t,\theta)}{\partial t} \\ \frac{\partial f_y(t,\bar{X}^i_t,\bar{Y}^i_t,\theta)}{\partial t} \end{pmatrix} dt. \tag{38}$$

Hence we can numerically reproduce trajectories of the process $(X., Y.)$ in G_m step by step after modeling corresponding $(\bar{X}., \bar{Y}.)$ trajectories in G_c.

3 Numerical Algorithm

We use the Euler method for modeling trajectories $(X., Y.)$ and $(\bar{X}., \bar{Y}.)$. In the interior of the i-th triangle we use the following transition formulas at a step with number k:

$$\begin{pmatrix} \Delta\bar{X}^i \\ \Delta\bar{Y}^i \end{pmatrix}_k = \frac{1}{\Delta^i_k} \begin{pmatrix} -B^i_y & A^i_y \\ B^i_x & -A^i_x \end{pmatrix}_k \left[h \begin{pmatrix} \bar{a}_1 - \frac{\partial f_x}{\partial t} \\ \bar{a}_2 - \frac{\partial f_y}{\partial t} \end{pmatrix}_k + \sqrt{h} \begin{pmatrix} \bar{\sigma}_{11} & \bar{\sigma}_{12} \\ \bar{\sigma}_{21} & \bar{\sigma}_{22} \end{pmatrix}_k \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}_k \right], \tag{39}$$

$$\begin{pmatrix} \bar{X}^i \\ \bar{Y}^i \end{pmatrix}_{k+1} = \begin{pmatrix} \bar{X}^i \\ \bar{Y}^i \end{pmatrix}_k + \begin{pmatrix} \Delta\bar{X}^i \\ \Delta\bar{Y}^i \end{pmatrix}_k, \tag{40}$$

where h is an integration step; ξ_l $(l = 1, 2)$ are independent standard normal random variables.

If $(\bar{X}^i, \bar{Y}^i)_k$ is outside of $\bar{T}_i$, then we calculate $(\bar{X}^i, \bar{Y}^i)_{k+1}$ in another way. Firstly we calculate

$$\begin{pmatrix} X_k^i \\ Y_k^i \end{pmatrix} = \begin{pmatrix} f_x\left(t_k, \bar{X}_k^i, \bar{Y}_k^i\right) \\ f_y\left(t_k, \bar{X}_k^i, \bar{Y}_k^i\right) \end{pmatrix} - \frac{1}{C^i} \begin{pmatrix} A_x^i & B_x^i \\ A_y^i & B_y^i \end{pmatrix}_k \begin{pmatrix} \Delta\bar{X}^i \\ \Delta\bar{Y}^i \end{pmatrix}_k + h \begin{pmatrix} \frac{\partial f_x(t_k, \bar{X}_k^i, \bar{Y}_k^i)}{\partial t} \\ \frac{\partial f_y(t_k, \bar{X}_k^i, \bar{Y}_k^i)}{\partial t} \end{pmatrix}.$$

Then we determine a triangle in G_m to which (X_k^i, Y_k^i) belongs. Let a number of this triangle be i_1. After that we calculate

$$\begin{pmatrix} \bar{X}^{i_1} \\ \bar{Y}^{i_1} \end{pmatrix}_{k+1} = \begin{pmatrix} f_x^{-1}(t_k, X_k^i, Y_k^i) \\ f_y^{-1}(t_k, X_k^i, Y_k^i) \end{pmatrix}.$$

Numerical modeling trajectories outside the triangle net we do by the Euler method as usual.

4 Differentiation with respect to Parameters

Using representation (f_x, f_y) of the process $(X_\cdot, Y_\cdot)$ in G_m through the process $(\bar{X}_\cdot, \bar{Y}_\cdot)$ in G_c we can reformulate the expression of $u(t, x, y)$ in (5) as follows

$$\begin{aligned} u(t, x, y, \theta) = E_{t,\bar{x},\bar{y}}\big[&\varphi(f_x(T, \bar{X}_T, \bar{Y}_T), f_y(T, \bar{X}_T, \bar{Y}_T), \theta)\chi_{(\tau>T)\&((\bar{X}_T,\bar{Y}_T)\in\cup\bar{T}_i)} \\ &+\varphi(\bar{X}_T, \bar{Y}_T), \theta)\chi_{(\tau>T)\&((\bar{X}_T,\bar{Y}_T)\notin\cup\bar{T}_i)}\big], \end{aligned} \tag{41}$$

where $\bar{x}$, $\bar{y}$ are coordinates of the point in G_c that corresponds to $(x, y) \in G_m$.

To obtain statistical estimation u_θ we derive its probability representation in the form

$$\begin{aligned} u_\theta(t, x, y, \theta) = E_{t,\bar{x},\bar{y}} \Big[&\Big(\tfrac{\partial\varphi}{\partial x}\left(\tfrac{\partial f_x}{\partial x}\left(\bar{X}_\theta\right)_T + \tfrac{\partial f_x}{\partial y}\left(\bar{Y}_\theta\right)_T + \tfrac{\partial f_x}{\partial\theta}\right) \\ &+\tfrac{\partial\varphi}{\partial y}\left(\tfrac{\partial f_y}{\partial x}\left(\bar{X}_\theta\right)_T + \tfrac{\partial f_y}{\partial y}\left(\bar{Y}_\theta\right)_T + \tfrac{\partial f_y}{\partial\theta}\right) + \left(\tfrac{\partial\varphi}{\partial\theta}\right)_T\Big)\chi_{(\tau>T)\&\left((\bar{X}_T,\bar{Y}_T)\in\cup\bar{T}_i\right)} \\ &+\Big(\left(\tfrac{\partial\varphi}{\partial x}\right)\left(\bar{X}_\theta\right)_T + \left(\tfrac{\partial\varphi}{\partial y}\right)\left(\bar{Y}_\theta\right)_T + \left(\tfrac{\partial\varphi}{\partial\theta}\right)_T\Big)\chi_{(\tau>T)\&\left((\bar{X}_T,\bar{Y}_T)\notin\cup\bar{T}_i\right)}\Big]. \end{aligned}$$

5 Examination of the Method

We applied the proposed method to estimation of a solution and its parametric derivative of 2D melting problem [2]. It is an ice melting problem for a circle of radius r_2 with heating clement at the centre. Phase transition is on distance $r_1(t) = \theta\sqrt{t}$ apart the centre. Heat exchange in the solid phase is described by the following boundary problem

$$\frac{\partial u}{\partial t} = K\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \quad \sqrt{x^2+y^2} \in (r_1(t), r_2), \ \ t \in [t_0, T], \tag{42}$$

$$u(t, x, y) = 0, \qquad \sqrt{x^2+y^2} = r_1(t), \tag{43}$$

$$u(t, x, y) = -\Theta, \qquad \sqrt{x^2+y^2} = r_2, \tag{44}$$

$$u(t_0, x, y) = u_0(x, y). \tag{45}$$

The exact solution of (47) – (50) is the function

$$u(t,x,y) = \Theta\left[\frac{Ei\left(-(x^2+y^2)/4Kt\right)}{Ei\left(-\theta^2/4K\right)} - 1\right], \quad x^2+y^2 > \theta^2 t, \tag{46}$$

where $Ei(z) \equiv \int\limits_{\infty}^{-z} \frac{e^{-s}}{s} ds$ is the exponential integral.

We did a change of the time variable $s = T + t_0 - t$ to obtain the problem with backward time. Then we defined a function w by the equation $w(s,x,y) = u(T+t_0-t,x,y)$ and derived the following boundary problem

$$\frac{\partial w}{\partial s} + K\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right), \quad \sqrt{x^2+y^2} \in (r_1(T+t_0-s), r_2), \;\; s \in [t_0, T], \tag{47}$$

$$w(s,x,y) = 0, \qquad \sqrt{x^2+y^2} = r_1(T+t_0-s), \tag{48}$$

$$w(s,x,y) = -\Theta, \qquad \sqrt{x^2+y^2} = r_2, \tag{49}$$

$$w(T,x,y) = u_0(x,y). \tag{50}$$

In calculations values of numerical parameters was taken as follows: $\theta = 0.06637$cm/c, $\Theta = 2^\circ$, $r_2 = 4$, $t_0 = 9$, $T = 20$.

We constructed a triangle net near the moving boundary by selecting r_m a radius of a middle circle such that $r_1(t) < r_m < r_2$. The ring formed by circles of radiuses $r_1(t)$ and r_m was uniformly split into specified number of sectors. In each sector three triangles were defined as it is shown in the fig.1.

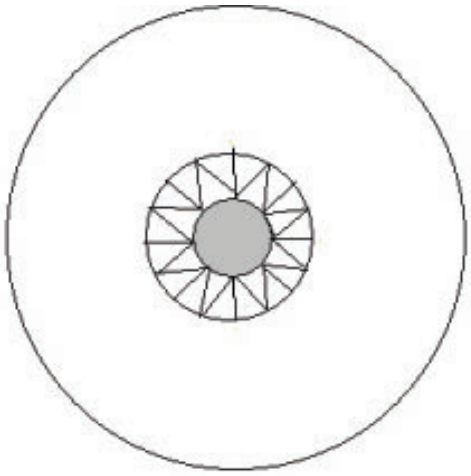


Figure 1: A triangle net in the ice melting problem.

In the fig.1 the ring is uniformly split into 8 sectors. We applied our method to obtain estimations of u and u_θ in different points. A good coincidence with the exact solution was obtained in the experiment.

References

[1] Gikhman I.I., Skorohod A.V. (1977). *Introduction to the theory of random processes*. Nauka, Moskow.

[2] S. Paterson Propogation of a boundary of fusion. (1952) *Proceedings of Glasgow Mathematical Association.* Vol. I, Part I, pp. 42-47.

Optimization of a Time-Sharing Queueing Process in Random Environment with Means of Computer Simulation

Andrei V. Zorine
N.I. Lobachevsky State University of Nizhni Novgorod – National Research University
Nizhni Novgorod, Russia
e-mail: zoav1602@gmail.com

Abstract

A service of conflict flows with time-sharing algorithm with readjustments in random environment is considered. A mathematical model is constructed as a homogeneous denumerable discrete-time Markov chain. Conditions for the stationary distribution existence are found. A computer simulation model is also built. With means of computer simulation a switching function can be found which minimizes several cost-type objective functionals.

Keywords: Conflict flows, time-sharing algorithm with readjustments, optimal switching.

Introduction

Time-sharing algorithm for service of conflict flows is used in many real controlling systems such as computers, road traffic control systems, micro-welding machines, customs inspection in large airports. One has to take into account variability of probabilistic characteristics of input flows to construct a mathematical model of a queueing system governed by time-sharing algorithm. The investigator has to calculate many important performance measures, namely, loading of the processor, mean sojourn time of an arbitrary customer in the queueing system, mean sojourn cost, etc. Moreover it is important to find the optimal switching rule for the server between the queues of waiting customers. No ultimate analytical solution is found for the whole complex of problems [1], thus it is the computer simulation of the time-sharing service process that can give answers to the investigator's problems. Also, computer simulation of queueing systems allows investigation of queueing systems with diverse input flows of customers both simulated and once observed.

1 Mathematical model of service process in the class of time-sharing algorithms with readjustments

A queueing system with $m < \infty$ conflict input flows $\Pi_1, \Pi_2, \ldots, \Pi_m$ is considered. The input flows are formed in a random external environment with $d < 5$ states $e^{(1)}, e^{(2)}, \ldots, e^{(d)}$. Customers in these flows are called primary. At a state $e^{(k)}$, $k = 1, 2, \ldots, d$, customers in a flow Π_j, $j = 1, 2,$

$\ldots$, m, arrive in batches so that the flow of batches in poissonian with intensity $\lambda_j^{(k)}$, batch sizes are independent random variables. A batch contains b customers with probability $p_b^{(j,k)}$, $b = 1, 2, \ldots$. In the computer simulation model it is assumed $p_1^{(j,k)} = 1 - b_j^{(k)}$, $p_b^{(j,k)} = b_j^{(k)}(1 - q_j^{(k)})(q_j^{(k)})^{b-2}$, $b = 2, 3, \ldots$, witch corresponds to Bartlett flow [2]. Customers in the flow Π_j are placed in a buffer O_j with infinite capacity. Service duration for a customer in O_j is a random variable with probability distribution function $B_j(t)$. After each service act the sever performs an inner readjustment. After service of the queue O_j the readjustment duration is also random with a probability distribution function $\bar{B}_j(t)$. In the computer simulation model there are five types of probability distributions for service and readjustment acts: exponential, Erlang of 2nd order, uniform, triangular and degenerate. When exponential distribution is selected, then $B_j(t) = 0$ for $t \leqslant 0$ and $B_j(t) = 1 - \exp\{-t\beta_{j,1}^{-1}\}$ for $t > 0$, $\bar{B}_j(t) = 0$ for $t \leqslant 0$ and $\bar{B}_j(t) = 1 - \exp\{-t\bar{\beta}_{j,1}^{-1}\}$ for $t > 0$, for each $j = 1, 2, \ldots, m$. When Erlang distribution is selected, $B_j(t) = 0$ for $t \leqslant 0$ and $B_j(t) = 1 - 2t\beta_{j1}^{-1}\exp\{-2t\beta_{j,1}^{-1}\} - \exp\{-2t\beta_{j,1}^{-1}\}$ for $t > 0$, $\bar{B}_j(t) = 0$ for $t \leqslant 0$ and $\bar{B}_j(t) = 1 - 2t\bar{\beta}_{j,1}^{-1}\exp\{-2t\bar{\beta}_{j,1}^{-1}\} - \exp\{-2t\bar{\beta}_{j,1}^{-1}\}$ for $t > 0$ for each $j = 1, 2, \ldots, m$. When uniform distribution is selected, $B_j(t) = 0$ for $t \leqslant 0$, $B_j(t) = t(2\beta_{j,1})^{-1}$ for $0 < t \leqslant 2\beta_{j,1}$, $B_j(t) = 1$ for $t > 2\beta_{j,1}$; $\bar{B}_j(t) = 0$ for $t < 0$, $\bar{B}_j(t) = t(2\bar{\beta}_{j,1})^{-1}$ for $0 \leqslant t < 2\bar{\beta}_{j,1}$, $B_j(t) = 1$ for $t \geqslant 2\bar{\beta}_{j,1}$. When triangular distribution is selected, $B_j(t) = 0$ for $t \leqslant 0$, $B_j(t) = t^2(2\beta_{j,1})^{-1}$ for $0 < t \leqslant b_{j,1}^{-1}$, $B_j(t) = 1 - (2\beta_{j,1} - t)^2(2\beta_{j,1}^2)^{-1}$ for $\beta_{j,1} < t \leqslant 1$, $B_j(t) = 1$ for $t > 2\beta_{j,1}$, $\bar{B}_j(t) = 0$ for $t \leqslant 0$, $\bar{B}_j(t) = t^2(2\bar{\beta}_{j,1})^{-1}$ for $0 < t \leqslant \bar{b}_{j,1}^{-1}$, $B_j(t) = 1 - (2\bar{\beta}_{j,1} - t)^2(2\bar{\beta}_{j,1}^2)^{-1}$ for $\bar{\beta}_{j,1} < t \leqslant 1$, $\bar{B}_j(t) = 1$ for $t > 2\bar{\beta}_{j,1}$ for each $j = 1, 2, \ldots, m$. When degenerate distribution is selected, $B_j(t) = 0$ for $t \leqslant \beta_{j,1}$, $B_j(t) = 1$ for $t > \beta_{j,1}$, $\bar{B}_j(t) = 0$ for $t \leqslant \bar{\beta}_{j,1}$, $\bar{B}_j(t) = 1$ for $t > \bar{\beta}_{j,1}$ for each $j = 1, 2, \ldots, m$. Notice that mathematical expectations for each of the distributions for service durations equal $\beta_{j,1}$, mathematical expectations for each of the distributions for service durations equal $\bar{\beta}_{j,1}$, but variances decrease ($\beta_{j,1}^2$, $\frac{1}{2}\beta_{j,1}^2$, $\frac{1}{3}\beta_{j,1}^2$, $\frac{1}{6}\beta_{j,1}^2$, 0 for the service durations, $\bar{\beta}_{j,1}^2$, $\frac{1}{2}\bar{\beta}_{j,1}^2$, $\frac{1}{3}\bar{\beta}_{j,1}^2$, $\frac{1}{6}\bar{\beta}_{j,1}^2$, 0 for the readjustment durations). If the queues are empty after an readjustment then the first incoming customer is chosen for service. On the contrary if the queues' lengths are represented with a nonzero vector $x = (x_1, x_2, \ldots, x_m)$ then a customer from the queue number $j = h(x)$ is chosen, where $h(\cdot)\colon X = \{0, 1, \ldots\}^m \to \{1, 2, \ldots, m+1\}$ is a given mapping of the nonnegative integer lattice such that the only inverse of the point $n = m + 1$ is the zero vector $\bar{0} = (0, 0, \ldots, 0) \in X$. In particular the computer simulation model includes the following switching functions: 1) let σ be a fixed permutation of the elements of $\{1, 2, \ldots, m\}$ determining the relative priorities for the queues, then for $x \neq \bar{0}$, $x_{\sigma(1)} = x_{\sigma(2)} = \ldots = x_{\sigma(j-1)} = 0$, $x_{\sigma(j)} > 0$ the mapping $h_\sigma(x)$ takes on value j; 2) service of the longest queue, $h_{\max}(x) = \min\{j\colon x_j = \max\{x_r\colon 1 \leqslant r \leqslant m\}\}$; 3) mixed switching function $h_r(\cdot)$, taking on value r when $x_r > 0$, and value $\min\{j\colon x_j = \max\{x_{j'}\colon 1 \leqslant j' \leqslant m, j' \neq r\}\}$ when $x_r = 0$. In the environment's state $e^{(k)}$ a served customer from the queue O_j with probability $\pi_{j,r}^{(k)}$ goes into the queue O_r for repeated service, and with probability $\pi_{j,n}^{(k)} = 1 - \sum_{r=1}^m \pi_{j,r}^{(k)}$ leaves the queueing system. Thus the input flows of the queueing system are superpositions of primary and secondary flows of customers. Changing of the random environment's state can occur only at instants of service and readjustment terminations, so that the sequence of the random environment's states makes

a irreducible aperiodic Markov chain. The probability of a transition from a state $e^{(l)}$, $l = 1, 2, \ldots, d$, into $e^{(k)}$, $k = 1, 2, \ldots, d$, equals $a_{l,k}$. The sojourn cost for a single customer per time unit in queue O_j is c_j. Put $\tau_0 = 0$, let τ_i be either an instant of service termination or readjustment termination, $i = 1, 2, \ldots$. An interval $(\tau_{i-1}, \tau_i]$ is called the i-th working tact of the queueing system. Denote by $\zeta_{j,i}$ the total sojourn time of all customers in the queue O_j during the i-th working tact; then the mean sojourn cost for all customers during the i-th working tact is $J_i(h) = \sum_{j=1}^{m} \mathbf{E}(c_j \zeta_{j,i})$ (here $\mathbf{E}$ denotes the mathematical expectation symbol). We will treat the functional $J_i(h)$ as the economic criterium for the system performance during the i-th working tact. A switching function $h(\cdot)$ is optimal if it minimizes the functional $J_i(h)$ in the stationary regime of operation.

Denote by $\chi_i \in \{e^{(1)}, e^{(2)}, \ldots, e^{(d)}\}$ the state of the external random environment during the interval $(\tau_i, \tau_{i+1}]$, by $\varkappa_{j,i}$ the length of the queue O_j at τ_i counting the arriving secondary customer, by $\Gamma_i \in \Gamma = \{\Gamma^{(1)}, \Gamma^{(2)}, \ldots, \Gamma^{(n)}\}$ the server state during the interval $(\tau_{i-1}, \tau_i]$. For $1 \leqslant s \leqslant m$ $i = 1, 2, \ldots$ the equation $\Gamma_i = \Gamma^{(s)}$ takes place if a service of a customer from O_s is taking place or has just terminated, and for $s = n$ a readjustment is taking place or has just terminated. An element Γ_0 with values in Γ defines the initial state of the server at τ_0. Put for convenience $\varkappa_i = (\varkappa_{1,i}, \varkappa_{2,i}, \ldots, \varkappa_{m,i})$, $\lambda_+^{(k)} = \lambda_1^{(k)} + \lambda_2^{(k)} + \ldots + \lambda_m^{(k)}$, $Q_i^{(s,k)}(w) = \mathbf{P}(\{\Gamma_i = \Gamma^{(s)}, \varkappa_i = w, \chi_i = e^{(k)}\})$. Using methods presented in [3] we get

Theorem 1. *A sequence*

$$\{(\Gamma_i, \varkappa_i, \chi_i); i = 0, 1, \ldots\} \tag{1}$$

given the probability distribution of the vector $(\Gamma_0, \varkappa_0, \chi_0)$ *is a Markov chain. The states of Markov chain (1) form a unique class of communicating periodic states.*

In the next section we will present necessary and sufficient conditions for the stationary distribution existence for the Markov chain (1) in order to solve the optimization problem.

2 Conditions for stationary distribution existence and solution of optimization problem in some particular cases

Denote by $\{\mathfrak{a}_l \colon l = 1, 2, \ldots, d\}$ the stationary probability distribution for the random environment. Put $\pi^{(k)} = (\pi_{j,r}^{(k)} \colon j, r = 1, 2, \ldots, m)$, $\pi = (\pi_{j,r} \colon j, r = 1, 2, \ldots, m) = \sum_{k=1}^{d} \mathfrak{a}_k \pi^{(k)}$, $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$, $\bar{\beta} = (\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_m)$, $\mu_{j,1}^{(k)} = \sum_{b=1}^{\infty} b p_b^{(k.j)}$, $\bar{\lambda}_j^{(k)} = \lambda_j^{(k)} \mu_{j,1}^{(k)}$,
$\bar{\lambda}^{(k)} = (\bar{\lambda}_1^{(k)}, \bar{\lambda}_2^{(k)}, \ldots, \bar{\lambda}_m^{(k)})^T$. Using methods in [1], one can prove Theorem 2, which contains in fact necessary conditions for the stationary distribution existence for Markov chain (1).

Theorem 2. *Assume that one of the following hypotheses holds: 1) the largest absolute value* R *of eigenvalues of the matrix* π *is 1; 2)* $R < 1$, $(\beta + \bar{\beta})(I_m - \pi^T)^{-1}\big(\sum_{l=1}^{d} \mathfrak{a}_l \bar{\lambda}^{(l)}\big) > 1$. *Then for every* $(\Gamma^{(s)}, w, e^{(k)}) \in \Gamma \times X \times \{e^{(1)}, e^{(2)}, \ldots, e^{(d)}\}$ *and independently of the initial probability*

distribution of $(\Gamma_0, \varkappa_0, \chi_0)$ *the following limit relation takes place:*

$$\lim_{i\to\infty} Q_i^{(s,k)}(y^{(n)}) = 0. \tag{2}$$

It is known [4] that a denumerable Markov chain either has a stationary probability distribution (not necessarily unique if the states form more that one positive recurrent class), or equation (2) takes place for each $(\Gamma^{(s)}, w, e^{(k)}) \in \Gamma \times X \times \{e^{(1)}, e^{(2)}, \ldots, e^{(d)}\}$ independently of the initial probability distribution. In the last case the sequence $\{\mathbf{E}(\sum_{j=1}^m \varkappa_{j,i}); i = 0, 1, \ldots\}$ increases unboundly. So, the boundness of the sequence of mean number of customers in the queueing system gives the sufficient condition for a stationary distribution existence.

Theorem 3. *Assume* $R < 1$ *and* $(\beta + \bar\beta)(I_m - \pi^T)^{-1}\Big(\sum_{l=1}^d \mathfrak{a}_l \bar\lambda^{(l)}\Big) < 1$. *Then the sequence* $\{\mathbf{E}(\sum_{j=1}^m \varkappa_{j,i}); i = 0, 1, \ldots\}$ *is bounded.*

Assuming that the conditions for a stationary distribution existence are fulfilled find the expression of the functional $J_i(h)$ with means of queueing system's parameters and random environment parameters.

Theorem 4. *If either* $a_{l,k} = \mathfrak{a}_k$ *for* l, $k = 1, 2, \ldots, d$, *or* $\bar\lambda^{(1)} = \bar\lambda^{(2)} = \ldots = \bar\lambda^{(d)}$, *then*

$$J_i(h) = J(h) = \sum_{j=1}^m c_j \sum_{r=1}^m (\beta_{r,1} + \bar\beta_{r,1}) x^{r,j} + \text{const} \tag{3}$$

and the constant in the right-hand side doesn't depend on $h(\cdot)$.

Theorem 5. *Let* $\beta_{r,2} = \int_0^\infty t^2\, dB_r(t)$, $\bar\beta_{r,2} = \int_0^\infty t^2\, d\bar B_r(t)$, $r = 1, 2, \ldots, m$, $a_{l,k} = \mathfrak{a}_k$. *Then quantities* $x^{g,j}$, g, $j = 1, 2, \ldots, \ldots, m$, *satisfy the system of equations*

$$x^{g,j} + x^{j,g} - \sum_{r=1}^m (\bar d^{\,r,j} x^{r,g} + \bar d^{\,r,g} x^{r,j}) = \bar c^{\,g,j}, \tag{4}$$

where $\bar d^{\,r,g} = (\beta_{r,1} + \bar\beta_{r,1}) \sum_{k=1}^d \mathfrak{a}_k \bar\lambda_g^{(k)} + \pi_{r,g}$, $\bar d^{\,r,j} = (\beta_{r,1} + \bar\beta_{r,1}) \sum_{k=1}^d \mathfrak{a}_k \bar\lambda_j^{(k)} + \pi_{r,j}$, *constants* $\bar c^{\,g,j}$ *do not depend on the mapping* $h(\cdot)$ *and* $\bar c^{\,g,j} = \bar c^{\,j,g}$.

Theorem 6. *Let* $\bar\lambda^{(1)} = \bar\lambda^{(2)} = \ldots = \bar\lambda^{(d)}$, $\pi_{j,r}^{(1)} = \pi_{j,r}^{(2)} = \ldots = p_{j,r}^{(d)}$, $\lambda_r^{(1)} b_r^{(1)} (1 - q_r^{(1)})^{-2} = \lambda_r^{(2)} b_r^{(2)} (1 - q_r^{(2)})^{-2} = \ldots = \lambda_r^{(d)} b_r^{(d)} (1 - q_r^{(d)})^{-2}$ *for* j, $r = 1, 2, \ldots, m$. *Then quantities* $x^{g,j}$, g, $j = 1, 2, \ldots, \ldots, m$, *satisfy the system of equations*

$$x^{g,j} + x^{j,g} - \sum_{r=1}^m (\bar{\bar d}^{\,r,j} x^{r,g} + \bar{\bar d}^{\,r,g} x^{r,j}) = \bar{\bar c}^{\,g,j}, \tag{5}$$

where $\bar{\bar d}^{\,r,g} = (\beta_{r,1} + \bar\beta_{r,1}) \sum_{k=1}^d \mathfrak{a}_k \bar\lambda_g^{(k)} + \pi_{r,g}$, $\bar{\bar d}^{\,r,j} = (\beta_{r,1} + \bar\beta_{r,1}) \bar\lambda_j^{(1)} + \pi_{r,j}$, *constants* $\bar{\bar c}^{\,g,j}$ *do not depend on the mapping* $h(\cdot)$ *and* $\bar{\bar c}^{\,g,j} = \bar{\bar c}^{\,j,g}$.

Thus we get the linear programming problem: find the minimum for the functional

$$J^0(h) = \sum_{j=1}^{m} c_j \sum_{r=1}^{m} (\beta_{r1} + \bar{\beta}_{r1}) x^{rj}$$

under linear constraints (4) or (5) and constraints $x^{rj} \geqslant 0$. An optimization problem of this type was first solved by G. P. Klimov in [5]. Recall that the queueing system under study differs much from that of G. P. Klimov in presence of the random stationary environment, non-ordinary input flows, varying redirection probabilities of secondary customers and readjustments. Despite of this the linear programming problems differ only in particular form of the coefficients in expressions for the objective functionals and linear constraints. Despite of the wide class of admissible controls, the optimal switching function $h^*(\cdot)$ belongs to the class of priority control. Priority indices can be identified with means of the next algorithm. Put $D_1 = \{1, 2, \ldots, m\}$, $c_r(D_1) = c_r$. Next, let $\{\vartheta_r(D_s) \colon r \in D_s\}$ be the solution to the system

$$\vartheta_r(D_s) = \sum_{j \in D_s} \vartheta_j(D_s) \pi_{r,j} + \beta_{r,1} + \bar{\beta}_{r,1}, \quad r \in D_s,$$

$r(D_s) = \arg\min\{c_j(D_s)/\vartheta_j(D_s) \colon j \in D_s\}$, $D_{s+1} = D_s \backslash \{r(D_s)\}$, $c_r(D_{s+1}) = (c_r(D_s) \times (\vartheta_r(D_s))^{-1} - \min\{c_j(D_s)/\vartheta_j(D_s) \colon j \in D_s\})\vartheta_r(D_s)$. Put $h^*(x) \in D_s$, when $\sum_{r \in D_{s+1}} x_r$ is zero and $\sum_{r \in D_s} x_r > 0$. Then the switching function $h^*(\cdot)$ is optimal.

3 Description of computer simulation model and numerical experiments

Computer simulation of the queueing system is done in intervals $(\tau_i, \tau_{i+1}]$, $i = 0, 1, \ldots$. Here we describe the steps of the simulation algorithm. Assume that the environment state χ_i is $e^{(k)}$ during the interval and the server is in the state $\Gamma_i = \Gamma^{(s)}$. Arrival instants $\widehat{\tau}_\iota$ of primary customers during the interval are generated and kept in appropriate queues. The number of customers in the queue O_j at the instant τ_i is determined by the variable $\varkappa_{i,j}$. If at the instant τ_i a readjustment has terminated and the queues are empty then the following independent random variables are generated: the index of the flow with the earliest arrival is a discrete random variable taking on value j with probability $\lambda_j^{(k)}/\lambda_+^{(k)}$; the batch size for the first arrival is a random variable taking on value 1 with probability $1 - b_j^{(k)}$ and value $\vartheta \geqslant 2$ with probability $b_j^{(k)}(1 - q_j^{(k)})(q_j^{(k)})^{\vartheta - 2}$; and finally the time from τ_i to the earliest arrival is a random variable with exponential distribution with parameter $\lambda_+^{(k)}$. All customers from the first batch are kept is the right queue with record of their arrival time. After this according to the switching function $\tilde{s} = h(x)$ the next queue O_s is selected for service, $\Gamma_{i+1} = \Gamma^{(\tilde{s})}$. This step is also done if τ_i a readjustment termination occurred and not all queues are empty. If at instant τ_i a service terminated, then s is determined from $\Gamma_i = \Gamma^{(s)}$, and put $\Gamma_{i+1} = \Gamma^{(n)}$. Now we generate the service or readjustment duration using an appropriate probability distribution function $B_{\tilde{s}}(t)$ for service interval, and $\bar{B}_s(t)$ for readjustment

interval. Thus we get a value t_Δ of a random variable Δ_i. For a queue O_j we generate the arrival instants of primary customers in the interval $(\tau_{i+1} - \Delta_i, \tau_i]$. To do this first determine the number of batches with Poisson probability distribution with parameter $\lambda_j^{(k)} t_\Delta$. Then for each batch we randomly generate its size as above and its arrival instant which has a uniform conditional probability distribution in the interval $(\tau_{i+1} - \Delta_i, \tau_{i+1})$. If $\Gamma_{i+1} \neq \Gamma^{(n)}$ and $\chi_i = e^{(k)}$, then the redirection of the served customer is simulated. We sample a value r of a discrete random variable with probabilities $\pi_{\tilde{s},1}^{(k)}$, $\pi_{\tilde{s},2}^{(k)}, \ldots, \pi_{\tilde{s},n}^{(k)}$ of the values 1, 2, …, n accordingly. If a value $r < n$ is obtained, then we move the customer from $O_{\tilde{s}}$ into O_r, otherwise we calculate the sojourn time for the customer as the difference between τ_{i+1} and its arrival time and place the customer in the output flow memorizing its sojourn time. Finally we simulate the random environment state change.

To estimate stationary characteristics of the queueing system performance the simulation of one realization was accomplished in two stages. At the first stage the termination instant of the transition process is determined, after this instant the process is considered to be in the quasi-stationary mode. To carry out this instant determination with given parameters of the external random environment, input flows, service and readjustments durations, redirection for repeated service two queueing systems are run simultaneously. Initially the queues of the first queueing system are empty while the lengths of the queues of the second queueing system are described with the nonzero vector $x^* \in X$. At the i-th simulation step a departure of a predefined number of served customers from each system is awaited. Then the estimates $\hat{\gamma}_i(y^{(n)})$ and $\hat{\gamma}_i(x^*)$ of the mean sojourn time of an arbitrary customer in each system and the loading estimates $\hat{\rho}_i(y^{(n)})$, $\hat{\rho}_i(x^*)$ (as the ratios of working time to the total simulation time) in each system are found. Given accuracies $\varepsilon_{\text{mean}} \in (0,\ 1)$ and $\varepsilon_{\text{load}} \in (0,\ 1)$, if the inequalities $|\hat{\gamma}_i(y^{(n)}) - \hat{\gamma}_i(x^*)| \geqslant \varepsilon_{\text{mean}} \hat{\gamma}_i(y^{(n)})$ or $|\hat{\rho}_i(y^{(n)}) - \hat{\rho}_i(x^*)| \geqslant \varepsilon_{\text{load}} \hat{\rho}_i(y^{(n)})$ take place then we proceed to the $(i+1)$-th step. But if both $|\hat{\gamma}_i(y^{(n)}) - \hat{\gamma}_i(x^*)| < \varepsilon_{\text{mean}} \hat{\gamma}_i(y^{(n)})$ and $|\hat{\rho}_i(y^{(n)}) - \hat{\rho}_i(x^*)| < \varepsilon_{\text{load}} \hat{\rho}_i(y^{(n)})$ then it is considered as the entrance into quasi-stationary mode. In out experiments we put $\varepsilon_{\text{mean}} = 0.05$, $\varepsilon_{\text{load}} = 0.05$. At the second stage the simulation of the work of the first queueing system is continued to compute estimated of performance characteristics in the quasi-stationary regime. For each realization the following performance characteristics are estimated: an estimate $\widehat{J}$ of the mean sojourn cost of all customers during one working tact, an estimate $\widehat{\gamma}$ of the mean sojourn time for a single customer. Thus the computer simulation program allows to conduct the numerical optimization of the time-sharing servicing process with respect to two objectives: minimization of the mean sojourn time of a single customer and minimization of the mean sojourn cost for all customers during one working tact. Notice that both objective functionals are different from one in [5]. The program is written in C++ and uses `pthread` library to allow parallel simulation of different realizations on multicore processors.

We would like to remind that the optimal switching function was found in Section 2 only for aperiodic external environment and mean characteristics of primary and secondary customers independent of the random environment. In practice the we have observed cases when we may use a modified Klimov's algorithm and cases when we may not use this algorithm. Consider the following numerical example. Let $d = 3$, $p_{1,2}^{(1)} = p_{2,3}^{(1)} = p_{3,1}^{(1)} = 0.9$, $p_{1,1}^{(2)} = p_{1,2}^{(2)} = p_{1,3}^{(2)} = 0.1$,

$p^{(2)}_{2,1} = 0.1$, $p^{(2)}_{3,1} = p^{(2)}_{3,2} = 0.125$, $p^{(3)}_{1,1} = 0.5$, $p^{(3)}_{1,3} = 0.2$, $p^{(3)}_{3,1} = p^{(3)}_{3,2} = 0.125$, $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 2$, $\bar{\beta}_1 = 0.15$, $\bar{\beta}_2 = 0.25$, $\bar{\beta}_3 = 1$, $\lambda^{(1)}_1 = 0.145$, $\lambda^{(1)}_2 = 0.13$, $\lambda^{(1)}_3 = 0.125$, $b^{(1)}_1 = b^{(1)}_2 = b^{(1)}_3 = 0.5$, $q^{(1)}_1 = q^{(1)}_2 = q^{(1)}_3 = 0$, $\lambda^{(2)}_1 = 0.0125$, $\lambda^{(2)}_2 = 0.0105$, $\lambda^{(2)}_3 = 0.025$, $b^{(2)}_1 = b^{(2)}_2 = b^{(2)}_3 = 0.5$, $q^{(2)}_1 = q^{(2)}_2 = q^{(2)}_3 = 0$, $\lambda^{(3)}_1 = \lambda^{(3)}_2 = \lambda^{(3)}_3 = 0.05$, $b^{(3)}_1 = b^{(3)}_2 = b^{(3)}_3 = 0$, $q^{(3)}_1 = q^{(3)}_2 = q^{(3)}_3 = 0$, and the transition probabilities for the environment are given by one of the following sets: 1) $a_{1,1} = 0.1$, $a_{1,2} = 0.8$, $a_{1,3} = 0.1$, $a_{2,1} = 0.2$, $a_{2,2} = 0.7$, $a_{2,3} = 0.1$, $a_{3,1} = 1$; 2) $a_{1,2} = a_{2,3} = a_{3,1} = 1$. The sojourn costs per unit time are given by $c = (2, 1, 5)$. In Table 1 the estimates for the mean sojourn time of a single customer, the mean sojourn cost of all customers in the system and the load are shown for both sets of transition probabilities of the environment. The smallest computed values are printed in bold face.

Table 1: for the mean sojourn time of a single customer, the mean sojourn cost of all customers in the system and the load for different transition probabilities of the environment

Switching function	$\hat{\gamma}$		$\hat{J}$		$\hat{\rho}$	
	set 1	set 2	set 1	set 2	set 1	set 2
$h_{\max}$	6.31	11.25	6.03	8.59	0.310	0.569
h_σ, $\sigma = \begin{pmatrix}1&2&3\\1&2&3\end{pmatrix}$	5.51	11.45	6.29	11.61	**0.309**	0.609
h_σ, $\sigma = \begin{pmatrix}1&2&3\\1&3&2\end{pmatrix}$	7.04	21.22	**5.22**	8.88	**0.309**	0.657
h_σ, $\sigma = \begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}$	**5.50**	**8.38**	6.34	8.65	0.311	0.538
h_σ, $\sigma = \begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}$	7.39	10.37	6.41	7.70	0.311	**0.483**
h_σ, $\sigma = \begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}$	7.64	15.16	5.41	7.08	0.311	0.593
h_σ, $\sigma = \begin{pmatrix}1&2&3\\3&2&1\end{pmatrix}$	7.94	12.19	6.07	7.81	0.311	0.501
h_1	5.92	14.81	5.90	11.51	0.311	0.629
h_2	6.09	8.38	6.37	7.69	0.312	0.515
h_3	7.70	11.97	5.62	**6.81**	0.311	0.542

From the table one can see that a 'fair' algorithm choosing the longest queue for service is worse that at least one priority algorithm. Since π is computed as $\mathfrak{a}_1\pi^{(1)} + \mathfrak{a}_2\pi^{(2)} + \mathfrak{a}_3\pi^{(3)}$, the algorithm from the end of Section 2 suggests h_σ with

$$\sigma = \begin{pmatrix}1&2&3\\1&3&2\end{pmatrix}.$$

If we run the algorithm for $c = (1, 1, 1)$, we get

$$\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \qquad \sigma = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$$

for sets 1 and 2 correspondingly. In Table 1 we see that in several cases this prediction is not correct.

Conclusions

In this work we studied a time-sharing queueing system with readjustments with input flows formed in a random environment. We have found the stationarity conditions and we also found the optimal switching function in certain constrained cases. However we have found that optimal switching for a time-sharing servicing process with readjustments can not be found using Klimov's algorithm so more theoretical study into the subject is required.

Acknowledgements

This work was done as a part of state budget R&D of Lobachevsky State University of Nizhni Novgorod – National Research University on the theme № 0120.0602598 "Analysis of discrete controlling queueing systems and systems for boolean functions computations".

References

[1] Zorine A. V., Fedotkin M. A. (2005). Optimization of control of doubly stochastic non-ordinary flows in time-sharing systems, *Automation and remote control*. No. 7, pp. 1115–1124.

[2] Fedotkin M. A. (1981) Incomplete description of flows of nonhomogenous demands. *Queueing theory*. MGU Publishing, VNIISI, Moscow. pp. 113–118.

[3] Zorine A. V., Fedotkin M. A. (2003). Analysis and optimization of time-sharing processes, functioning in a random environment. *Vestnik of N.I.Lobachevsky State University of Nizhni Novgorod. Ser. Matematika*. Vol. 1., pp. 18–28.

[4] Shiriaev A. N. (2004). *Probability*. MCCME, Moscow.

[5] Klimov G. P. (1974) Time-sharing queueing systems I. *Theory of Probability and its Applications*. Vol. XIX, pp. 532–551.

Numerical Simulation of the Sea Surface and Extreme Ocean Waves with Stochastic Spectral Models

Sergei M. Prigarin, Kristina V. Litvenko
Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of Russian Academy of Sciences, Novosibirsk State University, Novosibirsk, Russia
e-mail: sergeim.prigarin@gmail.com

Abstract

In this paper, we make an attempt to apply stochastic spatial-temporal conditional spectral models of the sea surface undulation to study features of formation and development of the ocean extreme waves.

Keywords: numerical simulation, sea surface undulation, extreme waves, rogue waves, spectral models, random fields.

Introduction

Extreme (rogue, freak) waves are known as a natural ocean phenomenon and hazards for mariners. Extreme waves are distinct from tsunami, as they are usually highly localized in space and duration, and unexpectedly occur far out at sea. The shapes of extreme waves can vary: sometimes they look like solitary waves, sometimes they appear as a group of waves (like the 'Three Sisters' on the Great Lakes) or as a wall of water (see details and further references in [14], for instance). The MaxWave research project [20] reads: 'Within the last years a high number of large ships has been lost. The causes of accidents are in many cases believed to be rogue waves'. The phenomenon of extreme waves has not been studied sufficiently well (for the first time a rogue wave was instrumentally detected only in 1995), and nowadays serious efforts are mounted to observe the extreme waves and to study the phenomenon both theoretically and experimentally.

In this paper, we attempt to simulate extreme waves by using a specific technique of conditional stochastic spectral models of random fields. The necessary information for the stochastic models proposed is a spectrum of the time-space random field, which describes the sea surface undulation. This information enables us to numerically construct time-space realizations of the surface of the sea that is statistically homogeneous in space and stationary in time. Similar models were constructed in [1, 3, 4, 5] to study optical properties of the sea swell by Monte Carlo method. To simulate waves of abnormal height in addition to the spectrum, we need to know the "true" level of the sea surface for a number of fixed points at given moments of time. The technique of conditional spectral models proposed in [15] allows one to construct an ensemble of random time-space realizations of the sea swell passing through these points. Thus, demanding the sea level to reach considerable heights, we can simulate and study the processes of occurrence and

attenuation of extreme waves. In particular, one of the most fascinating results that was obtained by numerical experiments is the existence of accompanying waves before and after the main giant wave. This group of three waves can, probably, be interpreted as a successful simulation of the 'Three Sisters'.

1 Stochastic spectral models of the sea surface undulation

The experimental data of statistical properties of wind-driven waves indicate to the fact that they may be described with a high precision by a homogeneous Gaussian random field of surface deviations from a mean level [2, 8]. Therefore, numerical models of homogeneous Gaussian fields appear to be efficient means for the sea swell simulation. In the next two subsections, we present the basic principles of constructing spectral models and specific features of spectral models for the sea surface.

1.1 Spectral models of Gaussian random fields

One of the most promising general approaches for the simulation of homogeneous Gaussian processes and fields is based on the spectral decomposition of random functions. Among a rich variety of publications devoted to this approach, we would like to mention the earliest publications [6, 11, 13, 19], paper [12], where the spectrum randomization was combined with the spectrum partitioning, papers [9, 10] with advanced results on convergence, recent publication [7], and monographs [17, 18], where the detailed information and more references to numerical spectral models can be found.

Let us consider a real-valued homogeneous random Gaussian field $w(x)$, $x \in R^k$, with zero mean, unit variance and correlation function $R(x) = Ew(x+y)w(y)$. The spectral representations of the random field and its correlation function can be written down in the following way

$$w(x) = \int\limits_P \cos < x, \lambda > \xi(d\lambda) + \int\limits_P \sin < x, \lambda > \eta(d\lambda), \tag{1}$$

$$R(x) = \int\limits_P \cos < x, \lambda > \nu(d\lambda). \tag{2}$$

Here $\xi(d\lambda)$, $\eta(d\lambda)$ are real-valued orthogonal stochastic Gaussian measures in a "spectral space" P (i.e., P is a measurable set such that $P \cap (-P) = \{0\}$, $P \cup (-P) = R^k$), $\nu(d\lambda)$ is a 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 (1) for simulation of 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: $P = \sum\limits_{j=1}^{n} Q_j, \quad Q_j \cap Q_i =$

$\emptyset$ for $i \neq j$. As an approximation of (1) we consider

$$w_n(x) = \sum_{j=1}^{n} a_j \left[\xi_j \cos < \lambda(j), x > + \eta_j \sin < \lambda(j), x >\right], \quad a_j = \nu^{1/2}(Q_j) \tag{3}$$

where ξ_j, η_j are independent standard normal variables, while vectors $\lambda(j) \in P$ belong to the corresponding sets Q_j. A specific algorithm of simulation is defined by partitioning of the spectral space P and by a method of choosing $\lambda(j) \in Q_j$. Various modifications of spectral models, their properties and convergence are discussed, for example, in [17].

1.2 Simulation of the sea surface undulation

Various investigations were performed on the spectrum of the wind-driven see surface undulation (see, for example [2, 8]). In this paper, we make use of a spectral model of the sea surface $w(x_1, x_2)$ with the following approximation of the spectrum [2]:

$$\nu(d\lambda_1 d\lambda_2) = f(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2, \quad \lambda_1 \in (0, +\infty),\ \lambda_2 \in (-\infty, +\infty),$$

$$f(\lambda_1, \lambda_2) = S_{\rho\theta}(\rho, \theta))\rho^{-1}, \quad \rho = (\lambda_1^2 + \lambda_2^2)^{1/2},\ \theta = \arg(\lambda_1 + i\lambda_2) \in [-\pi/2, \pi/2],$$

$$S_{\rho\theta}(\rho, \theta) = Q(\mu, \theta) S_\rho(\rho), \quad S_\rho(\rho) = 0.5(g/\rho)^{1/2} S((g\rho)^{1/2}),$$

$$S(\mu) = \begin{cases} 6m_0 \left(\mu_{\max}/\mu\right)^5 \mu^{-1} \exp\left\{-1.2\left[\left(\mu_{\max}/\mu\right)^5 - \left(\mu_{\max}/\mu_1\right)^5\right]\right\}, & \mu \in [0, \mu_1], \\ S(\mu_1) + \left(S(\mu_2) - S(\mu_1)\right)(\mu - \mu_1)/(\mu_2 - \mu_1), & \mu \in (\mu_1, \mu_2), \\ 0.0078 g^2 \mu^{-5}, & \mu \in [\mu_2, \mu_3); \end{cases}$$

$$Q(\mu, \theta) = 2\pi^{-1} \cos^2 \theta.$$

Here $S(\mu)$ is the "frequency" spectrum, $Q(\mu, \theta)$ is the "angular" spectrum, $\mu_3 \approx 30$ sec^{-1} is the upper boundary of the frequency spectrum of gravity waves, $\mu_{\max}$ is the frequency of a spectral maximum for S, g is the gravity acceleration,

$$\mu_1 = 1.8\mu_{\max}\widetilde{\mu}^{(-0.7)}, \quad \mu_2 = 2.0\mu_{\max}\widetilde{\mu}^{(-0.7)},$$

$$\widetilde{\mu} = v\mu_{\max}/g, \quad m_0 = 0.00127 g^{-2} v^4 \widetilde{\mu}^{(-3.19)},$$

v is the wind speed (m/sec) at the height of 10 meters above the sea level. Within this spectrum approximation, statistical properties of the sea undulation are determined by $\mu_{\max}$ and by the wind speed v. A realization of a spectral model of the surface of a wind-ruffled sea is presented in Figure 1.

A spatial-temporal spectral model of the sea surface with zero mean level can be written down in the form:

$$w_n(x_1, x_2, t) = \sum_{j=1}^{n} a_j \left[\xi_j \cos(\varphi_j(x_1, x_2, t)) + \eta_j \sin(\varphi_j(x_1, x_2, t))\right], \tag{4}$$

$$\varphi_j(x_1, x_2, t) = \lambda_1(j)x_1 + \lambda_2(j)x_2 + \mu(j)t,$$

where $\mu(j)$ are connected with $\lambda_1(j), \lambda_2(j)$ by the dispersion relation

$$\mu(j)^2 = g\rho_j \tanh(\rho_j H), \quad \rho_j = [\lambda_1^2(j) + \lambda_2^2(j)]^{1/2}.$$

In case of the deep water (for large values of depth H), the dispersion relation has the form $\mu(j) = \sqrt{g\rho_j}$. It means that if the monochromatic wavelength is equal to $L = 2\pi/\rho$, then the velocity of the wave is equal to $\sqrt{gL/(2\pi)} = \mu/\rho$, where $\mu = \sqrt{g\rho}$.

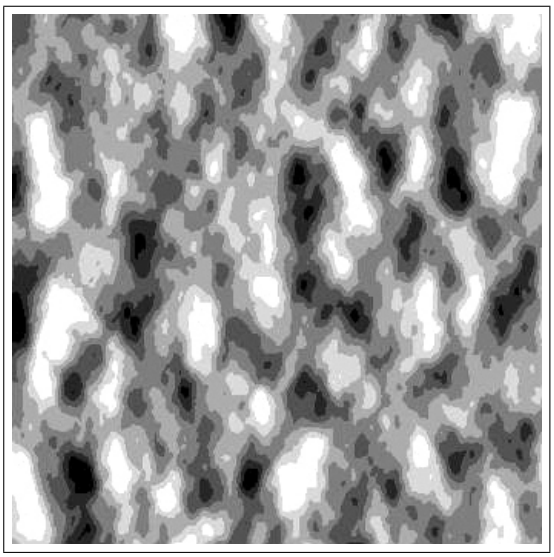


Figure 1: An example of the simulated topography of the sea surface roughness (a spectral model). The wind direction is parallel to the horizontal axis

2 Conditional spectral models and simulation of the sea surface with abnormally high points of elevation

Below we present a novel approach that can be used to simulate ensembles of temporal-spatial realizations of the sea surface with abnormally high waves (see Figures 2,3). This approach is based on conditional spectral models.

Assume, that it is necessary to simulate Gaussian random field (1) which satisfies the condition

$$w(x_m) = b_m, \quad m \in \{1, 2, \ldots, M\}. \tag{5}$$

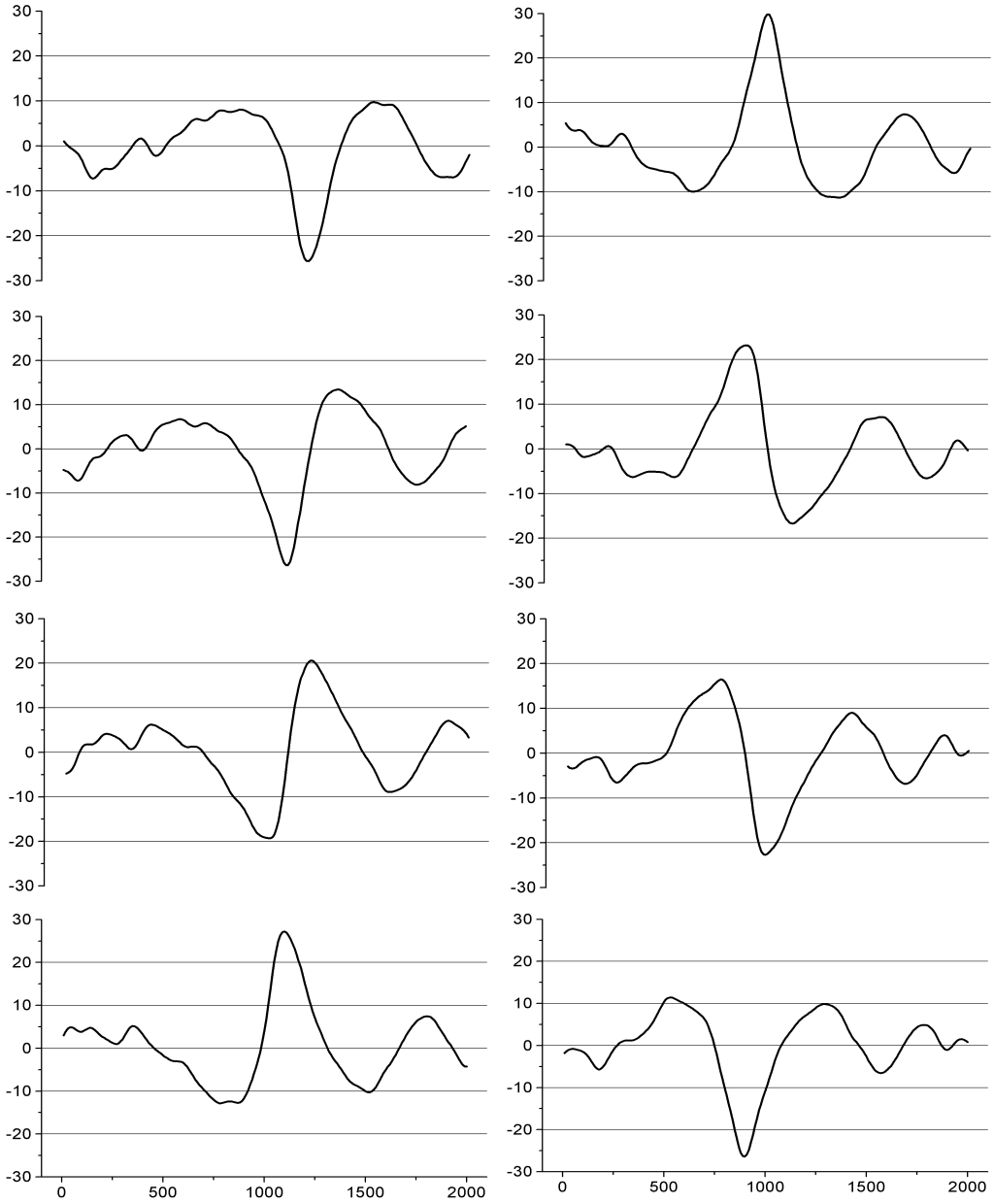


Figure 2: An extreme wave with 30m maximum height and deep troughs simulated by a conditional spectral model: 8 profiles of 2000 meters length at time 0, 4, 8, 12, 16, 20, 24, 28 sec. The first four profiles are on the left (sequentially from top to bottom) and the next four profiles are on the right. The extreme wave moves from right to left

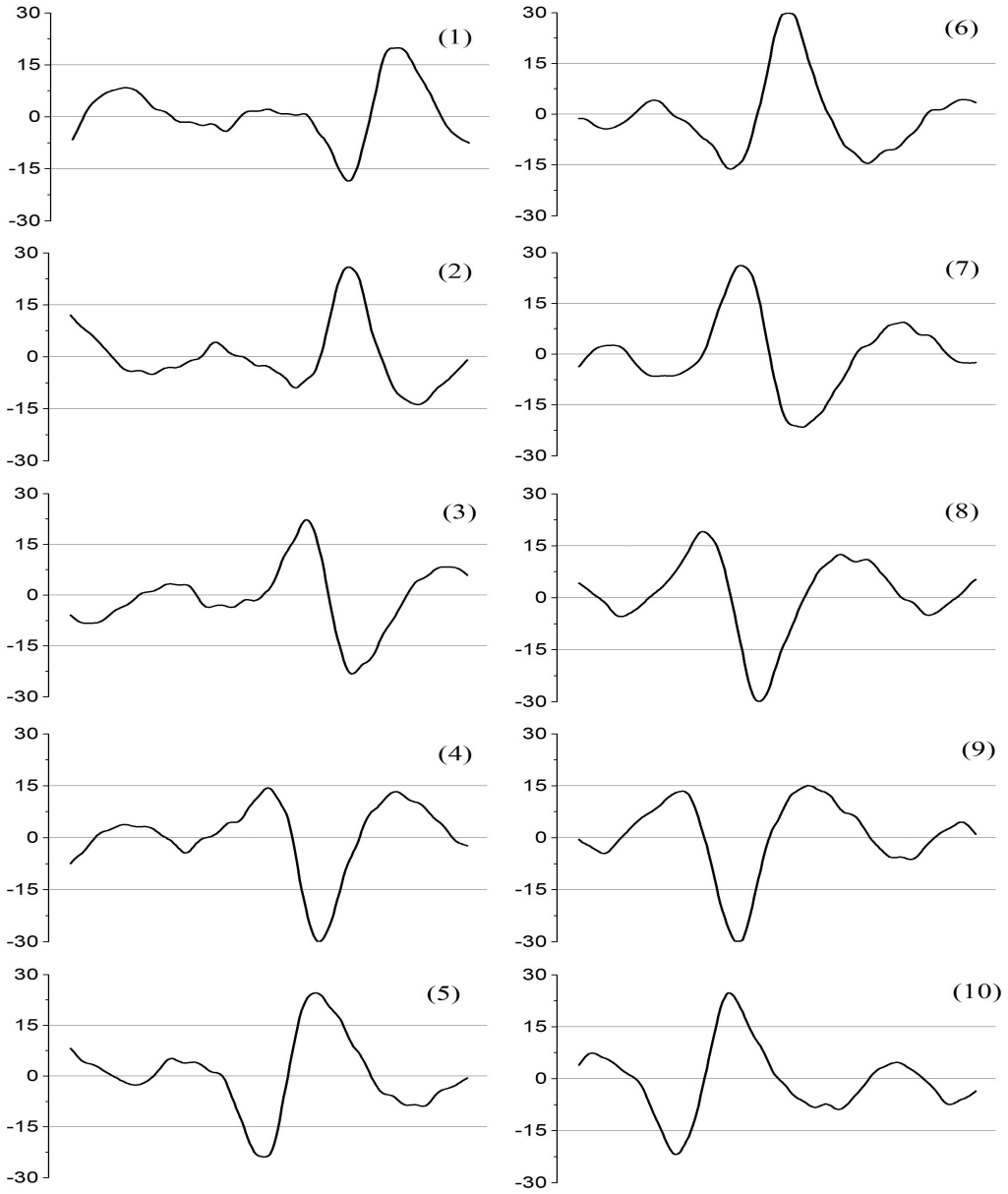


Figure 3: The "Three Sisters" with 30m maximum height simulated by a conditional spectral model: 10 profiles of 2000 meters length at the instants 0, 8, 16, 22, 32, 38, 46, 52, 56, 66 sec. The first five profiles are on the left (sequentially from top to bottom) and the next five profiles are on the right. The first wave (images 1, 2), the second (main) wave (images 5, 6, 7), and the third wave (image 10) can be well defined

This means that in spectral model (3), the standard normal random variables ξ_j, η_j must be simulated under additional constraints:

$$\sum_{j=1}^{N} d_{mj}\xi_j + d'_{mj}\eta_j = b_m, \quad m = 1, \ldots, M,$$

$$d_{mj} = a_j \cos < \lambda(j), x_m >, d'_{mj} = a_j \sin < \lambda(j), x_m > .$$

The corresponding simulation methods for the conditional spectral models were proposed in [15] (see, also, [16, 17]).

Figures 2, 3 present 2000 meter long realizations of conditional spectral models with the parameters $\mu_{\max} = 0.3$ sec^{-1}, $v = 10$ m/sec. These parameters correspond to the sea swell with waves up to 10 meters high. The extreme wave in Figure 2 was simulated at the instants 0, 4, 8, 12, 16, 20, 24, 28 sec. under additional condition (5) that at the time of 14 sec., its height is equal to 30 meters. One can see deep troughs just before and after the simulated extreme wave. Similarly, the extreme wave in Figure 3 was simulated at the instants 0, 8, 16, 22, 32, 38, 46, 52, 56, 66 sec. subject to the requirement that at the time of 38 sec., its height is equal to 30 meters (see image 6 in Figure 3). The profiles in Figure 3 show that two smaller waves appear before and after the main extreme wave. This can be interpreted as a shape of the freak wave known as 'Three Sisters'.

Conclusions

Preliminary results of stochastic simulation of the sea surface undulation with abnormally high waves based on conditional spectral models show that this approach can be an efficient tool when studying the formation and development of the ocean extreme waves.

Acknowledgements

We take this opportunity to thank the Russian Foundation for Basic Research for the financial support (grants 09-05-00963, 11-01-00641).

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] Davidan I.M., Lopatukhin L.I., Rozhkov V.A. (1985). *Wind-Driven Undulation in the World Ocean*. Gidrometeoizdat, Leningrad, [in Russian].

[3] Kanter R.R, Prigarin S.M. (1989). *Numerical modelling of the wind driven undulation for investigation of the field of reflected optical radiation*. Preprint 829. Compt. Cent. Sib. Division, USSR Acad. Sci., Novosibirsk, pp.25 [in Russian].

[4] Kargin B.A., Oppel U.G., Prigarin S.M. (1999). Simulation of the undulated sea surface and study of its optical properties by Monte Carlo method. *Proc. SPIE*. Vol. **3583**, pp.316-324.

[5] Kargin B.A., Prigarin S.M. (1992). Simulation of the sea undulation surface and study of its optical properties by Monte Carlo method. *Atmospheric and Oceanic Optics*. **Vol.5**, No.3, pp.186-190.

[6] Kraichnan R.H. (1970), Diffusion by a random velocity field, *Phys. Fluids*, **13**, No 1, P.22–31.

[7] Kramer P., Kurbanmuradov O., Sabelfeld K. (2007). Comparative analysis of multiscale Gaussian random field simulation algorithms, *J. Comp. Phys.*, **226**, P.897–924.

[8] Krylov Yu.M. (1966). *Spectral Methods of Investigation and Calculation of Wind-Driven Waves*. Gidrometeoizdat, Leningrad [in Russian].

[9] Kurbanmuradov O. (1995). Convergence of numerical models for the Gaussian fields, *Russ. J. Num. Anal. and Math. Modell.*, **10** (1995), No 4, P. 311–324.

[10] Kurbanmuradov O. (1995). Weak convergence of approximate models of random fields, *Russ. J. Num. Anal. and Math. Modell.*, **10**, No 6, P. 500–517.

[11] Matheron G. (1973). The intrinsic random functions and their applications, *Advances in Applied Probability*, **5**, P.439–468.

[12] Mikhailov G.A. (1978). Numerical construction of a random field with given spectral density, *Dokl. USSR Ac. Sci.*, **238**, No 4, pp.793–795.

[13] Orfeuil J.P. (1972), *Simulation du Wiener-Lévi et de ses intégrales*, Internal report, Centre de Morphologie Mathématique, Fontaineblau.

[14] Pelinovsky E., Kharif Ch., Eds. (2008), *Extreme Ocean Waves*. Springer, Berlin.

[15] Prigarin S.M. (1997), *Conditional Spectral Models of Gaussian Homogeneous Fields*. Preprint 1098. Compt. Cent. Sib. Division, USSR Acad.Sci, Novosibirsk, pp.19 [in Russian].

[16] 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.

[17] Prigarin S.M. (2001), *Spectral Models of Random Fields in Monte Carlo Methods*. VSP, Utrecht.

[18] Prigarin S.M. (2005), *Numerical Modeling of Random Processes and Fields*, Inst. of Comp. Math. and Math. Geoph. Publ., Novosibirsk, [in Russian].

[19] Shinozuka M. (1971), Simulation of multivariate and multidimensional random processes. *J. of Acoust. Soc. Am.*. Vol.**49**, pp.357–368.

[20] Web Page of MaxWave Poject *http://coast.gkss.de/projects/maxwave/*

Stochastic Models of Broken Clouds (A Few Simulation Examples)

Sergei M. Prigarin
Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of Russian Academy of Sciences, Novosibirsk State University
Novosibirsk, Russia
e-mail: sergeim.prigarin@gmail.com

Abstract

This paper deals with numerical simulation of stochastic indicator fields corresponding to satellite images of broken clouds in the atmosphere. Numerical spectral models and a method based on thresholds of the Gaussian functions are used to simulate the indicator fields. An additional example of simulation of a binary pattern of granules on the photosphere of the Sun is presented as well.

Keywords: Broken clouds, stochastic geometry, numerical simulation, random fields, threshold and spectral models, Sun photosphere granules.

Introduction

Cloudiness is a major factor affecting the radiation balance in the Earth's atmosphere, and the stochastic structure of clouds brings about a considerable uncertainty in climate models. That is why it is a challenging problem to construct numerical models of clouds taking into account their random optical properties and geometry. There are a lot of different approaches to solve this problem (see, for example [14], and references therein). This paper deals with a method first proposed in [1] and then developed in [2, 11, 12, 13, 14, 17]. The method in point is based on the statistical analysis of planar indicator fields for broken clouds. The corresponding software can be found at the Internet [10]. The main objective of this paper is to present some new simulation examples to demonstrate, in addition to the previous investigations, how the method works. The method seems to be universal: see [3, 6, 7, 4, 5, 18], where the same approach was used for rather different applications. In this paper, as an additional example, simulation results are presented for a pattern of granules on the photosphere of the Sun.

1 Description of the method

Assume that $\varepsilon(x)$ is a binary random field, $\varepsilon(x) \in \{0, 1\}$, with a (multidimensional) parameter x. We consider binary models of the following type. Let $u(x)$ be a real-valued Gaussian random field, $A_0 \cup A_1 = (-\infty, \infty)$ be splitting of the real axis to two disjoint sets, and

$$\varepsilon(x) = \begin{cases} 0, & \text{for } u(x) \in A_0, \\ 1, & \text{for } u(x) \in A_1. \end{cases} \tag{1}$$

Furthermore, we make two other assumptions to simplify our considerations. First, the Gaussian field u is supposed to be homogeneous with mean zero and correlation function $K_u(x) = Eu(x+y)u(x)$ and, second, the sets A_0, A_1 are supposed to be of the form

$$A_0 = (-\infty, d), \quad A_1 = [d, +\infty), \quad d \in (-\infty, +\infty). \tag{2}$$

Parameter d will be called a threshold level and model (1) of the binary fields will be called **threshold model**.

Remark. Evidently, more general threshold models can be considered, but we restrict this study by this model because of its simplicity and adaptability for a considerably large number of applied problems. In papers [8, 11, 12, 15], the following splitting of the real axis was considered in addition to (2):

$$A_0 = (-d, d), \quad A_1 = (-\infty, d] \cup [d, +\infty), \quad d \in (0, +\infty).$$

However, numerical experiments have showen that this model is more complicated and unstable in tuning than model (2).

For the threshold model, the binary field ε (just as the Gaussian field u) is a homogeneous one. Its average is

$$m_\varepsilon = P(\varepsilon = 1) = P(u \in A_1) = 1 - \Phi(d).$$

Here Φ is the function of the standard normal distribution. For the covariance function $K_\varepsilon(x) = E\varepsilon(x+y)\varepsilon(y)$ of the field $\varepsilon(x)$, we have

$$K_\varepsilon(x) = R\left(K_u(x)\right),$$

where

$$R(\rho) = \int\limits_{\{(\xi,\eta):\, \varepsilon(\xi)\varepsilon(\eta)=1\}} \varphi_\rho(\xi, \eta) d\xi d\eta, \tag{3}$$

and

$$\varphi_\rho(\xi, \eta) = \left[2\pi\sqrt{1-\rho^2}\exp\left(\frac{\xi^2 + \eta^2 - 2\rho\xi\eta}{2(1-\rho^2)}\right)\right]^{-1}$$

is the probability density of the two-dimensional Gaussian random vector with zero mean, unit variance of the components and correlation coefficient ρ between the components. Note, that the covariance distortion for the general point-wise nonlinear transformations of the Gaussian functions was studied in [7, 8].

Thereby, the parameters of the threshold model for a binary field $\varepsilon(x)$ are the following: the threshold level d and the correlation function $K_u(x)$ of the homogeneous Gaussian field $u(x)$.

2 Estimation of parameters and isotropic spectral models

If a set of realizations of the binary field ε is available, then parameters of the model can be found in the following way. First, we compute an estimation m_ε^* of the mean value m_ε for the field ε.

Then approximation d^* for the threshold level can be found from the equation $m_\varepsilon^* = 1 - \Phi(d^*)$. Second, we compute an estimate $K_\varepsilon^*(x)$ of the covariance function $K_\varepsilon(x)$ and find approximation $K_u^*(x)$ for the correlation function $K_u(x)$ of the Gaussian field $u(x)$ using the equation

$$K_\varepsilon^*(x) = R(K_u^*(x)), \tag{4}$$

where the function R is defined in (3). For the defined parameters of the threshold model, the numerical algorithm consists in the simulation of the Gaussian field $u(x)$ with the estimated correlation function $K_u^*(x)$ and subsequent transformation (1).

Remark. The direct inversion of (4) can fail because the result K_u^* may not be positive definite and the inversion can be ill-defined. Thus, to construct the estimate K_u^* in addition to equation (4), it is reasonable to take into account different considerations such as positive definiteness, smoothness, the rate of convergence to zero, etc.

For the simulation of a binary field $\varepsilon(x)$, according to threshold model (1), it is necessary to construct realizations of the homogeneous Gaussian random field $u(x)$. Various numerical methods are known to simulate Gaussian random fields. Below we give a brief information about spectral models of the isotropic homogeneous Gaussian fields on the plane. Exactly these models are presented in [10], and we used them in the numerical experiments described in the next section.

A correlation function of any homogeneous isotropic field $u(x_1, x_2)$ on the plane can be written down in the form

$$K_u(x_1, x_2) = Eu(x_1, x_2)u(0, 0) = \sigma^2 \int_0^\infty J_0\Big(z\sqrt{x_1^2 + x_2^2}\Big) G(dz),$$

where σ^2 is variance of the field, J_0 is the Bessel function of the first kind, $G(dz)$ is a "radial" spectral measure on $[0, \infty)$, $G[0, \infty) = 1$. Further we assume that the measure G has the density: $G(dz) = g(z)dz$. (A table of spectral densities and the corresponding correlation functions of isotropic fields on the plane can be found in [9].)

Assume that $0 = Z_0 < Z_1 < \ldots < Z_{N-1} < Z_N = \infty$. For an approximate simulation of the isotropic homogeneous Gaussian field $u(x_1, x_2)$ on the plane we consider the following spectral model [9]:

$$\begin{aligned} u_{NM}(x_1, x_2) = \sigma \sum_{n=1}^{N} c_n M^{-1/2} \sum_{m=1}^{M} (-2 \ \ln \ \alpha_{nm})^{1/2} \\ \times \cos \ [x_1 z_n \ \cos \ \omega_{nm} + x_2 z_n \ \sin \ \omega_{nm} + 2\pi\beta_{nm}] \,, \end{aligned} \tag{5}$$

where

$$c_n^2 = G[Z_{n-1}, Z_n) = \int_{Z_{n-1}}^{Z_n} g(z)dz, \quad \omega_{nm} = \frac{\pi(m - \gamma_n)}{M},$$

z_n are random variables distributed in $[Z_{n-1}, Z_n)$ according to the probability density $g(z)/c_n^2$, and α_{nm}, β_{nm}, γ_n are independent random variables uniformly distributed in $[0,1]$. Thus, the field u is approximated by the spectral model u_{NM}, which is the sum of $N \times M$ harmonics with random amplitudes and random frequencies. The randomness of frequencies ensures the coincidence of correlation functions of the field u and its spectral model u_{NM}. On the other hand, random frequencies make the distributions of the field u_{NM} to be non-Gaussian. However, a sequence of the fields u_{NM} is asymptotically Gaussian if $NM \to \infty$ and $\max\limits_{n \le N}(c_n^2/M) \to 0$ (see [9] about the convergence of spectral models).

The simulation algorithm provides the computation of the following arrays:

$$A(n,m) = \sigma c_n \left(-2(\ln\ \alpha_{nm})/M\right)^{1/2}, \quad D(n,m) = 2\pi\beta_{nm},$$

$$B(n,m) = z_n\ \cos\ \omega_{nm}, \quad C(n,m) = z_n\ \sin\ \omega_{nm},$$

and the value of the field at the desired point (x_1, x_2) is calculated by the formula

$$u_{NM}(x_1,x_2) = \sum_{n=1}^{N}\sum_{m=1}^{M} A(n,m)\ \cos\left[B(n,m)x_1 + C(n,m)x_2 + D(n,m)\right].$$

Different modifications of the spectral model can be used to simplify the algorithm or to reproduce specific properties of the field. Algorithmically simpler, but less flexible is model (5), where $c_n = N^{-1/2}$ and z_n are independent and distributed on the whole semi-axis $[0,\infty)$ with probability density $g(z)$. The choice of a version of the simulation algorithm and its parameters (particularly, the number of harmonics) is specified by the intension to represent in detail the corresponding parts of the spectrum.

In order to calculate the radial spectral density $g(z)$ for isotropic spectral models on the plane, a numerical approximation of the following representation is used

$$g(z) = z\int_0^{+\infty} rJ_0(zr)K_0(r)dr, \quad K_0\left(\sqrt{x_1^2+x_2^2}\right) = K_u^*(x_1,x_2).$$

3 Simulation examples and additional remarks

The results of numerical experiments for broken clouds are presented in Figs.1-2. On the left, there is a realization of the observed binary field obtained from a satellite image of clouds. On the right, there is a realization of the isotropic threshold model constructed on the basis of statistical analysis of the observed binary field by the method described in the previous Sections. The satellite images were borrowed from the NASA Earth Observatory Site.

Fig. 3 presents the results of simulation of a binary texture generated by an image of granules on the photosphere of the Sun. The source image was borrowed from NASA-Marshall Solar Physics Site.

To simulate Gaussian random fields on the plane, an isotropic spectral model of the form of (5) was used with 2500 harmonics for the images presented in Figs. 1,3 and with 5000 harmonics for the image presented in Fig. 2.

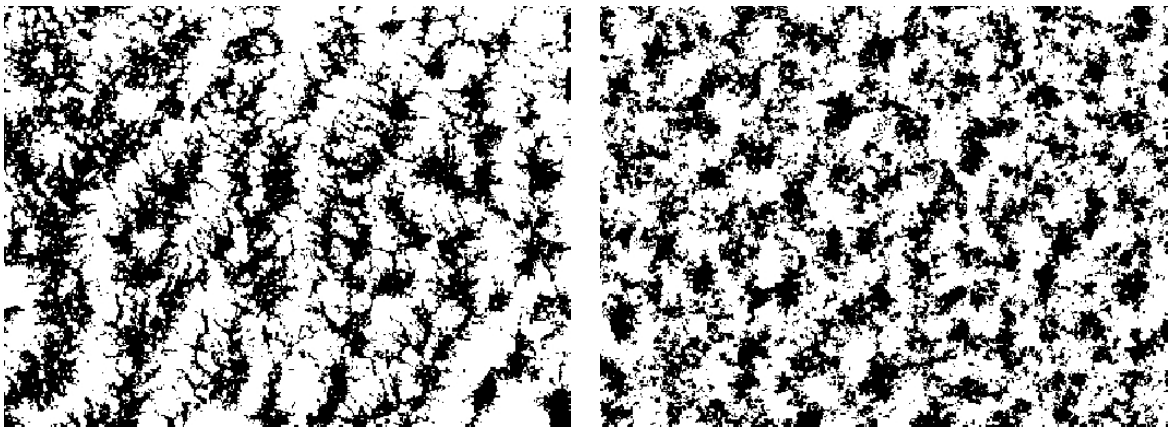


Figure 1: A satellite image of a 540km × 400km region of cyclonic clouds over the South Atlantic Ocean (on the left), and results of simulation (on the right). The cloud fraction is 0.62.

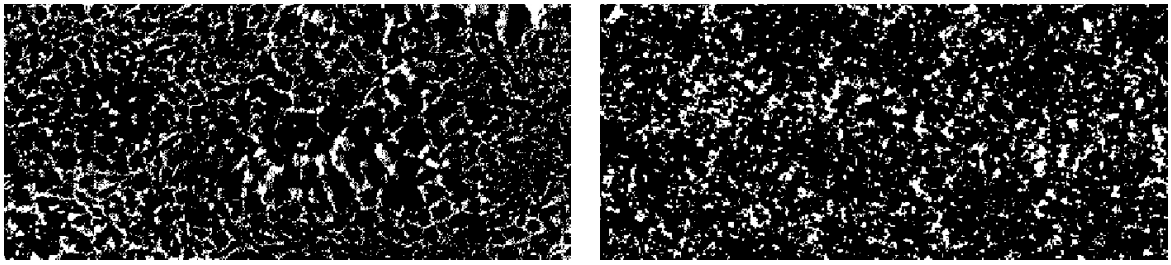


Figure 2: A satellite image of open-cell clouds (480km × 215km) off the west coast of South America (the upper image), and results of simulation (the lower image). The cloud fraction is 0.15.

Note, that the Gaussian threshold models are appropriate only for the simulation of "random" but not "regular" fields (examples of unsuccessful simulation can be found in [16]). The proposed binary models are based on the Gaussian field transformations. The Gaussian fields have a maximum entropy among the fields with fixed moments of the first and second order. This maximum entropy property, in a sense, is inherited by the threshold models. According to the simulation results, for many of random binary fields, the threshold models can generate fairly good replications reproducing the mean value and the correlation structure.

Finally, let us mention here that the described threshold method can be applied to anisotropic and, even, to non-homogeneous random fields. In this case, one should construct numerical models of the corresponding anisotropic or non-homogeneous Gaussian functions (see, for example,

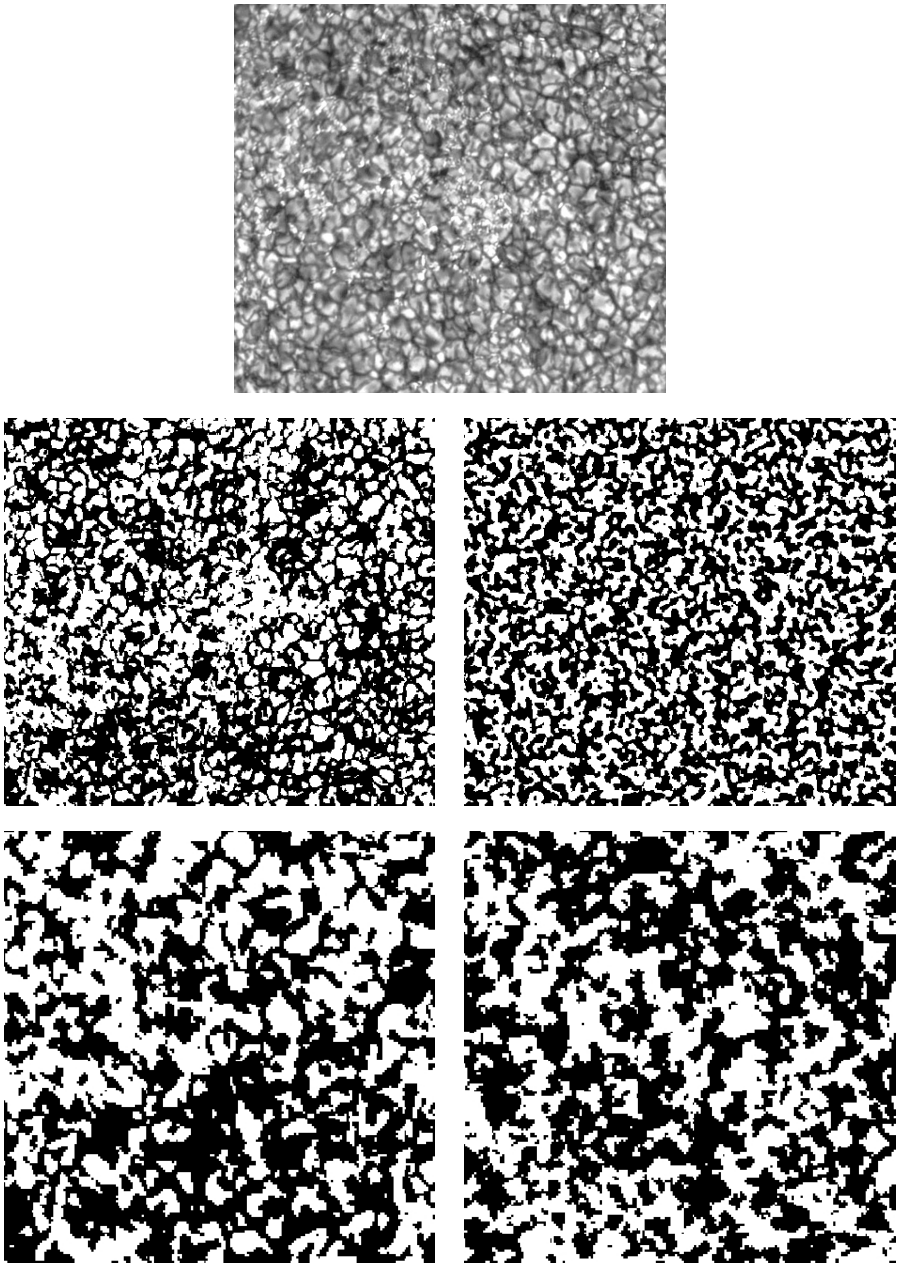


Figure 3: Granules on the photosphere of the Sun: the source (upper) image of $40000 \times 40000\text{km}^2$ area, the corresponding bitmap (the left middle), a result of its simulation (the right middle), the left lower corner of the bitmap (the left lower image), and its simulation (the right lower image)

[7, 9]). More complicated structures, like semi-binary fields (whose single-point marginal distributions are mixtures of singular and continuous distributions) and vector-valued fields, can be simulated as well [13, 14].

Acknowledgements

The author would like to thank Prof. B. A. Kargin, Prof. A. L. Marshak, Prof. G. A. Mikhailov, and Prof. V. A. Ogorodnikov for many fruitful discussions and their collaboration. In addition, the author is grateful to the Russian Foundation for Basic Research for the financial support (09-05-00963, 11-01-00641).

References

[1] Kargin B.A., Prigarin S.M. (1988). *Modelling of stochastic fields of heap clouds and study of their radiative properties with Monte Carlo method*. Preprint 817. Compt. Cent. Sib. Division, USSR Acad. Sci., Novosibirsk, pp.18 [in Russian].

[2] Kargin B.A., Prigarin S.M. (1994). Imitational simulation of cumulus clouds for studying solar radiative transfer in the atmosphere by the Monte Carlo method, *Atmospheric and Oceanic Optics*. Vol.**7**, No.9, pp.690-696.

[3] Koutsourelakis P.-S., Deodatis G. (2005). Simulation of binary random fields with applications to two-phase random media, *J. Eng. Mech.*, Vol.**131**, No.4, pp.397–412.

[4] Nott D.J., Wilson R.J. (1997). Parameter estimation for excursion set texture models. *Signal Process*. Vol.**63**, pp.199-210.

[5] Nott D.J., Wilson R.J. (2000). Multi-phase image modelling with excursion sets. *Signal Process*. Vol.**80**, pp.125-139.

[6] Ogorodnikov V.A (1989), Testing of stochastic models for series of dry and rainy days, *Proceedings ZapSibNIGMI Goskomgidrometa*, Vol.**86**, pp.74–79 [in Russian].

[7] Ogorodnikov V.A., Prigarin S.M. (1996). *Numerical Modelling of Random Processes and Fields: Algorithms and Applications*, VSP, Utrecht.

[8] Prigarin S.M. (1986). On modelling random vectors with given correlations and marginal distributions. In: *Methods of Statistical Modelling*, Comp. Cent. Sib. Division, USSR Acad. Sci., Novosibirsk, pp. 3-15 [in Russian].

[9] Prigarin S.M. (2001), *Spectral Models of Random Fields in Monte Carlo Methods*. VSP, Utrecht.

[10] Prigarin S.M. (2009), Simulation of binary textures and stochastic structure of clouds. *Webpage http* : *//smp.sscc.ru/* ∼ *smp*2/*Prigarin_CloudSim.pdf*.

[11] Prigarin S.M., Kargin B.A., Oppel U.G. (1998). Random fields of broken clouds and their associated direct solar radiation, scattered transmission, and albedo. *Pure and Applied Optics*. A., Vol. **7**, No.6, pp.1389-1402.

[12] Prigarin S.M., Marshak A.L. (2005). Numerical model of broken clouds adapted to observations. *Atmospheric and Oceanic Optics*. Vol.18 (2005), No.3, pp.236-242.

[13] Prigarin S.M., Marshak A.L. (2008), Simulation of vector semibinary homogeneous random fields and modeling of broken clouds, *Numerical Analysis and Applications*. Vol. **1**, No. 3, pp. 285–292.

[14] Prigarin S.M., Marshak A.L. (2009), A simple stochastic model for generating broken cloud optical depth and cloud top height fields. *Journal of the Atmospheric Sciences*. Vol. **66**, No. 1, pp. 92–104.

[15] Prigarin S.M., Martin A., Winkler G. (2004). Numerical models of binary random fields on the basis of thresholds of Gaussian functions. *Siberian Journal of Numerical Mathematics*. Vol. **7**, No.2, pp.165-175.

[16] Prigarin S.M., Martin A., Winkler G. (2010). Simulation of binary random fields with Gaussian numerical models. *Monte Carlo Methods and Applications*. Vol. **16**, Issue 2, pp.129-142.

[17] Prigarin S.M., Titov G.A. (1996). Spectral methods of numerical modeling of geophysical fields. *Atmospheric and Oceanic Optics*, Vol.**9**, No.7, pp.629-635.

[18] Wilson R.J.,Nott D.J. (2001). Review of recent results on excursion set models. *Image Analysis and Stereology*, Vol.**20**, No.2, pp.71-78.

Statistical Modeling Method for Kinetic Traffic Flow Model with Acceleration Variable

Aleksandr Burmistrov, Mariya Korotchenko
Institute of Computational Mathematics and Mathematical Geophysics SB RAS
Novosibirsk State University
Novosibirsk, Russia
e-mail: burm@osmf.sscc.ru, kmaria@osmf.sscc.ru

Abstract

We consider a kinetic vehicular traffic flow (VTF) model with acceleration variable and study evolution of the N-particle systems, which are governed by a homogeneous Boltzmann-like equation. For this model we obtain a linear integral equation of the second kind and suggest to solve it by the statistical modeling method. The numerical results show that the approach to simulation suggested by the authors is reasonable to apply to the vehicular traffic problems. Moreover, authors managed to exclude from simulation an external to the initial model parameter – a discrete time interval, which was used previously. It resulted in a simpler simulation process.

Keywords: N-particle system, vehicular traffic flow, Monte Carlo method.

Introduction

This paper is devoted to the study and simulation of the vehicular traffic flow (VTF). This study appears to be significant due to the constant growth of traffic in most parts of the world nowadays. It results in the necessity for improvement of the transportation network, considering the principles of its growth and distribution of load on its sections.

There are two main approaches to the VTF simulation – a deterministic and a stochastic ones. A functional relation between some parameters, such as, for example, velocity and distance between the cars in the flow, underlies the *deterministic* type of models. On the other hand, in the frame of *stochastic* models, VTF is considered as a random process. Moreover, the models describing the VTF can be further classified into three categories: micro-, macro-, and mesoscopic ones (for more details see [5]).

Mesoscopic (or *kinetic*) models, a type of models we use in our paper, consider the VTF as a random process. Moreover, these models regard the VTF as a gas, which consists of interacting particles and every particle in this gas corresponds to a car. By an interaction of two cars we understand an event when their state, determined by a number of parameters, is changed. There are two main types of interactions in the kinetic models between the cars in the system, depending on velocity of the leading car: acceleration and breaking. The possibility of overtaking is usually introduced into the model by means of a probability, depending on the density of cars on the road. The equations describing kinetic models are similar to the gas kinetic equations, in particular, to

the Boltzmann equation. However, unlike the latter one, the momentum and energy conservation laws do not hold in case of the VTF.

In this paper we develop our methods in the frame of the kinetic VTF model suggested in [8]. A distinctive feature of this model consists in introducing of the acceleration variable into the set of phase coordinates along with the space and velocity coordinates of the car. Such a modification of the phase space allowed in [8] to apply this acceleration oriented model to a wider range of the VTF types. This model adequately describes not only a constrained traffic but also a higher car density regimes.

In order to verify approach to the study and simulation of the VTF suggested in this paper further we will consider a single-lane traffic in a spatially homogeneous case without overtaking. Note that the obtained results will be compared with a known analytical solution in case of stochastic equilibrium (i. e. stationary distribution). We would like to underline that information about the equilibrium velocity can be of a great importance, for example, in planning the road capacity.

In the framework of [8], distribution of a single car with acceleration a and velocity v has the probability density $f(a, v, t)$, which solves the integro-differential equation of Boltzmann type:

$$\frac{\partial f}{\partial t}(a,v,t) + a\frac{\partial f}{\partial v}(a,v,t) = \int\limits_{\bar{a},\bar{v},a'} [\Sigma(a|a',v,\bar{a},\bar{v},\mathbf{m}_f(t))f(a',v,t) - \tag{1}$$

$$- \Sigma(a'|a,v,\bar{a},\bar{v},\mathbf{m}_f(t))f(a,v,t)]\, f(\bar{a},\bar{v},t)\, \mathtt{d}\bar{a}\, \mathtt{d}\bar{v}\, \mathtt{d}a',$$

with the initial distribution $f(a, v, 0) = f_0(a, v)$. Here $\bar{a}$ and $\bar{v}$ are the acceleration and the velocity of the leading car (*leader*), correspondingly. By a *leader* here and further on we understand the car situated straight ahead to the current car, which we will call the *follower*. It is the leader and the follower who interact. The function $\Sigma(a|a', v, \bar{a}, \bar{v}, \mathbf{m}_f(t)) = \Sigma(a' \to a|v, \bar{a}, \bar{v}, \mathbf{m}_f(t))$ is a weighted interaction rate function and it has the following form

$$\Sigma(a|a',v,\bar{a},\bar{v},\mathbf{m}_f(t)) = \int\limits_{h_{\min}}^{\infty} \sigma(a|h,a',v,\bar{a},\bar{v})Q(h,a',v,\bar{a},\bar{v})D(h|a',v,\mathbf{m}_f(t))\, \mathtt{d}h. \tag{2}$$

Here we used the notations:
$h_{\min}$ is the minimal distance between two cars at rest, (the mean length of a car);
$\sigma(\cdot)$ is the probability density of the follower's acceleration in case the interaction between the cars with states (a', v) and $(\bar{a}, \bar{v})$ takes place at distance h;
$Q(\cdot)$ is the interaction rate, it depends on a current microscopic state of the interacting car pair and the distance h between them;
$D(\cdot)$ is a conditioned probability density of the distance h. It depends on the follower's state (a', v) and a vector $\mathbf{m}_f(t)$, which value is determined by some moments of the solution f (such as mean velocity, velocity scattering, mean acceleration etc.). Further on the function $D(\cdot)$ will also depend on the car density $\mathcal{K}$.

We should note that in this model, suggested in [8], the car acceleration a is added to the phase coordinates as an independent variable in contrast to the gas dynamics. As a result of this modification there are only acceleration jumps (no velocity jumps as in other kinetic models) produced by the pairwise interactions in the system. Moreover, after the interaction takes place the leader does not change its acceleration. Therefore the function $\Sigma(\cdot)$ is not symmetric. We suggest to designate the interacting cars as ordered pairs (i, j), where the first number stands for the follower and the second one stands for the leader.

This paper aims at constructing the basic integral equation of the second kind. The latter equation will enable us to use well-developed techniques of the weight statistical modelling (see e. g. [6]) for estimating the functionals of solution to the equation (1).

1 Basic integral equation of the second kind

The simulation process of stochastic kinetics of the N-particle system is a homogeneous Markov chain in which transitions are due to elementary pair interactions. Note that we deliberately do not use a gas dynamic term *collision* because it has evidently a confusing meaning in case of the vehicular traffic flow.

The integral equation, which describes evolution of the particle (car in this case) ensemble, uniquely defines all the transition densities in the Markov chain. It means that the distribution density of time intervals between elementary interactions in the system can also be determined using this integral equation.

In order to construct the required basic integral equation of the second kind we introduce a phase space Λ of velocities and accelerations for the ensemble of N cars: $(A, V) = (a_1, v_1, \ldots, a_N, v_N) \in \Lambda$. Let us consider the distribution density of the N-particle system $P(A, V, t)$. Further we omit the dependence of the function $\Sigma(\cdot)$ on the vector $\mathbf{m}_f(t)$ without loss of generality. In this case the function $P(A, V, t)$ satisfies a master equation (see [3]) of the form

$$\frac{\partial P}{\partial t} + A\frac{\partial P}{\partial V} = \sum_{i \neq j} \int \frac{\Sigma(a_i|a_i', v_i, a_j, v_j)P(A_i', V, t) - \Sigma(a_i'|a_i, v_i, a_j, v_j)P(A, V, t)}{N-1}\, \mathtt{d}a_i', \qquad (3)$$

here $A_i' = (a_1, \ldots, a_{i-1}, a_i', a_{i+1}, \ldots, a_N)$. To complete the problem statement we add an *initial condition* $P(A, V, 0) = P_0(A, V)$ as well as *boundary conditions* to the equation (3). The latter conditions should eliminate both negative velocities and ones exceeding some maximum value $V_{\max}$: $P(A, V, t) = 0$ if there is such a number i that either condition ($v_i = 0$ and $a_i < 0$) or condition ($v_i = V_{\max}$ and $a_i > 0$) is fulfilled.

Let us now rewrite the equation (3) in the form

$$\frac{\partial P}{\partial t}(A, V, t) + A\frac{\partial P}{\partial V}(A, V, t) + \upsilon(A, V)P(A, V, t) = J_N(A, V, t), \qquad (4)$$

here we used the following notations: $J_N(A, V, t) = \int F(A' \to A|V)P(A', V, t)\, \mathtt{d}A'$,

$$F(A' \to A|V) = \tfrac{1}{N-1} \sum_{i \neq j} \Sigma(a'_i \to a_i | v_i, a_j, v_j) \cdot \left\{ \prod_{m \neq i} \delta(a'_m - a_m) \right\};$$
$$\upsilon(A, V) = \tfrac{1}{N-1} \sum_{i \neq j} \upsilon_{(i,j)}; \;\; \upsilon_{(i,j)} = \int \Sigma(a_i \to a''_i | v_i, a_j, v_j) \, \mathtt{d}a''_i.$$

Here $\delta(\cdot)$ is a Dirac delta function. Taking into account the initial conditions and parametric dependence between velocity, acceleration and time $V = V' + A(t - t')$, we can integrate the equation (4) with respect to time.

Let us consider in our system the interaction density $\Phi(A, V, t) = \upsilon(A, V) P(A, V, t)$ and the function $\Psi(A, V, t)$, for which the following integral relation holds

$$\Phi(A, V, t) = \int_0^t \int K_t(t' \to t | A, V') K_V(V' \to V | A, t - t') \Psi(A, V', t') \, \mathtt{d}V' \, \mathtt{d}t'. \tag{5}$$

Then $\Psi(A, V, t)$ satisfies the equation $\Psi = \mathbf{K}_1 \Psi + \Psi_0$ with a free term $\Psi_0(A, V, t) = \delta(t) P_0(A, V)$ and the kernel

$$K_1(A, V, t | A', V', t') = K_t(t' \to t | A', V') K_V(V' \to V | A', t - t') K_A(A' \to A | V).$$

Note that the kernel K_1 is a product of distribution densities of new values t, V, A:

$$K_t(t' \to t | A', V') = \Theta(t - t') \upsilon(A', V' + A'(t - t')) E_\upsilon(A', V', t, t'),$$
$$K_V(V' \to V | A', t - t') = \delta_V(A', V', t, t'), \;\; K_A(A' \to A | V) = \tfrac{F(A' \to A|V)}{\upsilon(A', V)},$$

here we used the following notations $E_\upsilon(A, V', t, t') = \exp\{-\int_{t'}^{t} \upsilon(A, V' + A(\tau - t'))\mathtt{d}\tau\}$, $\delta_V(A, V', t, t') = \prod_{m=1}^{N} \delta(v_m - v'_m - a_m(t - t'))$, and $\Theta(\cdot)$ is a Heaviside step function.

Thus, the transition in our Markov chain consists of several elementary transitions in the following order: $(A', V', t') \to (A', V', t) \to (A', V, t) \to \{\pi\} \to (A, V, t)$.

Let us denote $\Phi(A, V, t) = \upsilon(A, V) P(A, V, t) = \sum_\pi \frac{\upsilon(\pi)}{N-1} P(A, V, t) = \sum_\pi F_\Phi(\pi, A, V, t)$, here the summation is performed over indices $\pi = (i, j)$ of all possible ordered pairs of cars in the system. The function $F_\Phi(\cdot)$ here is related to the function $F_\Psi(\cdot)$ similarly to the formula (5).

Let us now introduce the number of an interacting pair π to the set of phase coordinates (A, V, t) of our system (see [7] for more details). Further we will consider Markov chain in this modified phase space $\mathbf{Z} \times [0, T] \ni (Z, t) = (\pi, A, V, t)$.

The initial state $Z_0 = (\pi_0, A_0, V_0)$ (i. e. the point of the first interaction in the system at $t_0 = 0$) in our modified phase space is simulated according to the distribution density $P_0(A, V) \cdot \delta(\pi_0)$. Note, that π_0 can be chosen arbitrary since it does not affect the distribution of the next interaction. The density function of the point (Z_0, t_0) is denoted by $F_0(Z, t) = \delta(t) \cdot P_0(A, V) \cdot \delta(\pi_0)$.

The mentioned above modification results in decomposition of the phase space according to the pair number π and makes it possible to derive a new basic integral equation of the second

kind for the function $F(Z,t) = F_\Psi(Z,t)$: $F = \mathbf{K}F + F_0$. We can rewrite the latter equation as follows

$$F(Z,t) = \int\limits_0^t \int\limits_{\mathbf{Z}} F(Z',t')K(Z',t' \to Z,t)\,\mathtt{d}Z'\,\mathtt{d}t' + F_0(Z,t). \tag{6}$$

Here $\mathtt{d}Z = \mathtt{d}V\,\mathtt{d}A\,\mathtt{d}\mu(\pi)$ and integration with respect to μ means the summation over all possible ordered pairs (i,j). The kernel K of the equation (6) is a product of transitional densities $K(Z',t' \to Z,t) = K_t(t' \to t|A',V') \cdot K_V(V' \to V|A',t-t') \cdot K_\pi(\pi) \cdot K_a(a_i' \to a_i|\pi = (i,j),V) \cdot \left\{ \prod\limits_{m \neq i} \delta(a_m' - a_m) \right\}$, i. e. it contains δ-functions as factors only.

Despite the presence of general functions, it is possible to treat $\mathbf{K}$ as an operator from $L_1(\mathbf{Z} \times [0,T])$ to $L_1(\mathbf{Z} \times [0,T])$ (see [6]). Moreover, due to the finiteness of T, the norm $\|\mathbf{K}\|_{L_1} < 1$ and the Neumann series $F(Z,t) = \sum\limits_{n=0}^{\infty} \mathbf{K}^n F_0(Z,t) = \sum\limits_{n=0}^{\infty} F_n(Z,t)$ for the integral equation (6) converges with respect to the L_1 norm. Note, that $F_n(Z,t)$ is a distribution density of the nth interaction in the system. This fact makes it possible to construct weight estimates using the integral equation (6) rather than the equation for the function Ψ.

The transition of the system from the state Z' to the state Z is performed as follows: 1. the instant t of the next interaction in the system is chosen according to K_t;
2. the velocities of all cars are calculated at time t according to K_V;
3. the pair number $\pi = (i,j)$ of the interacting cars is chosen by the probabilities $p(\pi) = p(i,j) = \upsilon_{(i,j)}/\upsilon(A',V)(N-1)$;
4. new accelerations a_i of the car with number i is changed according to $K_a(a_i' \to a_i|\pi,V) = \Sigma(a_i' \to a_i|v_i,a_j,v_j)/\upsilon_{(i,j)}$; the accelerations of other cars do not change.

2 Estimation of functionals

Usually when solving the equation (1) the functionals $I_{\mathbf{h}}(T)$ of one-particle distribution function f in following form

$$I_{\mathbf{h}}(T) = \int\int \mathbf{h}(a_1,v_1)f(a_1,v_1,T)\,\mathtt{d}a_1\,\mathtt{d}v_1 = \int\limits_{\Lambda} \mathbf{h}(a_1,v_1)P(A,V,T)\,\mathtt{d}A\,\mathtt{d}V$$

are of interest. Let us denote $\mathbf{H}(A,V) = \frac{1}{N}\sum\limits_{i=1}^{N} \mathbf{h}(a_i,v_i)$. Then, by analogy with [7], we use the relation between the functions P, Ψ, F and obtain a formula for the functional $I_{\mathbf{h}}(T)$ of solution to the equation (6):

$$I_{\mathbf{h}}(T) = \int\limits_{\mathbf{Z}} \int\limits_0^T \mathbf{H}(A,V + A(T-t'))E_v(A,V,T,t')F(Z,t')\,\mathtt{d}Z\,\mathtt{d}t'.$$

Since we have at our disposal an integral equation of the second kind and a Markov chain corresponding to it, we can apply a well-developed techniques of weight statistical simulation (see [6], e. g.). This enables us to study dependence of our model on various parameters, estimate parametric derivatives and reduce computational costs of statistical methods (e. g. with the help of the majorant frequency principle [2] and the value modelling algorithms [4]).

Let us introduce a Markov chain $\{Z_n, t_n\}$, $n = 0, 1..., \kappa$, where κ is the number of interaction preceding the passage of the system beyond the time boundary T, with the normalized transition density $P(Z', t' \to Z, t)$:

$$P = P_1(t|A', V', t')P_2(V|A', V', t)P_3(\pi|A', V, t)P_4(a_i|\pi, A', V, t) \cdot \left\{ \prod_{m \neq i} \delta(a'_m - a_m) \right\},$$

and the normalized distribution density $P^{(0)}(A, V)\delta(t)\delta(\pi_0)$ of the initial state (Z_0, t_0). We define random weights Q_n by the formulas

$$Q_0 = \tfrac{P_0(A_0,V_0)}{P^{(0)}(A_0,V_0)}, \;\; Q_n = Q_{n-1}Q(Z_{n-1}, t_{n-1}; Z_n, t_n),$$
$$Q(Z', t'; Z, t) = \left\{ \tfrac{K_t(t' \to t|A',V')}{P_1(t|A',V',t')} \right\} \left\{ \tfrac{K_V(V' \to V|A',t-t')}{P_2(V|A',V',t)} \right\} \left\{ \tfrac{K_\pi(\pi)}{P_3(\pi|A',V,t)} \right\} \left\{ \tfrac{K_a(a'_i \to a_i|\pi=(i,j),V}{P_4(a_i|\pi,A',V,t)} \right\}.$$

For numerical estimation of the functional $I_{\mathbf{h}}(T)$ we can use the collision estimator ξ or absorption estimator η, which are functionals of the Markov chain trajectory [6, 7]:

$$\xi = \sum_{n=0}^{\kappa} Q_n \tilde{\mathbf{H}}(A_n, V_n, T - t_n), \;\; \eta = \tfrac{Q_\kappa \tilde{\mathbf{H}}(A_\kappa, V_\kappa, T - t_\kappa)}{q(A_\kappa, V_\kappa, t_\kappa)}, \;\; q(A, V, t') = 1 - \int_0^{T-t'} P_1(\tau|A, V, t')\, \mathtt{d}\tau.$$

Theorem 1 ([7]). *If $P^{(0)}(A, V) \neq 0$ for $P_0(A, V) \neq 0$; and $Q(Z', t'; Z, t) < +\infty$ for $Z', Z \in \mathbf{Z}$, $t', t < T$, then $\mathbf{E}\xi = I_{\mathbf{h}}(T)$. If, additionally, $q(A, V, t') > 0$ for $(A, V) \in \Lambda$ and $t' < T$, then $\mathbf{E}\eta = I_{\mathbf{h}}(T)$. Moreover, if the weights Q_n are uniformly bounded and $\mathbf{H} \in L_\infty$, then there exists such T^* that $\mathbf{V}\xi < +\infty$ and $\mathbf{V}\eta < +\infty$ for $T < T^*$.*

3 Numerical results

The numerical results in this section show the efficiency of transition to the basic integral equation in VTF problems with *distance threshold* interaction. Moreover, we succeeded not to use in the simulation procedure an external discrete time parameter. In [1] we successfully tested two types of interactions (*maxwellian* and *hard sphere*) for a spatially homogeneous nearly free stationary VTF. In this paper as a test for the algorithm described at the end of the Sect. 1 we consider a distance oriented interaction model [9] with the following parameters: $Q = \upsilon(\pi) = 1/\mathcal{T}$,

$$D(h) = \tfrac{1}{\bar{H} - h_{\min}} \exp\left\{ -\tfrac{h - h_{\min}}{\bar{H} - h_{\min}} \right\} \Theta(h - h_{\min}), \;\; \bar{H} = 1/\mathcal{K},$$
$$\sigma(a|h, v) = \Theta(h - H(v)) \cdot \delta(a - a_0) + \Theta(H(v) - h) \cdot \delta(a + a_0)$$

and a simple distance interaction threshold $H(v) = \alpha \cdot v + h_{\min}$. Taking these functions into account we find the form of the weighted interaction density (2)

$$\Sigma(a_i' \to a_i | v_i, a_j, v_j) = \tfrac{p}{\mathcal{T}}\delta(a + a_0) + \tfrac{(1-p)}{\mathcal{T}}\delta(a - a_0), \;\; p = \int\limits_{h_{\min}}^{H(v_i)} D(h)\, \mathtt{d}h.$$

For such coefficient $\Sigma(\cdot)$ the solution to the equation (1) is given by (see [9]):

$$f(v) = \tfrac{1}{a_0\mathcal{T}} \exp\left\{-\tfrac{v}{a_0\mathcal{T}} - 2\beta \mathrm{e}^{-\frac{v}{a_0\mathcal{T}\beta}}\right\} \left(\mathrm{e}^{-2\beta} + \beta(2\beta)^{-\beta}\gamma(\beta, 2\beta)\right)^{-1}, \;\; \beta = \tfrac{\bar{H} - h_{\min}}{\alpha a_0 \mathcal{T}},$$

here $\gamma(\beta, 2\beta)$ is an incomplete gamma function.

Velocity distribution. First we estimate the velocity distribution in the flow. In this case we choose functions $\mathbf{h}(a, v)$ equal to indicators of some partitioning of the velocity interval $0 \le v_i \le V_{max} = 40$ m/s. As an initial velocity distribution we use a mixture of two normal distributions with the means $\mathrm{V}_1 = 15$ m/s, $\mathrm{V}_2 = 25$ m/s and the variance $\sigma_0 = 1$ m/s. Initial accelerations are equal to 0. We simulated $M = 10^3$ trajectories of our system consisting of $N = 10^3$ cars. The numerical estimate for evolution of the velocity distribution is shown in Fig. 1 ($\mathcal{K} = 0.025$ m^{-1}, $\mathcal{T} = 2.5$ s, $h_{\min} = 6.5$ m, $a_0 = 0.3$ m/s^2, $\alpha = 1.2$ s).

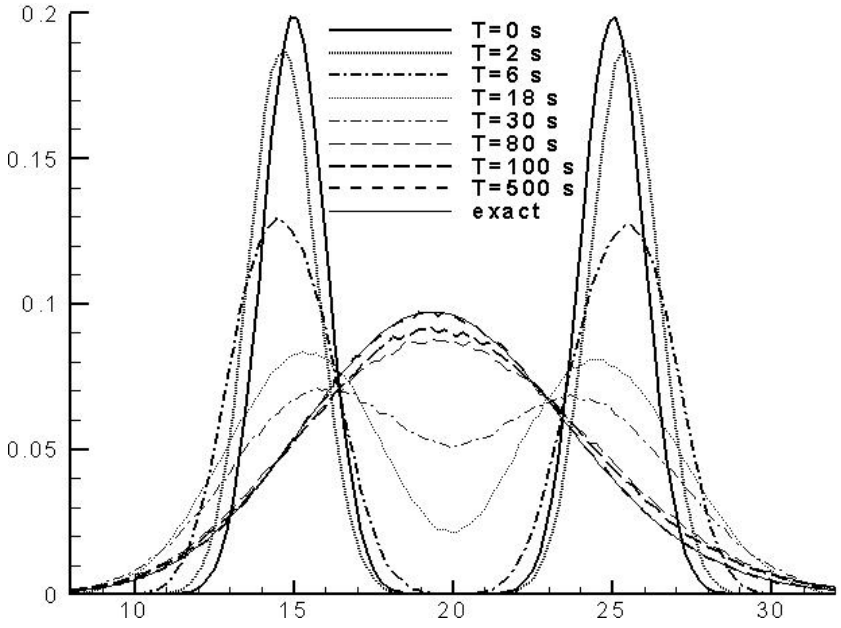


Figure 1: Numerical estimate for evolution of the velocity distribution $f(v, T)$.

Fundamental diagram. Next we consider a numerical estimation of the traffic density $\mathcal{K}$V dependence (here V is the mean velocity which is estimated with the help of corresponding function $\mathbf{h}(a, v) = v$) on the car density $\mathcal{K}$ which is called a *fundamental diagram*. Fig. 2 shows a typical shape of this curve for the following parameters: $\mathcal{T} = 2.5$ s, $h_{\min} = 6.5$ m, $a_0 = 0.1$ m/s^2, $\alpha_1 = 1.2$ s, $\alpha_2 = 1.5$ s, $\alpha_3 = 1.8$ s. For some value of $\mathcal{K}$ there is a change from a free flow (with no dependence on α) to an interaction oriented flow (with strong dependence on α). For the latter flow cars can not drive in their own way, but they should agree their velocity with the flow velocity. Note that low values of α correspond to a more aggressive driver, while high values of

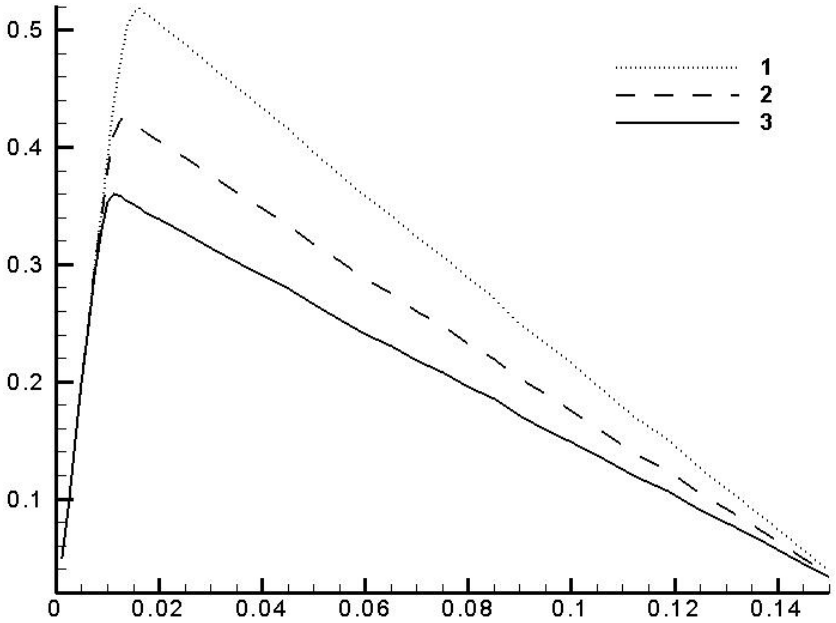


Figure 2: Fundamental diagram ($M = 10^2$, $N = 10^2$): 1 – α_1, 2 – α_2, 3 – α_3.

this parameter stand for a more conservative driving manner.

Acknowledgements

The authors acknowledge the kind hospitality of the Novosibirsk State Technical University and the AMSA'2011 conference organizers.

This work was partly supported by Russian Foundation for Basic Research (grants 09-01-00035, 09-01-00639, 11-01-00252) and SB RAS (Integration Grant № 22).

References

[1] Burmistrov A.V., Korotchenko M.A. (2010). Application of statistical methods for investigation of the kinetic traffic flow. *Proceedings of the ICM&MG Young Scientists Conference*,

Novosibirsk, pp. 22-31. [In Russian]

[2] Ivanov M.S., Rogasinsky S.V. (1988). *The Direct Statistical Simulation Method in Dilute Gas Dynamics*. Publ. Comput. Center, Sib. Branch, USSR Acad. Sci., Novosibirsk. [In Russian]

[3] Kac M. (1959). *Probability and Related Topics in Physical Sciences*. Interscience, New York.

[4] Korotchenko M.A., Mikhailov G.A., Rogasinsky S.V. (2007). Value modifications of weighted statistical modeling for solving nonlinear kinetic equations. *Russ. J. Numer. Anal. Math. Modelling*. Vol. **22**, pp. 471-486.

[5] Mahnke R., Kaupužs J., Lubashevsky I. (2005). Probabilistic description of traffic flow. *Phys. Rep*. Vol. **408**, pp. 1-130.

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

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

[8] Waldeer K.T. (2003). The direct simulation Monte Carlo method applied to a Boltzmann-like vehicular traffic flow model. *Comput. Phys. Commun*. Vol. **156**, pp. 1-12.

[9] Waldeer K.T. (2004). A vehicular traffic flow model based on a stochastic acceleration process. *Transp. Theory Stat. Phys*. Vol. **33**, pp. 7-30.

Stochastic Models of the Price Series for Trade Algorithms

Mikhail A. Yakunin
Institute of Computational Mathematics and Mathematical Geophysics SB RAS
Novosibirsk, Russia
e-mail: yma@osmf.sscc.ru

Abstract

We investigate stochastic models of series of price increments for trade algorithms. The models are constracted based on stochastic differential equations. The estimates of unknown parameters of the model of price increments with jumps are obtained by using the method of moments.

Keywords: price increments, probability density, estimates of parameters.

Introduction

Price series models in the form of stochastic differential equations (SDEs) are widely used in financial mathematics because such models are flexible and are able to reproduce various hypothetical market situations with a proper choice of parameters. A simple model of price increments of a financial instrument is obtained by solving a linear SDE with additive noise on a uniform time grid. This model is generalized to price series with jumps on random intervals between these. In real stock exchange, such jumps are observed, for instance, with the opening of some trade sessions, or upon receiving unexpected pieces of news of economic or political character. Such model of price increments is an approximation to the solution of a linear SDE with a Poisson component on a uniform time grid [2]. Estimates of the unknown parameters of this model are computed based on observations of price increments with simple formulas obtained by the method of moments.

Price series models are used in the theory of trade algorithms. The main characteristic of a trade algorithm is its total profitability. The complexity of analysis of this random sequence is connected with unknown distribution laws for its two components: the number of buy-sell transactions per given number of the algorithm steps and the profitability of the algorithm at each step. Under the assumption of the stationary nature and m-dependence of a sequence of price increments, it is possible to establish the asymptotic normality of the total profitability of the trade algorithm. The usage of a stationary model for a series of price increments allows one to reduce the analysis of the mean of the number of transactions to analysis of the constant probability of transaction closing [1].

1 Analysis of the price series model

The easy-to-use stochastic model of a price series is given in a recurrent form [1]

$$P_{n+1} = P_n \exp(r_{n+1}),$$

where

$$r_{n+1} = \rho r_n + \sigma\sqrt{1-\rho^2}\xi_{n+1}, \quad n = 0, 1, 2, \ldots, \quad r_0 = \sigma\xi_0. \tag{1}$$

Here ξ_{n+1} are mutually independent and independent of r_k, $k \le n$, random values having the same symmetric distribution with zero mean and unit variance, ρ is the correlation coefficient of two consecutive price increments, σ is the price volatility, $|\rho| < 1$, $\sigma > 0$.

The Markov sequence of price increments given according to (1) is homogeneous with zero mean and constant variance σ^2. If random values r_n have the same one-dimensional distribution, it is stationary. In this case $f_r(x) \equiv (1/\sigma)\ f_\xi(x/\sigma)$, where $f_r(x)$ and $f_\xi(x)$ are the probability densities of the random values r_n and ξ_n, respectively. The random sequence $\{r_n\}$ is stationary if $f_\xi(x)$ is a normal probability density. The normal random sequence (1) for $\rho > 0$ on a uniform time grid with step h is the solution of a linear SDE with additive noise

$$dr(t) = \big((\ln\rho)/h\big)r(t)dt + \sqrt{-2(\ln\rho)/h}\,\sigma\,dw(t), \quad r(0) = r_0.$$

The following theorem is valid for stationary random sequence (1).

Theorem. *Let ξ_n, $n = 0, 1, 2, \ldots$, be mutually independent continuous random values having the same symmetric distribution with zero mean, unit variance, and finite $\mathbf{E}|\xi_n|^{2+\delta}$ for some $\delta > 0$. If random values r_n defined according to (1) are equally distributed for all n and all $\rho \in [0, 1)$, then the probability density $f_\xi(x)$ is normal.*

Proof. Let there exists a nonnormal probability density $f_\xi(x)$ where the random values r_n satisfy the requirements of the theorem. Form the recurrent sequence

$$s_0 = \sigma\xi_0, \quad s_1 = \frac{1}{\sqrt{2}}s_0 + \frac{\sigma}{\sqrt{2}}\xi_1, \ldots, s_n = \frac{\sqrt{n}}{\sqrt{n+1}}s_{n-1} + \frac{\sigma}{\sqrt{n+1}}\xi_n, \ldots. \tag{2}$$

Since random sequence (2) is obtained from (1) by replacing ρ in the latter one by $\rho_n = \sqrt{n}/\sqrt{n+1}$ for each n, then s_n and r_n are equally distributed, i.e., $f_s(x) \equiv f_r(x)$ independently of n. On the other hand, (2) is reduced to the form

$$s_n = \tfrac{\sigma}{\sqrt{n+1}} \sum_{i=0}^{n} \xi_i$$

and by the central limit theorem the probability density $f_s(x)$ converges for $n \to \infty$ to the normal one with the parameters $(0,\ \sigma^2)$. The contradiction obtained shows that the probability density $f_\xi(x)$ can be only normal. Theorem is proved.

Remark. Theorem does not state that there are no random values r_n with the probability density $f_r(x)$ being nonnormal and the same for all n and some particular values of the parameter $\rho \ne 0$ in (1); however symmetric distributions convenient for practical usage in a price series model do not satisfy this requirement. For example, for the Laplace distribution with $\sigma = 1$ we have

$$f_\xi(x) = \tfrac{1}{\sqrt{2}}\, e^{-\sqrt{2}|x|}, \quad -\infty < x < \infty$$

and the random value r_1 has the probability density

$$f_{r_1}(x) = \tfrac{1}{\sqrt{2}(\mu^2-\rho^2)} \left(\mu\, e^{-\frac{\sqrt{2}}{\mu}|x|} - |\rho|\, e^{-\frac{\sqrt{2}}{|\rho|}|x|}\right)$$

where $\mu = \sqrt{1-\rho^2}$, which does not relate to the Laplace distribution for any values of ρ, except for $\rho = 0$. The same is valid for the uniform distribution whose convolution, for example, for $\rho = \mu = 1/\sqrt{2}$ gives a Simpson (triangular) distribution.

For stationary Markov sequence (1) the probability density $f_\xi(x)$ has to satisfy the integral equation

$$f_\xi(y) = \tfrac{1}{\mu} \int\limits_{-\infty}^{\infty} f_\xi\left(\tfrac{y-\rho x}{\mu}\right) f_\xi(x)dx.$$

This is valid for the normal distribution of random values ξ_n.

2 Price series model with jumps

Suppose that the price increment undergoes jumps at random times. The time interval between neighbouring jumps is distributed in accordance with an exponential law with intensity λ

$$f(\tau) = \lambda e^{-\lambda\tau}, \quad \tau \geq 0, \tag{3}$$

and the value of jump is distributed in accordance with a normal law with zero mean and variance s^2. Then instead of stationary normal random sequence (1) we obtain

$$r_{n+1} = \rho r_n + \sqrt{1-\rho^2}\big(\sigma\xi_{n+1} + sI_{n+1}\eta_{n+1}\big), \quad n = 0, 1, 2, \ldots, \quad r_0 = \sqrt{\sigma^2 + ps^2}\xi_0. \tag{4}$$

Here $\{\xi_k\}$ and $\{\eta_k\}$ are sequences of mutually independent standard normal random values, I_{n+1} are random indicators of jumps independent of each other, and also of r_i, $i \leq n$, $\{\xi_k\}$, and $\{\eta_k\}$. We have $\mathbf{E}I_{n+1} \equiv p$, where p is the probability of a jump on the half-interval $(t_n,\ t_{n+1}]$, and in view of (3), $p = 1 - e^{-\lambda h}$. The model of price increments specified in (4) according to [2] is an approximation to the solution of the linear SDE with a Poisson component. Specifically, between neighbouring nodes t_n and $t_{n+1} = t_n + h$ on a given uniform time grid, there may appear at most one jump, and that jump is fixed at t_{n+1}. A random sequence of form (4) is a homogeneous Markov sequence with zero mean and constant variance $D_r = \sigma^2 + ps^2$.

Now we determine a transition probability density for the Markov sequence in (4). Consider the following random values:

$$\beta_n = \sigma\sqrt{1-\rho^2}\xi_n, \quad \gamma_n = s\sqrt{1-\rho^2}I_n\eta_n.$$

The probability density $f_\beta(x)$ of the random value β_n is normal. The random value γ_n is normal with probability p and is zero with probability $1-p$. For the probability density of γ_n, therefore, we have

$$f_\gamma(x) = p\ f_s(x) + (1-p)\ \delta(x),$$

where

$$f_s(x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}s} e^{-x^2/(2(1-\rho^2)s^2)},$$

and $\delta(x)$ is the delta function. The probability density of a sum of independent random values, $\beta_n + \gamma_n$, is determined by the convolution of the densities

$$f_{\beta+\gamma}(x) = p \int_{-\infty}^{\infty} f_s(\tau) f_\beta(x-\tau)\, d\tau + (1-p) \int_{-\infty}^{\infty} \delta(\tau) f_\beta(x-\tau)\, d\tau = p\ f_1(x) + (1-p)\ f_\beta(x), \quad (5)$$

where

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}}, \quad f_\beta(x) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{x^2}{2\sigma_2^2}},$$
$$\sigma_1^2 = \sigma_2^2 + (1-\rho^2)s^2, \quad \sigma_2^2 = (1-\rho^2)\sigma^2. \quad (6)$$

Thus, a distribution of the sum $\beta_n + \gamma_n$ of random values is a mixture of normal distributions. For the transition density of sequence (4), in view of (5), we obtain

$$f(y|x) = p\ f_1(y-\rho x) + (1-p)\ f_\beta(y-\rho x). \quad (7)$$

3 Estimation of parameters of the price series model

A multidimensional probability density of the Markov sequence $\{r_0, r_1, \ldots, r_N\}$ is the following:

$$f(y_0, y_1, \ldots, y_N) = f_0(y_0) \prod_{n=0}^{N-1} f(y_{n+1} \mid y_n), \quad (8)$$

where $f_0(y_0)$ is the probability density of a random value r_0, and the transition density $f(y_{n+1} \mid y_n)$ is determined in accordance with (7). A maximum likelihood estimate with observations $\{r_0, r_1, \ldots, r_N\}$ for the parameter vector $\Theta = (\rho, \sigma_1^2, \sigma_2^2, p)^T$ is that value of the vector at which the likelihood function $L(\Theta) = \ln f(r_0, r_1, \ldots, r_N)$ reaches maximum. Assuming that the probability density $f_0(y_0)$ is independent of Θ, from (7) and (8) we obtain

$$L_1(\Theta) = \sum_{n=0}^{N-1} \ln\big(p f_1(r_{n+1} - \rho r_n) + (1-p) f_\beta(r_{n+1} - \rho r_n)\big). \quad (9)$$

To find the maximum likelihood estimate for the vector of parameters Θ, the four-variable function in (9) must be maximized and inspected for having a global maximum, which is rather difficult and does not generally lead to satisfactory results. Even if there is only one unknown parameter

p, an algebraic equation $dL_1/dp = 0$ of order $N-1$ must be solved for p. The method of approximate estimation in this case involves finding an estimate that is asymptotically equivalent to a maximum likelihood one [3].

Thus, another more convenient method should be used to estimate unknown parameters of model (4). For this problem, the method of moments can be applied. For the parameter ρ in (4), we have $\mathbf{E}r_{n+1}r_n = \rho\mathbf{E}r_n^2$ for any n. Replacing the mathematical expectations in this equality by suitable sample moments in one realization of the random sequence $\{r_n\}$, we obtain the estimate

$$\widehat{\rho} = \sum_{n=0}^{N-1} r_{n+1}r_n \Big/ \sum_{n=0}^{N-1} r_n^2. \tag{10}$$

With independent random values $r_{n+1} - \rho r_n$, in accordance with their probability density given by (5), for the second, fourth, and sixth moments we have

$$\begin{aligned} \mathbf{E}(r_{n+1} - \rho r_n)^2 &= p\sigma_1^2 + (1-p)\sigma_2^2, \\ \mathbf{E}(r_{n+1} - \rho r_n)^4 &= 3(p\sigma_1^4 + (1-p)\sigma_2^4), \\ \mathbf{E}(r_{n+1} - \rho r_n)^6 &= 15(p\sigma_1^6 + (1-p)\sigma_2^6). \end{aligned} \tag{11}$$

In the same way as we did in deriving estimate (10), from (11) we obtain the following system for the three remaining unknown parameters:

$$p\sigma_1^2 + (1-p)\sigma_2^2 = S_1, \tag{12}$$

$$p\sigma_1^4 + (1-p)\sigma_2^4 = S_2, \tag{13}$$

$$p\sigma_1^6 + (1-p)\sigma_2^6 = S_3, \tag{14}$$

where

$$S_1 = \tfrac{1}{N}\sum_{n=0}^{N-1}(r_{n+1} - \widehat{\rho} r_n)^2, \quad S_2 = \tfrac{1}{3N}\sum_{n=0}^{N-1}(r_{n+1} - \widehat{\rho} r_n)^4, \quad S_3 = \tfrac{1}{15N}\sum_{n=0}^{N-1}(r_{n+1} - \widehat{\rho} r_n)^6.$$

Substituting p from (12) into equations (13) and (14), after transformations, we arrive at the following system of two equations:

$$S_1(\sigma_1^2 + \sigma_2^2) - \sigma_1^2\sigma_2^2 = S_2, \quad \sigma_1^2\sigma_2^2 = (S_1S_3 - S_2^2)/(S_2 - S_1^2).$$

Its solution is

$$\widehat{\sigma}_{1,2}^2 = \frac{S_3 - S_1S_2 \pm \sqrt{S_3^2 + 4S_2^3 + 4S_1^3S_3 - 3S_1^2S_2^2 - 6S_1S_2S_3}}{2(S_2 - S_1^2)}, \tag{15}$$

where "+" is chosen for $\widehat{\sigma}_1^2$ since $\sigma_1^2 > \sigma_2^2$ in accordance with (6). In view of (15), from (12) and formulas (6) we derive estimates

$$\widehat{p} = \frac{S_1 - \widehat{\sigma}_2^2}{\widehat{\sigma}_1^2 - \widehat{\sigma}_2^2}, \quad \widehat{\sigma} = \frac{\widehat{\sigma}_2}{\sqrt{1 - \widehat{\rho}^2}}, \quad \widehat{s} = \frac{\sqrt{\widehat{\sigma}_1^2 - \widehat{\sigma}_2^2}}{\sqrt{1 - \widehat{\rho}^2}}, \tag{16}$$

and $\widehat{\lambda} = -(1/h)\ln(1-\widehat{p})$.

Remark. For the price increment model without jumps in (1) (i.e., with $p = 0$), (12) immediately implies $\widehat{\sigma}^2 = S_1/(1-\widehat{\rho}^2)$. In this case the estimates $\widehat{\sigma}$ and $\widehat{\rho}$ in (10) are also maximum likelihood estimates, which maximize function (9) for variables σ and ρ.

The parameter estimates were computed based on observations for one trajectory of price increments by the statistical simulation algorithm presented in [2]. An example of such calculations is presented below. The number of observations on a uniform time grid is $N = 10^5$, which corresponds to about 10 months in real stock exchange. A time grid step is $h = 1$(minute). Parameter values are

$$\rho = 0.1, \quad \sigma = 0.001, \quad s = 0.01, \quad p = 0.05.$$

Parameter estimates by formulas (10) and (16) are

$$\widehat{\rho} = 0.099168, \quad \widehat{\sigma} = 0.0011042, \quad \widehat{s} = 0.010289, \quad \widehat{p} = 0.046480.$$

Under the decreasing values of the jump probability p and the ratio s/σ, estimation errors of the parameters p and s increase.

In the future, we intend to investigate properties of the above parameter estimates.

Acknowledgements

Author thanks T.A. Averina for computation of parameter estimates by her statistical simulation algorithm.

References

[1] Artemiev S.S., Yakunin M.A. (2008). Analysis of the total profitability asymptotic distribution for a trade algorithm. *Russ. J. Numer. Anal. Math. Modelling*. Vol. **23**, No. 1, pp. 1-19.

[2] Averina T.A., Yakunin M.A. (2009). Parameter Estimation for a Price Series Model by Solving a Linear SDE with a Poisson Component. *Numerical Analysis and Applications*. Vol. **2**, No. 2, pp. 99-105.

[3] Borovkov A.A. (1984). *Mathematical Statistics*. Nauka, Moscow. (in Russian)

Monte Carlo Modeling the Radiation Heat Transfer with Temperature Correction

Sergei A. Brednikhin, Boris A. Kargin
Institute of computational Mathematics and Mathematical Geophysics SB RAS
Novosibirsk State University
Novosibirsk, Russia
e-mail: bkargin@osmf.sscc.ru

Abstract

In the paper the results of modeling of radiation heat transfer in systems of dust protoplanetary clouds with the Monte Carlo method using an algorithm of temperature correction [1] with the NMC code are presented. The results of a series of standard calculations are presented. Also within the bounds of the paper an attempt is made to apply the mentioned modeling method with use of external iterations to planet atmospheres by means of introducing a gas component into the modeled system. The results of calculations of altitude distribution of temperature in a simplified model of Venus' atmosphere without taking into account convection and reradiation with the gas component are presented. The obtained results show a qualitative correspondence with experimental data.

Keywords: Radiation heat transfer, dust circumstellar clouds, Venus, Monte Carlo method.

The problem of modeling radiation heat transfer in dust and gas-dust media in the presence of an external radiation source is one of the main problems in modeling the radiation balance in stellar and planetary atmospheres and dust circumstellar clouds. The Monte Carlo method has an important place among numerical methods use for solving this problem. This is due to its ease of use in systems with geometry different from spherical or flat. In the last two decades several algorithms were introduced for the solution of the problem of temperature balance in circumstellar dust clouds with the Monte Carlo method. Earlier versions of the algorithms required an iteration approach and, as a consequence, multiple modeling of trajectories. An algorithm proposed later in the paper [2] reduced the iterative process to solving balance equations. In 2001 the so-called algorithm of temperature correction was proposed [1], which requires neither solving of the balance equation nor external iterations for media, optical properties of which are not dependent on temperature. In the current paper an attempt is undertaken to implement this algorithm for solving the problem of radiation balance in a planet atmosphere.

Modeling was conducted with the NMC code, based on which an algorithm of temperature correction was realized and validation was carried out on the example of standard calculations. The application of this algorithm without significant changes to the case of a dense atmosphere required certain nonphysical assumptions which led to a considerable difference of obtained results from experimental data; however this did not prevent the obtainment of qualitatively correct properties of the temperature dependence.

Let us briefly discuss the application of the algorithm of temperature correction to a system with luminosity Lin a unit time interval Δt, surrounded with a dust medium with absorption coefficient $k_\nu^{abs}$and scattering coefficient k_ν^{sca} for a unit of mass. Let us model the source's radiation in the form of N_γ parcels such that the energy of a single parcel will be $E_\nu = L\Delta t/N_\gamma$. The frequency spacing of a parcel d_ν is chosen according to a distribution density corresponding to the source's spectrum. Let's say that as a result of modeling in the volume i, n_i parcels were absorbed so that the full absorbed energy in this volume is $E_{abs} = n_i E_\gamma$. Supposing that the medium in a volume i is in a local thermodynamic equilibrium and is radiation according to the law

$$j_\nu = k_\nu^{abs} \rho B_\nu (T) \tag{1}$$

where $B_\nu (T)$is the Planck function, the expression for the amount of radiated energy may be written out as

$$E_i^{em} = 4\pi\Delta t \int dV_i \int \rho k_\nu^{abs} B_\nu (T)\, d\nu = 4\Delta t \int k_p (T)\, \sigma T^4 \rho dV_i \tag{2}$$

where ρ is the density of matter in the volume i and $k_p (T) = \pi \int k_\nu^{abs} B_\nu (T) d\nu / \sigma T^4$. Choosing the volume i sufficiently small that in it the temperature T_i may be considered constant, we have

$$E_i^{em} = 4\Delta t k_P (T_i)\, \sigma T_i^4 m_i, \tag{3}$$

where m_i is the mass of matter in a volume. Equating the absorbed and radiated energies we obtain an equation for the temperature of the volume i after absorption

$$\sigma T_i^4 = \frac{n_i L}{4 N_\gamma k_P (T_i)\, m_i}. \tag{4}$$

The value $k_P (T_i)\, \sigma T_i^4$ grows steadily with the growth of T_i. The solving of equation (4) at each collision event presents a considerably labor-intensive problem, however with k_ν^{abs} not dependent on temperature it's possible to tabulate the solutions of the mentioned equation for different values of n_i/N_γ, m_i , k_ν^{abs} using multi-dimensional interpolation directly during the computation process.

Knowing the temperature after an absorption event we have to determine the probability density according to which we shall choose the frequency of a reradiated portion of energy. The frequency is defined by the difference in energy fluxes emitted by a cell before and after an absorption event

$$\Delta j_\nu = j_\nu - j'_\nu = k_\nu^{abs} \left(B_\nu (T_i) - B_\nu (T_i - \Delta T) \right), \tag{5}$$

where $\Delta T$is the temperature increase as a result of absorption. Choose E_γ sufficiently small so that the temperature increase is marginal, then

$$\Delta j_\nu \approx k_\nu^{abs} \Delta T \frac{dB_\nu}{dT}. \tag{6}$$

Normalizing to a unit, we have

$$\frac{dP_i}{d\nu} = \frac{k_\nu}{K}\left(\frac{dB_\nu}{dT}\right)_{T=T_i}, \tag{7}$$

where $dP_i/d\nu$is the probability to emit a parcel with frequency in the interval $[\nu, \nu + d\nu]$, and $K = \int_0^\infty k_\nu^{abs}\,(dB_\nu/dT)\,d\nu$is the normalizing coefficient. In the process of calculation the initial temperature distribution in small volumes $\{T_i\}$iteratively relaxes to the sought for distribution. The initial values of $\{T_i\}$must be nonnegative, there are also certain restrictions on the difference of the initial distribution from the final for a given value of N_γ, the mentioned restrictions are considered thoroughly in the work [4]. An algorithm was realized using the NMC code. Validation of the algorithm was conducted (fig. 1) on benchmarks proposed in the works [1,3].

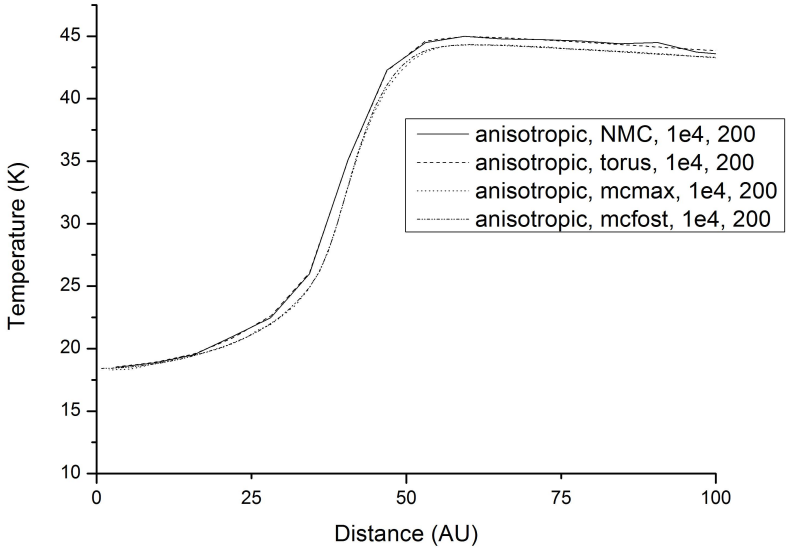


Figure 1: The result of calculation of vertical temperature structure at a distance of $R = 200AU$in a system with optical thickness $\tau = 10^4$, benchmark P04, work [3].

When the calculations were conducted the assumptions were stipulated that a sufficiently small volume is an atmosphere is in a local thermodynamic equilibrium and that in the volume absorption is made by gas and aerosol and emission – only by aerosol. Convection and Rayleigh scattering are also not taken into account. It's necessary to note, however, that a similar algorithm for solving of the problem may be constructed for systems not conforming to mentioned restrictions, but it will require solving a dynamical problem and a considerable complication of

the process of solving the equation (4). Apart from this the conditions of energy balance will no longer take place in independent small volumes, but will be true for the system in whole.

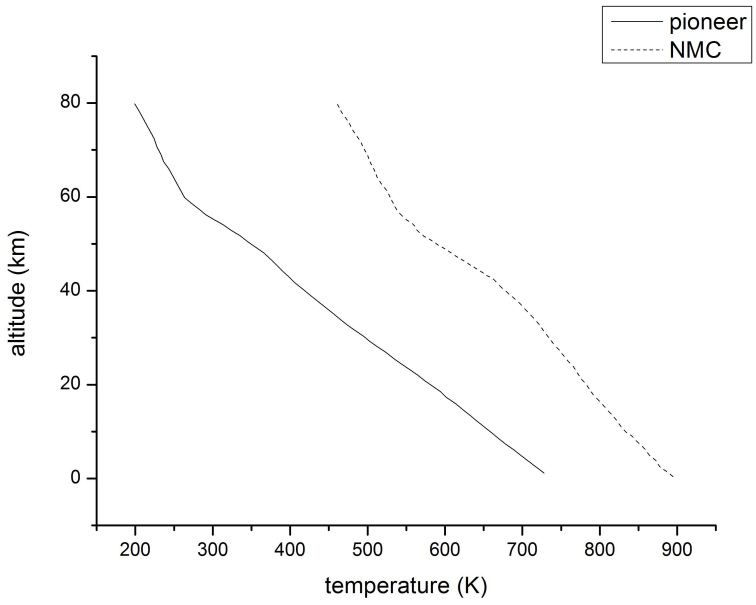


Figure 2: The result of calculating the vertical profile of temperature. The measurement results of the pioneer probe are shown in solid line.

The atmosphere is represented in the form of a hemisphere layer 90 km thick with a given pressure profile illuminated from the pole by a radiation flux with a spectrum corresponding to the Planck spectrum at a temperature of 5780 K. The atmosphere consists of gases and aerosol. The modeling algorithm is supplemented with an external iteration by temperature. It's connected with the fact that the attenuation coefficient in gas is strongly dependent on temperature and after each calculation of temperature we have to conduct a re-calculation of the attenuation coefficient in gas and concentration of gas molecules. Using formally the dependence of $k_\nu^{abs}(T_i)$ may in some cases lead to divergence while solving the equation (4) as well as while modeling the free path length. Apart from external iterations the character of absorption of energy in atmosphere layers also changes. A continuous gas absorption is introduced, so that a part of the energy is absorbed even in the absence of interactions in a layer. While propagating a path of length $l$the energy absorbed by the gas will be equal to $E_\gamma^{in} - E_\gamma^{out} = E_\gamma^{in}\left(1 - exp(-K_\nu^{abs} l)\right)$, where $K_\nu^{abs}$is the monochromatic attenuation coefficient for gas. The presence of this procedure leads to a necessity to discard trajectories with a too small weight. The equation (4) also changes: in the j-th step by temperature the value of k_ν^{abs} is considered to be constant, $k_\nu^{abs} = k_\nu^{abs}(T_{j-1})$.The remaining steps

of the algorithm remain the same thanks to the absence of reradiation by the gas component and local thermodynamic equilibrium.

Besides the changes mentioned above, optimization of the algorithm was realized using an asymptotic solution of the Milne problem [5] for the case of not much transparent media proposed in [6,7] based on the "value-modeling" theory. Modification allowed reducing the computation time by 30%. The data about Venus' atmosphere composition as well as the data about the fractions of gas, pressure, composition, form and size of aerosol particles were taken from [5]. The optical properties of aerosol were calculated according to Mie theory, the refraction coefficient of the mixture of sulfuric acid and water were taken from the database HITRAN [6]. The parameters of gas were calculated with the help of multilinear calculations using the databases HITRAN and HITEMP [7]. The data obtained were compared to results from other codes.

In fig.2 the result of calculation of the vertical profile of temperature in the atmosphere is presented.

The obtained dependence shows higher temperature values with respect to the experiment, this is explained by the absence in the model of a reradiation channel connected to the radiation from gas, which makes a significant contribution due to high temperatures. Regretfully introducing this process into the considered model leads to the necessity of solving a non-stationary problem. Effects connected to convection are also not taken into account.

The work has been done under the financial support of the RFFR (grant 09-01-00035-a), Projects of RAS (2.4) and DMS of RAS (1.3.2).

References

[1] Bjorkman J.E., Wood K. (2001). Radiative equilibrium and temperature correction in Monte Carlo radiation transfer. *Astrophysical Journal*. Vol. **334**,pp. 282-288.

[2] Lucy L.G. (1999).Computing radiative equilibria with Monte Carlo technique. *Astronomy and Astrophysics*. Vol. **554**,pp. 615-639.

[3] Pinte C., Harries T.J., Min M., Watson A.M., Dullemond C.P., Woitke P., Ménard F., Durán-Rojas M.C. (1999).Benchmark problems for continuum radiative transfer. High optical depths, anisotropic scattering, and polarisation. *Astronomy and Astrophysics*. Vol. **498**,pp. 967-980.

[4] Maarten Baes,Dimitris Stamatellos, Jonathan I. Davies, Anthony P. Whitworth, Sabina Sabatini, Sarah Roberts, Suzanne M. Linder, Rhodri Evans. (2005).Radiative equilibrium in Monte Carlo radiative transfer using frequency distribution adjustment. *New Astronomy*. Vol. **10 I. 6**,pp. 523-533.

[5] B. Davison. (1957). *Neutron Transport Theory*. Oxford University Press.

[6] G. A. Mikhailov. (1992). *Optimization of Weighted Monte Carlo Methods*. Springer-Verlag: New-York, Berlin, Heidelberg.

[7] B. A. Kargin, G. A. Mikhailov. (1972).An analysis of the efficiency of using asymptotic solutions in Monte Carlo computations. *Zh. Vychisl. Mat. Mat. Fiz.*. Vol. **12**,No. **1**,pp. 150-158 (in Russian).

[8] Kliore A. J., Moroz V, I., Keating J. M. (1985).Venus International Reference Atmosphere (VIRA). *Advances in space research*. Vol. **15**,No. **11**,pp. 150-158.

[9] (2009).The HITRAN 2008 molecular spectroscopic database. *Journal of Quantitative Spectroscopy & Radiative Transfer*. Vol. **110**,pp. 533-572.

[10] Rothman L. S., Gordon I. E., Barber R. J., Dothe H., Gamache R. R., Goldman A., Perevalov V. I., Tashkun S. A., Tennyson J.. (2011). HITEMP, the High-Temperature Molecular Spectroscopic Database. *JQSRT*. (in press)

Научное издание

ПРИКЛАДНЫЕ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА. ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ И СТАТИСТИЧЕСКИЕ ВЫВОДЫ

ТРУДЫ

Международного семинара

Новосибирск, 20-22 сентября 2011г.

На английском языке

Подписано в печать 20.07.2011. Формат 60 × 84 1/8.
Усл. печ. л. 54,0. Уч.-изд. л. 50,22.
Тираж 75 экз. Заказ № 1168.