APPLIED METHODS OF STATISTICAL ANALYSIS.

NONPARAMETRIC METHODS IN CYBERNETICS AND SYSTEM ANALYSIS

PROCEEDINGS

OF THE INTERNATIONAL WORKSHOP

18-22 September 2017

Krasnoyarsk

2017

UDC 519.22(063)
A 67

E d i t o r s:

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

A 67 **Applied Methods of Statistical Analysis. Nonparametric Methods in Cybernetics and System Analysis** - AMSA'2017, Krasnoyarsk, Russia, 18-22 September, 2017: Proceedings of the International Workshop. - Novosibirsk: NSTU publisher, 2017. - 379 pp.
ISSN 2313-870X

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

Applied Methods of Statistical Analysis. Nonparametric Methods in Cybernetics and System Analysis

Scientific Program Committee:

I. Kovalev,	Siberian State Aerospace University, Russia
A. Medvedev,	Siberian State Aerospace University, Russia
B. Lemeshko,	Novosibirsk State Technical University, Russia
M. Arana-Jimenez,	University of Cadiz, Spain
N. Balakrishnan,	McMaster University, Canada
E. Chimitova,	Novosibirsk State Technical University, Russia
Yu. Dmitriev,	Tomsk State University, Russia
G. Koshkin,	Tomsk State University, Russia
M. Krnjajić,	National University of Ireland, Ireland
H. Liero,	University of Potsdam, Germany
G. Mikhailov,	Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
V. Mkhitarian,	National Research University Higher School of Economics, Russia
M. Nikulin,	University of Bordeaux, France
V. Ogorodnikov,	Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
F. Tarasenko,	Tomsk State University, Russia
V. Timofeev,	Novosibirsk State Technical University, Russia

Local Organizing Committee:

Evgeny Agafonov, Ekaterina Chimitova, Olesya Chubarova, Ekaterina Chzhan, Anastasiya Raskina, Mariia Semenova

Preface

The Fourth International Workshop "Applied Methods of Statistical Analysis. Nonparametric Methods in Cybernetics and System Analysis" – AMSA'2017 is organized by Siberian State University of Science and Technologies called after academician M.F. Reshetnev and Novosibirsk State Technical University. It took place in Krasnoyarsk, which is one of the largest Siberian cities. The Workshop aims to bring together specialists interested in the development of statistical methods of data analysis and correct and efficient practical application of these methods.

Within the framework of AMSA'2017, the XVI International Symposium on Nonparametric Methods in Cybernetics and System Analysis was held. The first such Symposium was held in 1976 in Tomsk, and since then, after each two or three years, it was taken at various places in Siberia, collecting participants from all the Soviet Union, and later – from other countries. The Symposium is devoted to the development of modern mathematical methods for building intellectual computer systems for various purposes operating under incomplete knowledge of the studied process and problems of system analysis. The main topics of the Symposium: – nonparametric and robust statistics; – nonparametric adaptive and trained systems; – identification, modeling, classification and processing images; – simulation and organizational systems control; – applications in design and usage of computer systems of various purposes and automation systems.

The first three Workshops "Applied Methods of Statistical Analysis" were organized by Novosibirsk State Technical University in 2011, 2013 and 2015 years. These meetings had been focused on recent research in the areas of survival analysis, reliability, quality of life, and related topics, from both statistical and probabilistic points of view. The great attention was paid to applications of statistical methods in survival analysis, reliability and quality control.

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

The organization of the Fourth International Workshop "Applied Methods of Statistical Analysis. Nonparametric Methods in Cybernetics and System Analysis" – AMSA'2017 was funded by RFBR according to the research project №17-01-20474 and was supported by the Russian Ministry of Education and Science (project 1.1009.2017/4.6).

Prof. Boris Lemeshko

Contents

On Distribution Functionals Estimation with Auxiliary Information

Yury G. Dmitriev and Gennady M. Koshkin
National Research Tomsk State University, Tomsk, Russia
e-mail: dmit@mail.tsu.ru, kgm@mail.tsu.ru

Abstract

A class of nonparametric estimators of the main functional of distribution constructed with using auxiliary information available in a parametric model is proposed. It is shown that the use of auxiliary information as the knowledge of other distribution functionals in estimation of the main functional can often provide the mean square error smaller than that of estimators constructed without such auxiliary information. For example, the mathematical expectation of a random variable can be taken as the main functional and the value of its variance can be used as auxiliary information. The asymptotic normality of the proposed estimators is proved.

Keywords: Nonparametric estimator, auxiliary information, U-statistics, asymptotic normality.

Introduction

Consider estimating of the functional

$$\theta(F) = \int \cdots \int \varphi(x_1, \ldots, x_s) \prod_{l=1}^{s} dF(x_l) = \int \varphi(\vec{x}_s) dH(\vec{x}_s) = \theta'(H), \qquad (1)$$

by sample $X_1, \ldots, X_N$ from distribution $F(x), x \in R^1, \varphi : R^s \to R^1, \vec{x}_s = (x_1, \ldots, x_s)$, $H(\vec{x}_s) = \prod_{l=1}^{s} F(x_l)$, $dH(\vec{x}_s) = \prod_{l=1}^{s} dF(x_l)$. Let we have auxiliary information in the form of m prior functionals

$$b_j(F) = \int \cdots \int \psi_j(x_1, \ldots, x_s) \prod_{l=1}^{s} dF(x_l) =$$

$$= \int \psi_j(\vec{x}_s) dH(\vec{x}_s) = b'_j(H), \quad j = \overline{1, m}, \qquad (2)$$

and each of these m functionals (2) may take values from the given finite set

$$\{\beta_j\} = \{\beta_{j1}, \ldots, \beta_{jk_j}\}, \qquad (3)$$

where $\psi_j : R^s \to R^1$, $k_j \geq 1$ is a number of possible values of the jth functional.

Denote such prior class of distribution functions by

$$\mathcal{F}^a = \{F : b_j(F) \in \{\beta_{j1}, \ldots, \beta_{jk_j}\}, \quad j = \overline{1, m}\},$$

and write the prior conditions in the form

$$\Delta_j(F) = \Delta'_j(H) = \int \Psi_j(\vec{x}_{sk_j}) dH(\vec{x}_{sk_j}) = 0,\ j = \overline{1,m},\ F \in \mathcal{F}^a, \tag{4}$$

where

$$\Psi_j(\vec{x}_{sk_j}) = \prod_{t=1}^{k_j} (\psi_j(\vec{x}_s^t) - \beta_{jt}),\ \vec{x}_{sk_j} = (\vec{x}_s^1, \ldots, \vec{x}_s^{k_j}).$$

The problem is to estimate the functional (1) taking into account condition (4). Further, without loss of generality, we assume $k_j = k \geq 1,\ j = \overline{1,m}$. In the special case $k_j = k = 1$, the problem was studied by Levit [12] for the U-statistics and in [6] using other statistics. With aim of using auxiliary information (3), several authors [14, 2, 15] have employed the empirical likelihood method by restricting consideration to the distributions satisfying (3). In the article [16], there was considered the quantile estimation problem in the presence of auxiliary information (3) on the base of M-estimator in conjunction with the empirical likelihood method. In papers [3] and [4], modifications of nonparametric kernel estimators of probability density functionals constructed with using auxiliary information expressed by functionals from conditional and unconditional densities are studied. In all these works, it is shown that the estimators which take into account auxiliary information have smaller variances than the estimators without using auxiliary information.

1 Structure of estimators

To construct the estimators, we need initial estimators of the functionals θ and b. The problem of nonparametric estimation of functionals of types (1) and (4) was considered in [6], where a constructive method for synthesizing estimators was proposed, based on substituting instead of the unknown function $H(\vec{x}_s)(H(\vec{x}_{sk}))$ its nonparametric estimator. In the case $s > 1$, several estimators based on the different methods of forming an s-dimensional sample from the set of initial observations can be proposed for the s-dimensional function $H(\vec{x}_s)$. In the general form, the estimator can be expressed by the formula

$$\hat{H}_\tau(x_1, \ldots, x_s) = \frac{1}{|\omega_\tau|} \sum_{\{i_j\} \in \omega_\tau} \prod_{j=1}^{s} c(x_j - X_{i_j}), \tag{5}$$

where $c(t) = \{1 : t > 0;\ 0 : t \leq 0\}$, ω_τ is the set of index compositions $(i_1, \ldots, i_s)$, selected according to some rule with the index τ $(i_j = 1, \ldots, N;\ j = 1, \ldots, s)$, $|\omega_\tau|$ is the number of elements in the set ω_τ. In particular, if we consider all the sets $(i_1, \ldots, i_s)$ that do not contain any coinciding index: $i_j \neq i_l$ for all $j, l = 1, \ldots, N$ (denote this set by ω_1), then $|\omega_1| = C_N^s s!$, and the estimator is defined in the form

$$\hat{H}_1(x_1, \ldots, x_s) = \frac{1}{C_N^s s!} \sum_{\{i_j\} \in \omega_1} \prod_{j=1}^{s} c(x_j - X_{i_j}). \tag{6}$$

If we consider all possible sets of indices, then we obtain the estimator

$$\hat{H}_2(x_1,\dots,x_s)=\frac{1}{N^s}\sum_{i_1=1}^{N}\cdots\sum_{i_s=1}^{N}\prod_{t=1}^{s}c(x_t-X_{i_t})=\prod_{t=1}^{s}F_N(x_t).$$

The simplest in the number of compositions and computational operations is the estimator

$$\hat{H}_3(x_1,\dots,x_s)=\frac{1}{[N/s]}\sum_{i=1}^{[N/s]}\prod_{t=1}^{s}c(x_t-X_{t+(i-1)s}),$$

where $[z]$ is the integer part of z.

The multiplicity of estimators of the functions $\hat{H}_\tau(\vec{x}_s)(\hat{H}(\vec{x}_{sk}))$ leads to a multiplicity of estimators for the functionals $\theta'(H)$ and $\Delta_j(H)$. Thus, using $\hat{H}_1(\cdot)$ in (1) and (4), we obtain the U-statistics (see [8, 9]), and the application of $\hat{H}_2(\cdot)$ leads to the Mises functionals [13, 5]. The choice of one or another estimator of $\hat{H}_\tau$ depends on the practical possibilities of working with the sample and the required accuracy of estimating θ and b.

Now turn to the solution of the problem posed above. Consider the class of estimators

$$\hat{\theta}_\tau(\boldsymbol{\lambda})=\hat{\theta}_\tau-\sum_{j=1}^{m}\lambda_j\Delta'_j(\hat{H}_\tau),\tag{7}$$

where $\boldsymbol{\lambda}=(\lambda_1,\dots,\lambda_m)$ is a coefficient vector. Let us find the vector $\boldsymbol{\lambda}$ from the minimum of mean square error (MSE)

$$S_F\hat{\theta}_\tau(\boldsymbol{\lambda})=M_F[\hat{\theta}_\tau(\boldsymbol{\lambda})-\theta(F)]^2=M_F(\hat{\theta}_\tau-\theta(F))^2-\boldsymbol{\lambda}M_F(\hat{\theta}_\tau-\theta(F))\hat{\Delta}_\tau^T+\boldsymbol{\lambda}M_F\hat{\Delta}_\tau^T\hat{\Delta}_\tau\boldsymbol{\lambda}^T,$$

where $\hat{\Delta}_\tau=(\Delta'_1(\hat{H}_\tau),\dots,\Delta'_m(\hat{H}_\tau))^T$.

Let the matrix

$$V_\tau=M_F\hat{\Delta}_\tau\hat{\Delta}_\tau^T=\|\ M_F\Delta'_j(\hat{H}_\tau)\Delta'_l(\hat{H}_\tau)\ \|_{j,l=\overline{1,m}},\tag{8}$$

is a non-degenerate one, V_τ^{-1} is the inverse matrix with respect to V_τ,

$$C_\tau=\|\ M_F(\hat{\theta}_\tau-\theta(F))\Delta'_j(\hat{H}_\tau)\ \|_{j=\overline{1,m}}\tag{9}$$

is the matrix-column. The minimum of the MSE is achieved by

$$\boldsymbol{\lambda}=\boldsymbol{\lambda}_\tau=V_\tau^{-1}C_\tau,\tag{10}$$

for which

$$S_F\hat{\theta}_\tau(\boldsymbol{\lambda}_\tau)=M_F[\hat{\theta}_\tau-\theta(F)]^2-C_\tau^TV_\tau^{-1}C_\tau.\tag{11}$$

The non-negative quantity $C_\tau^TV_\tau^{-1}C_\tau$ determines the decrease of the MSE by attracting auxiliary information (4), and its value depends on the type of the used estimator $\hat{H}_\tau$.

2 Estimators based on U-statistics

Consider the case when in the functionals estimating we use $\hat{H}_1$. Then

$$\hat{\theta}_1 = \theta'(\hat{H}_1) = \frac{1}{s!C_N^s}\sum_{i_1\neq}^{N}\cdots\sum_{\neq i_s}^{N}\varphi(X_{i_1},\ldots,X_{i_s}) =$$

$$= (C_N^s)^{-1}\sum_{1\le i_1<\ldots<i_s\le N}\bar{\varphi}(X_{i_1},\ldots,X_{i_s}), \tag{12}$$

$$\Delta_j'(\hat{H}_1) = \frac{1}{(sk)!C_N^{sk}}\sum_{i_1\neq}^{N}\cdots\sum_{\neq i_{sk}}^{N}\Psi_j(X_{i_1},\ldots,X_{i_{sk}}) =$$

$$= \left(C_N^{sk}\right)^{-1}\sum_{1\le i_1<\ldots<i_{sk}\le N}\bar{\Psi}_j(X_{i_1},\ldots,X_{i_{sk}}), \tag{13}$$

where the bar above denotes the symmetrization of the function with respect to the permutations of its arguments. Formulas (12), (13) are the U-statistics [8], [9]. For each $F\in\mathcal{F}^a$

$$M_F\hat{\theta}_1 = \theta(F),\ M_F\Delta_j'(\hat{H}_1) = \Delta_j(F) = 0.$$

So, it follows that the estimators of the class (7) for $\tau = 1$ are unbiased. Now, find the expression for $\boldsymbol{\lambda}_1$ that minimizes the variance of the estimator. Let

$$M_F\varphi^2(X_1,\ldots,X_s) < \infty, \tag{14}$$

$$M_F\psi_j^2(X_1,\ldots,X_s) < \infty, j = \overline{1,m}, \tag{15}$$

Define

$$\bar{\Psi}_{jr}(\vec{x}_r) = M_F(\bar{\Psi}_j(\vec{X}_{sk})|\vec{X}_r = \vec{x}_r), r = \overline{1,sk}, \tag{16}$$

$$\bar{\varphi}_r(\vec{x}_r) = M_F(\bar{\varphi}(\vec{X}_s)|\vec{X}_r = \vec{x}_r), \bar{\psi}_r(\vec{x}_r) = M_F(\bar{\psi}(\vec{X}_s)|\vec{X}_r = \vec{x}_r), r = \overline{1,s},$$

$$v_{jl}^{(r)} = M_F\bar{\Psi}_{jr}(\vec{X}_r)\bar{\Psi}_{lr}(\vec{X}_r), r = \overline{1,sk},\ \ j,l = \overline{1,m}, \tag{17}$$

$$c_j^{(r)} = M_F\bar{\varphi}_r(\vec{X}_r)\bar{\Psi}_{jr}(\vec{X}_r), r = \overline{1,s},\ j = \overline{1,m}, \tag{18}$$

$$d^{(r)} = M_F(\bar{\varphi}_r(\vec{X}_r) - \theta(F))^2,\ r = \overline{1,s}. \tag{19}$$

According to [9]

$$M_F\Delta_j'(\hat{H}_1)\Delta_l'(\hat{H}_1) = v_{1,jl} = \left(C_N^{sk}\right)^{-1}\sum_{r=1}^{sk}C_{sk}^rC_{N-sk}^{sk-r}v_{jl}^{(r)}, \tag{20}$$

$$M_F(\hat{\theta}_1-\theta)\Delta_j'(\hat{H}_1) = c_{1,j} = (C_N^s)^{-1}\sum_{r=1}^{s}C_s^rC_{N-s}^sc_j^{(r)}, \tag{21}$$

$$D_F\hat{\theta}_1 = (C_N^s)^{-1}\sum_{r=1}^{s}C_s^rC_{N-s}^sd^{(r)}. \tag{22}$$

Then, from (20) and (21)

$$V_1 = \| \, v_{1,jl} \, \|_{j,l=\overline{1,m}}, \quad C_1 = \| \, c_{1,j} \, \|_{j=\overline{1,m}} \, .$$

Assuming the non-degeneracy of the matrix V_1, we have for all $F \in \mathcal{F}^a$

$$\boldsymbol{\lambda}_1 = V_1^{-1} C_1, M_F \hat{\theta}_1(\boldsymbol{\lambda}_1) = \theta(F), D_F \hat{\theta}_1(\boldsymbol{\lambda}_1) = D_F \hat{\theta}_1 - C_1^T V_1^{-1} C_1.$$

Now find the main part of $\boldsymbol{\lambda}_1$. For $N \to \infty$, we obtain

$$v_{1,jl} = \frac{(sk)^2 v_{jl}^{(1)}}{N} \left(1 + O(N^{-1})\right), \tag{23}$$

$$c_{1,j} = \frac{s^2 k c_j^{(1)}}{N} \left(1 + O(N^{-1})\right), \tag{24}$$

$$D_F \hat{\theta}_1 = N^{-1} s^2 d^{(1)} + O(N^{-2}). \tag{25}$$

In our problem

$$\bar{\Psi}_{j1}(x_1) = (\bar{\psi}_{j1}(x_1) - b_j(F)) a_{j,k}(F),$$

$a_{j,1}(F) = 1$, and for $k > 1$

$$a_{j,k}(F) = \frac{1}{k} \sum_{l=1}^{k} \Delta_j^{(l)}, \quad \Delta_j^{(l)} = \prod_{t=1, t \neq l}^{k} \Delta_{jt}(F) \, .$$

Consequently,

$$v_{jl}^{(1)} = \mathrm{cov}_F(\bar{\psi}_{j1}(X_1), \bar{\psi}_{l1}(X_1)) a_{j,k}(F) a_{l,k}(F), \tag{26}$$

$$c_j^{(1)} = \mathrm{cov}_F(\bar{\varphi}_1(X_1), \bar{\psi}_{j1}(X_1)) a_{j,k}(F), \tag{27}$$

$$d^{(1)} = D_F \bar{\varphi}_1(X_1). \tag{28}$$

Further

$$\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}^{(1)}(1 + O(N^{-1})), \tag{29}$$

$$\boldsymbol{\lambda}^{(1)} = \frac{1}{k} a^{-1} V^{-1} C, \tag{30}$$

where the matrix-column

$$C = \| \, c_j \, \|_{j=\overline{1,m}}, \; c_j = \mathrm{cov}_F(\bar{\varphi}_1(X_1), \bar{\psi}_{j1}(X_1)),$$

V^{-1} is the inverse matrix of the matrix

$$V = \| \, v_{jl} \, \|_{j,l=\overline{1,m}}, \; v_{jl} = \mathrm{cov}_F(\bar{\psi}_{j1}(X_1), \bar{\psi}_{l1}(X_1)),$$

a is the diagonal matrix with elements $a_{j,k}(F), j = \overline{1,m}$, on the main diagonal, a^{-1} is the inverse matrix of the matrix a.

Consider a random variable

$$\hat{\theta}_1(\boldsymbol{\lambda}^{(1)}) = \hat{\theta}_1 - \sum_{j=1}^{m} \lambda_j^{(1)} \Delta_j'(\hat{H}_1). \tag{31}$$

Lemma. *Suppose that the conditions (14) and (15) are satisfied,* $\det V \neq 0$,

$$\sigma_0^2 = s^2(d^{(1)} - C^T V^{-1} C) > 0. \tag{32}$$

Then for each $F \in \mathcal{F}^a$

$$M_F\hat{\theta}_1(\boldsymbol{\lambda}^{(1)}) = \theta_1,\ D_F\hat{\theta}_1(\boldsymbol{\lambda}^{(1)}) = N^{-1}\sigma_0^2 + O(N^{-2}),$$

$$\sqrt{N}(\hat{\theta}_1(\boldsymbol{\lambda}_1) - \hat{\theta}_1(\boldsymbol{\lambda}^{(1)})) \xrightarrow{p} 0, \quad N \to \infty, \tag{33}$$

and, consequently, random variables $\sqrt{N}(\hat{\theta}_1(\boldsymbol{\lambda}_1) - \theta)$ *and* $\sqrt{N}(\hat{\theta}_1(\boldsymbol{\lambda}^{(1)}) - \theta)$ *have the same asymptotic normal distribution with the mean* 0 *and variance* σ_0^2

$$\mathcal{L}(\sqrt{N}(\hat{\theta}_1(\boldsymbol{\lambda}^{(1)}) - \theta(F))) \to \mathcal{N}(0, \sigma_0^2), \tag{34}$$

where σ_0^2 *is defined by the formula (32).*

Proof. Unbiasedness of $\hat{\theta}_1(\boldsymbol{\lambda}^{(1)})$ follows from the U-statistics unbiasedness. The expression for the variance $D_F\hat{\theta}_1(\boldsymbol{\lambda}^{(1)})$ follows from (23)–(29). On the basis of (15),(16), and (29)

$$M_F(\hat{\theta}_1(\boldsymbol{\lambda}_1) - \hat{\theta}_1(\boldsymbol{\lambda}^{(1)}))^2 = O(N^{-2}).$$

This, together with the Chebyshev inequality, gives (33).

Further, $\hat{\theta}_1(\boldsymbol{\lambda}^{(1)})$ write as the following U-statistic:

$$\hat{\theta}_1(\boldsymbol{\lambda}^{(1)}) = (C_N^s)^{-1} \sum_{1 \le i_1 < \ldots < i_{sk} \le N} \bar{Q}(X_{i_1}, \ldots, X_{i_{sk}}) \tag{35}$$

with the kernel

$$\bar{Q}(x_1, \ldots, x_{sk}) = \frac{1}{(sk)!} \sum_{(p)} Q(x_{i_1}, \ldots, x_{i_{sk}}),$$

where

$$Q(\vec{X}_{sk}) = \varphi(\vec{X}_s) - \sum_{j=1}^{m} \lambda_j^{(1)} \Psi_j(\vec{X}_{sk}),$$

the sign (p) denotes summation over all $(sk)!$ permutations $(i_1, \ldots, i_{sk})$ of numbers $(1, \ldots, sk)$.

Let

$$\bar{Q}_1(x_1) = M_F(\bar{Q}(\vec{X}_{sk}) | X_1 = x_1) = \frac{\bar{\varphi}_1(x_1) + (k-1)\theta}{k} - \sum_{j=1}^{m} \lambda_j^{(1)} \bar{\Psi}_{j1}(x_1).$$

As $D_F\bar{Q}_1(X_1) = k^{-2}\sigma^2 > 0$, then $\bar{Q}(x_1, \ldots, x_{sk})$ is a non-degenerate kernel. Hence, on the basis of the theorem on asymptotic normality U-statistics (see [8, 9]), we obtain (34). Lemma is proved.

Note an important conclusion that follows from the foregoing. The main term of the variance $D_F\hat{\theta}_1(\boldsymbol{\lambda}^{(1)})$ (23) is independent of $\Delta_{jl},\ l = \overline{1, k_j},\ j = \overline{1, m}$, and also of the $k_j = k \geq 1$ values of the integrals $b_j(F),\ j = \overline{1, m}$ (see (3)), and coincides with the case when $(k_j = 1)$. This means that an increase of the uncertainty in the prior conditions (3) due to the multiplicity of the finite number of values of the functionals $b_j(F),\ j = \overline{1, m}$, does not make worse the asymptotic properties of $\hat{\theta}_1(\lambda_1)$ as compared with the case when it is *a priori* known that each functional $b_j(F),\ j = \overline{1, m}$ takes only one value. For a finite sample size and $k_j > 1$, the multiplicity of the values of $b_j(F)$ reveals itself only in terms of a higher order of smallness by the presence of a positive term of order $O(N^{-2})$. Let us illustrate this with the following example.

Example 1. Let $m = 1,\ k_1 = 2$. Omitting the extra indices, we obtain

$$\lambda_1 = \frac{\mathrm{cov}_F(\varphi, \psi)}{(\Delta_1 + \Delta_2)D_F\psi + 2D_F^2\psi/(\Delta_1 + \Delta_2)(N-1)},$$

$$D_F\theta_N(\lambda_1) = \frac{1}{N}\left[D_F\varphi - \frac{\mathrm{cov}_F^2(\varphi, \psi)}{D_F\psi + 2D_F^2\psi/(\Delta_1 + \Delta_2)^2(N-1)}\right] =$$

$$= \frac{1}{N}\left[D_F\varphi - \frac{\mathrm{cov}_F^2(\varphi, \psi)}{D_F\psi}\right] + \frac{2\mathrm{cov}_F^2(\varphi, \psi)}{N[2D_F\psi + (\Delta_1 + \Delta_2)^2(N-1)]}.$$

The first term in the last expression shows that the influence of polysemy $(k_1 = 2)$ in the prior condition with respect to the principal variance term is absent. The presence of polysemy manifests itself only in the second term.

Highlight the case when $s = 1,\ k = 1$, and it is known that among the components $\Delta_j,\ \overline{1, m}$, only one (which is unknown) is equal to zero, while the others are different from zero. This prior information is expressed by one condition

$$\Delta(F) = \prod_{j=1}^{m}\Delta_j(F) = \prod_{j=1}^{m}\int(\psi_j(x) - \beta_j)dF(x) = \Delta'(H) = \int\Psi(\vec{x}_m)dH(\vec{x}_m) = 0 \tag{36}$$

where $\Psi(\vec{x}_m)) = \prod_{j=1}^{m}(\psi_j(x_j) - \beta_j)$.

Using $\hat{H}_1$ in the estimation of functionals leads to

$$\hat{\theta}_1(\lambda_1{}^{(1)}) = \hat{\theta}_1 - \lambda_1^{(1)}\Delta'(\hat{H}_1), \tag{37}$$

where

$$\lambda_1^{(1)} = \frac{\sum_{j=1}^{m}\mathrm{cov}_F(\varphi, \psi_j)\Delta^{(j)}}{\sum_{j=1}^{m}(\Delta^{(j)})^2 D_F\psi_j)}, \tag{38}$$

$\Delta^{(j)} = \prod_{l=1,l\neq j}^{m} \Delta_l$. By Lemma, the quantity

$$\mathcal{L}(\sqrt{N}(\hat{\theta}_1(\lambda_1{}^{(1)})) - \theta(F)) \to \mathcal{N}(0, \sigma_0^2),\ N \to \infty,$$

where the variance

$$\sigma_0^2 = D_F\varphi(X_1) - \frac{\left[\sum_{j=1}^{m} \mathrm{cov}_F(\varphi, \psi_j)\Delta^{(j)}\right]^2}{\sum_{j=1}^{m}(\Delta^{(j)})^2 D_F\psi_j)}. \tag{39}$$

Example 2. Let the continuous distribution function $F(x)$ be symmetric, i.e. $F(x) = 1-F(2\alpha-x),\ x \in R^1$, where the center of symmetry α takes one of the values $\alpha_1, \ldots, \alpha_m$. Use this information in estimating $\theta(F) = \int \varphi(x)dF(x)$. Consider the functions

$$\psi_j(x) = [\varphi(x) - \varphi(2\alpha_j - x)]/2,\ j = \overline{1,m},$$

and the functionals $\Delta_j(F) = \int \psi_j(x)dF(x),\ j = \overline{1,m}$, where $\Delta_j(F) = 0$, if the center of symmetry is α_j. Hence, the prior condition has the form

$$\Delta(F) = \prod_{j=1}^{m} \Delta_j(F) = 0.$$

Estimator for $\theta(F)$ is constructed by the formula (37) with $\lambda^{(1)}$ (38).

3 Adaptive Estimators

Statistics $\hat{\theta}_1(\boldsymbol{\lambda}_1)$ and $\hat{\theta}_1(\boldsymbol{\lambda}^{(1)})$ can be used as estimators for $\theta(F)$ if you know $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}^{(1)}$; otherwise you need to build adaptive estimators. Consider the following adaptive estimator

$$\hat{\theta}_1(\hat{\boldsymbol{\lambda}}^{(1)}) = \hat{\theta}_1 - \hat{\boldsymbol{\lambda}}^{(1)}\hat{\Delta}_1^T$$

with

$$\hat{\boldsymbol{\lambda}}^{(1)} = \frac{1}{k}\hat{a}^{-1}\hat{V}^{-1}\hat{C}. \tag{40}$$

Here $\hat{C} = \| \hat{c}_j^{(1)} \|_{j=\overline{1,m}}$ is the matrix-column with elements

$$\hat{c}_j = \int [\bar{\varphi}(x_1, \ldots, x_s)\bar{\psi}_j(x_1, x_1' \ldots, x_s') - \bar{\varphi}(x_1, \ldots, x_s)\bar{\psi}_j(x_1', \ldots, x_s')]d\hat{H}_1(\vec{x}_{2s}),$$

$\hat{V}^{-1}$ is the inverse matrix of the matrix $\hat{V} = \| \hat{v}_{jl} \|_{j,l=\overline{1,m}}$ with elements

$$\hat{v}_{jl} = \int [\bar{\psi}_j(x_1, \ldots, x_s)\bar{\psi}_l(x_1, x_1' \ldots, x_s') - \bar{\psi}_j(x_1, \ldots, x_s)\bar{\psi}_l(x_1', \ldots, x_s')]d\hat{H}_1(\vec{x}_{2s}),$$

$\hat{a}^{-1}$ is the inverse matrix of the diagonal matrix $\hat{a}$ with elements

$$\hat{a}_{j,k} = \frac{1}{k}\sum_{l=1}^{k}\prod_{t=1,t\neq l}^{k}\hat{\Delta}_{jt},$$

where

$$\hat{\Delta}_{jt} = \int \psi_j(\vec{x}_s)d\hat{H}_1(\vec{X}_s) - \beta_{jt}.$$

Theorem. *Let* $F \in \mathcal{F}^a$, *the conditions (14) and (15) hold,* $\det V \neq 0$,

$$M_F\{M_F(\varphi^2(\vec{X}_s)|X_1)M_F(\psi_j^2(\vec{X}_s)|X_1)\} < \infty,\ j = \overline{1,m}. \tag{41}$$

$$M_F\{M_F(\psi_j^2(\vec{X}_s)|X_1)M_F(\psi_l^2(\vec{X}_s)|X_1)\} < \infty,\ j,l = \overline{1,m}. \tag{42}$$

Then as $N \to \infty$

$$\hat{\boldsymbol{\lambda}}^{(1)} \xrightarrow{p} \boldsymbol{\lambda}^{(1)}, \tag{43}$$

and if (32) is satisfied, then

$$\mathcal{L}(\sqrt{N}(\hat{\theta}_1(\hat{\boldsymbol{\lambda}}^{(1)}) - \theta(F))) \to \mathcal{N}(0,\sigma_0^2), \tag{44}$$

where σ_0^2 *is determined by the formula (32).*

Proof. Since $\det V \neq 0$, the components $\hat{\lambda}_j^{(1)}$ of the vector $\hat{\boldsymbol{\lambda}}^{(1)}$ are continuous functions of estimators (as variables) $\hat{v}_{jl}$, $\hat{c}_j$, $\hat{\Delta}_j$, in the true points v_{jl}, c_j, Δ_j, $j,l = \overline{1,m}$. All estimators are unbiased and according to (14), (15), (41), and (42) are consistent. Therefore, on the basis of the second continuity theorem [1] it follows (43).

Now prove (44). We have

$$\sqrt{N}(\hat{\theta}_1(\hat{\boldsymbol{\lambda}}^{(1)}) - \theta(F)) = \sqrt{N}(\hat{\theta}_1(\boldsymbol{\lambda}^{(1)}) - \theta(F)) + \hat{R}_N,$$

where $\hat{R}_N = (\boldsymbol{\lambda}^{(1)} - \hat{\boldsymbol{\lambda}}^{(1)})\sqrt{N}\hat{\Delta}_1^T$. The components of vector $\hat{\Delta}_1$ are the U-statistics (13) with non-degenerate kernels. On the basis of (14), (15), and Theorems 4.2.3. [9] $\sqrt{N}\hat{\Delta}_1 \Rightarrow \eta, N \to \infty$ (the sign $\Rightarrow$ denotes the convergence in distribution), where η is a normal m-dimensional vector with $M\eta = \mathbf{0}$ and the covariance matrix V. This and (43) imply $\hat{R}_N \Rightarrow 0$. Taking into account (34), we obtain (44). Theorem is proved.

References

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

[2] Chen J., Qin J. (1993). Empirical Likelihood Estimation for Finite Populations and the Effective Usage of Auxiliary Information. *Biometrika*. Vol. **80**, pp. 107–116.

[3] Dmitriev Yu.G., Koshkin G.M. (1987). On the Use of a Priori Information in Nonparametric Regression Estimation. *IFAC Proceedings Series*. Vol. **2**, pp. 223–228.

[4] Dmitriev Yu.G., Koshkin G.M. (1987). Using Additional Information in Nonparametric Estimation of Density Functionals. *Automat. and Remote Control*. Vol. **48**, No. **10**, pp. 1307–1316.

[5] Dmitriev Yu.G., Koshkin G.M., Simahin V.A., Tarasenko F.P., Shulenin V.P. (1974). *Nonparametric Estimation of Functionals by Stationary Samples*. Tomsk Univ. Press, Tomsk (in Russian).

[6] Dmitriev Yu.G., Tarasenko F.P. (1978). On the Use of a Priori Information in Estimated Linear Functionals of Distribution. *Problems Control and Inform. Theory*. Vol. **7**, No. **6**, pp. 459–469.

[7] Dobrovidov A.V., Koshkin G.M., Vasiliev V.A. (2012). *Non-parametric State Space Models*. Kendrick Press, Heber City, UT.

[8] Hoeffding W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *Ann. Math. Statist.*. Vol. **19**, No. **3**, pp. 293–325.

[9] Koroljuk V.S., Borovskich Yu.V. (1994). *Theory of U-Statistics*. Springer Science+Business Media, Dordrecht.

[10] Koshkin G.M. (1993). Stable Estimation of Ratios of Random Functions from Experimental Data. *Russian Physics Journal*. Vol. **36**, No. **10**, pp. 1008–1015.

[11] Koshkin G.M. (1999). Deviation Moments of the Substitution Estimator and of its Piecewise-Smooth Approximations. *Siberian Mathematical Journal*. Vol. **40**, No. **3**, pp. 515–527.

[12] Levit B.Ya. (1975). Conditional Estimation of Linear Functionals. *Problems Inform. Trans.* Vol. **11**, No. **4**, pp. 291–302.

[13] von Mises R. (1947). On the Asymptotic Distribution of Differentiable Statistical Function. *Ann. Math. Statist.* Vol. **18**, No. **3**, pp. 309–348.

[14] Owen A.B. (1991). Empirical Likelihood for Linear Models. *Ann. Statist.* Vol. **19**, pp. 1725–1747.

[15] Qin J., Lawless J. (1994). Empirical Likelihood and General Estimating Equations. *Ann. Statist.* Vol. **22**, pp. 300–325.

[16] Zhang B. (1995). M-estimation and Quantile Estimation in the Presence of Auxiliary Information. *Journal of Statistical Planning and Inference.* Vol. **44**, pp. 77–94.

Combined Identification Algorithms

Yury G. Dmitriev, Gennady M. Koshkin and Vadim Yu. Lukov
National Research Tomsk State University, Tomsk, Russia
e-mail: dmit@mail.tsu.ru, kgm@mail.tsu.ru, v-lukov@rambler.ru

Abstract

In many applied problems it is required to construct a mathematical model of the dependence of output variables on input variables of the stochastic object. To solve this problem, both parametric and nonparametric approaches are used. Each of these approaches has advantages and disadvantages. In the paper, we consider combined algorithms for the identification of stochastic objects using jointly nonparametric and parametric estimates of regression.

Keywords: Nadaraya–Watson statistic, parametric estimate, regression, combined algorithm, identification, bootstrap.

Introduction

Suppose that a stochastic object is described by a regression function

$$r(\vec{x}) = M(Y|\vec{X} = \vec{x}) = \int yp(y|\vec{x})dy = \frac{\int yp(\vec{x}, y)dy}{p(\vec{x})}, \tag{1}$$

where $\left(\vec{X}, Y\right) = \left(X^{(1)}, ..., X^{(p)}, Y\right)$ is a $(p+1)$-dimensional vector of p object's inputs and output, $p(\vec{x}, y)$ is their joint distribution density, $p(\vec{x})$ is a distribution density of inputs, and $p(y|\vec{x})$ is the conditional distribution density.

Let there be independent observations $\left(\vec{X}_i, Y_i\right) = \left(X_i^{(1)}, ..., X_i^{(p)}, Y_i\right)$, $i = 1, ..., n$, of the random vector $\left(\vec{X}, Y\right)$. Let us consider the nonparametric Nadaraya–Watson [2, 11] estimate of the regression function (1)

$$\hat{r}(\vec{x}) = \hat{r}(\vec{x}; \vec{X}_1, ..., \vec{X}_n) = \frac{\sum\limits_{i=1}^{n} Y_i K\left(\dfrac{\vec{x} - \vec{X}_i}{\vec{h}_n}\right)}{\sum\limits_{t=1}^{n} K\left(\dfrac{\vec{x} - \vec{X}_i}{\vec{h}_n}\right)}, \tag{2}$$

where $K\left(\dfrac{\vec{x} - \vec{X}_i}{\vec{h}_n}\right) = K\left(\dfrac{x^{(1)} - X_i^{(1)}}{h_n^{(1)}}\right) \cdots K\left(\dfrac{x^{(p)} - X_i^{(p)}}{h_n^{(p)}}\right)$ is a p-dimensional kernel (the product of p one-dimensional kernels), $\vec{h}_n = (h_n^{(1)}, \ldots, h_n^{(p)})$ is a p-dimensional vector of bandwidth parameters.

Usually the researcher has some information about the nature of the dependence of the output of the object from the inputs. Suppose that he can express this knowledge

in the form of a given function $\varphi(\vec{x},\vec{\theta})$, where $\vec{\theta}=(\theta^{(1)},\ldots,\theta^{(s)})$ is the vector of the known parameters. This type of information we call as a prior guess.

Consider the task of sharing the nonparametric estimation of regression and a prior guess. The approach using combinations of different estimates was studied, for example, in [1]–[3],[9].

1 Combined estimators

1.1 Static model

As a combined regression estimate, we take [2, 9]

$$\hat{R}_\lambda(\vec{x})=(1-\lambda)\hat{r}(\vec{x})+\lambda\varphi(\vec{x},\vec{\theta}), \tag{3}$$

where λ is the weight coefficient determined from minimum of the criterion

$$M\{\hat{R}_\lambda(\vec{x})-r(\vec{x})\}^2. \tag{4}$$

So, from (4) we obtain the optimal λ:

$$\lambda(\vec{x})=\frac{M\{(\hat{r}(\vec{x})-r(\vec{x}))(\hat{r}(\vec{x})-\varphi(\vec{x},\vec{\theta}))\}}{M\{\hat{r}(\vec{x})-\varphi(\vec{x},\vec{\theta})\}^2}. \tag{5}$$

Substituting (5) into (4) and making the transformations, we get:

$$M\{\hat{R}_\lambda(\vec{x})-r(\vec{x})\}^2=M\{\hat{r}(\vec{x})-r(\vec{x})\}^2-\frac{[M\{(\hat{r}(\vec{x})-r(\vec{x}))(\hat{r}(\vec{x})-\varphi(\vec{x},\vec{\theta}))\}]^2}{M\{\hat{r}(\vec{x})-\varphi(\vec{x},\vec{\theta})\}^2}. \tag{6}$$

The second term in (6) shows how much the MSE of the combined estimate $\hat{R}_\lambda(\vec{x})$, taking into account the prior guess $\varphi(\vec{x},\vec{\theta})$, decreases compared to $\hat{r}(\vec{x})$ for each $\vec{x}\in R^p$. Since the optimal $\lambda(\vec{x})$ (5) is usually unknown, it becomes necessary to construct an estimate $\hat{\lambda}(\vec{x})$ of this coefficient, which leads to an adaptive combined estimate

$$\hat{R}_{\hat{\lambda}}(\vec{x})=(1-\hat{\lambda}(\vec{x}))\hat{r}(\vec{x})+\hat{\lambda}(\vec{x})\varphi(\vec{x},\vec{\theta}). \tag{7}$$

Let us consider an estimate of a weight coefficient by a bootstrap method. We write (5) in the form:

$$\lambda(\vec{x})=\frac{M\psi_1(\vec{x})}{M\psi_2(\vec{x})}, \tag{8}$$

where

$$M\psi_1(\vec{x})=M[(\hat{r}(\vec{x})-r(\vec{x}))(\hat{r}(\vec{x})-\varphi(\vec{x},\vec{\theta}))],\quad M\psi_2(\vec{x})=M\{\hat{r}(\vec{x})-\varphi(\vec{x},\vec{\theta})\}^2.$$

Generate a bootstrap sample $\left(\vec{X}_j^*,Y_j^*\right)$, $\vec{X}_j^*=(\vec{X}_{1,j}^*,...,\vec{X}_{n,j}^*)$, $j=1,\ldots,B$, for the numerator and denominator in (8). Then we have:

$$M\psi_1(\vec{x})\simeq\frac{1}{B}\sum_{j=1}^{B}[(\hat{r}(\vec{x};\vec{X}_j^*)-\hat{r}(\vec{x}))(\hat{r}(\vec{x};\vec{X}_j^*)-\varphi(\vec{x},\vec{\theta}))],$$

$$M\psi_2(\vec{x}) \simeq \frac{1}{B}\sum_{j=1}^{B}[\hat{r}(\vec{x};\vec{X}_j^*) - \varphi(\vec{x},\vec{\theta})]^2.$$

As a result, we obtain the following estimate of the weight coefficient (5):

$$\hat{\lambda}_B(\vec{x}) = \frac{\sum_{j=1}^{B}(\hat{r}(\vec{x};\vec{X}_j^*) - \hat{r}(\vec{x}))(\hat{r}(\vec{x};\vec{X}_j^*) - \varphi(\vec{x},\vec{\theta}))}{\sum_{j=1}^{B}[\hat{r}(\vec{x};\vec{X}_j^*) - \varphi(\vec{x},\vec{\theta})]^2}. \tag{9}$$

The usage of (9) in (7) leads to an adaptive combined estimate

$$\hat{R}_{\hat{\lambda}_B}(\vec{x}) = (1 - \hat{\lambda}_B(\vec{x}))\hat{r}(\vec{x}) + \hat{\lambda}_B(\vec{x})\varphi(\vec{x},\vec{\theta}). \tag{10}$$

If θ is evaluated by a sample, then the estimate (10) we will denote as $\tilde{R}_{\hat{\lambda}_B}(\vec{x})$. The properties of these estimates are illustrated below in section 3 by simulation.

1.2 Dynamic model

Consider the dynamic model (cf. [4]–[7],[10])

$$Y_t = f(\vec{X}_t) + \xi_t, \tag{11}$$

where Y_t is the output of the object at the time moment t, $\vec{X}_t = \left(X_t^{(1)}, \ldots, X_t^{(p)}\right)$ is the p-dimensional vector of the inputs at the time moment t, f is an unknown function, ξ_t is the sequence of the i.i.d. random variables with a nonnegative distribution density, $M\xi_t = 0$, $M\xi_t^2 < \infty, M\xi_t^3 = 0$, and $M\xi_t^4 < \infty$.

Assume that f is bounded and its form does not change in the time interval under study. As an prior guess about the form of f, take the function $\varphi(\vec{x},\vec{\theta})$ and consider the following combined adaptive estimate:

$$\hat{R}_{\hat{\lambda}_B}(\vec{x}_t) = (1 - \hat{\lambda}_B(\vec{x}_t))\hat{r}(\vec{x}_t) + \hat{\lambda}_B(\vec{x}_t)\varphi(\vec{x}_t,\vec{\theta}), \tag{12}$$

where

$$\hat{\lambda}_B(\vec{x}_t) = \frac{\sum_{j=1}^{B}(\hat{r}(\vec{x}_t;\vec{X}_j^*) - \hat{r}(\vec{x}_t))(\hat{r}(\vec{x}_t;\vec{X}_j^*) - \varphi(\vec{x}_t,\vec{\theta}))}{\sum_{j=1}^{B}[\hat{r}(\vec{x}_t;\vec{X}_j^*) - \varphi(\vec{x}_t,\vec{\theta})]^2}.$$

The estimate (12) is applied in section 4 for the analysis of stock prices on real data.

2 Modeling

Consider an illustrative example. Let

$$Y(x) = 10 + 1.8x^2 + \xi, \quad \varphi(\vec{x}; \vec{\theta}) = \theta^{(1)} + \theta^{(2)}x, \quad \theta^{(1)} = 8, \quad \theta^{(2)} = 10.8,$$

where ξ is normally distributed random variable with $M\xi = 0$ and $D\xi = \sigma^2$.

For different noise variances and samples sizes n, we investigate the behavior of the combined regression estimate for this model. The qualities of models identification and forecasting will be characterized using average relative errors:

$$\delta(\hat{r}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{r}(X_i)|}{|Y_i|}100\%.$$

For $n = 10$ and $\sigma = 1$, the plots of realizations $Y(x)$, $\varphi(x; \theta^{(1)}, \theta^{(2)})$, $r(x)$ and combined estimate for different $x \in [0, 1]$ are shown in Fig. 1. The behavior of the estimate of the weight coefficient (9) is shown in Fig. 2.

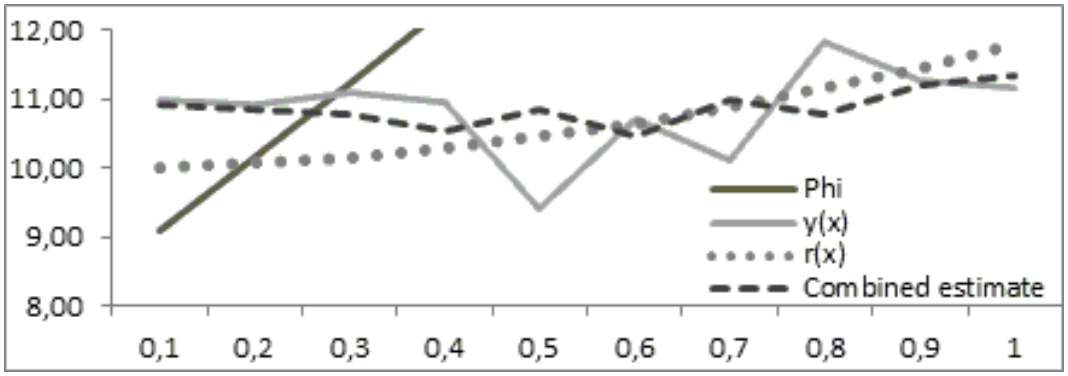


Figure 1: Plots of realizations $Y(x)$, $\varphi(x; \theta^{(1)}, \theta^{(2)})$, $r(x)$ and combined estimate for $n = 10$ and $\sigma = 1$

Let $\varphi(x; \hat{\theta}^{(1)}, \hat{\theta}^{(2)}) = \hat{\theta}^{(1)} + \hat{\theta}^{(2)}x$, where $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}$ are the least mean square (LMS) estimates. For this case, Table 1 gives average relative identification errors for various variances and samples sizes.

Table 1: Average relative identification errors $\delta(\hat{r})$, $\delta(\hat{R})$, and $\delta(\tilde{R})$

σ^2	1			3			5		
n	10	50	100	10	50	100	10	50	100
$\delta(\hat{r})$	4.99	3.87	2.33	23.88	15.43	8.96	56.14	30.20	17.44
$\delta(\hat{R})$	4.53	2.99	2.12	22.86	14.99	8.78	49.90	29.60	16.98
$\delta(\tilde{R})$	4.18	2.69	2.15	19.72	14.07	8.68	48.72	30.11	17.22

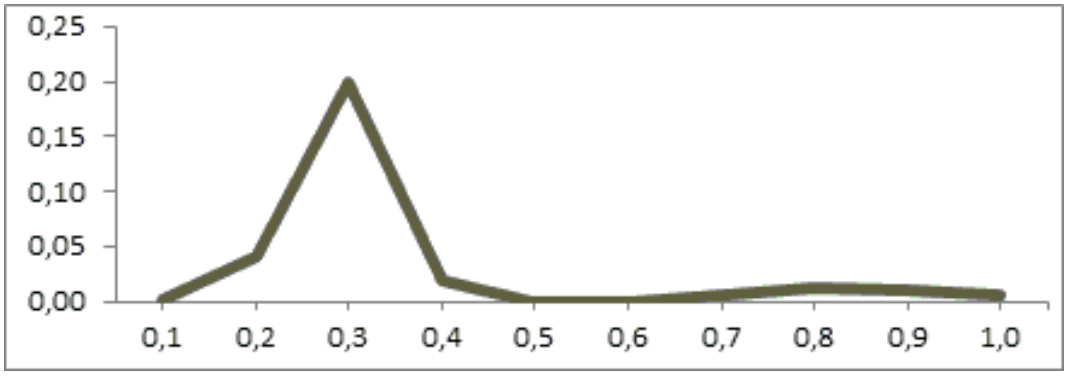


Figure 2: Plot of the dependence of the estimate of the weight coefficient (9) on x

In practice, from Table 1 it follows preferable applying a combined estimate in comparison with a nonparametric estimate in the case of small sample sizes and/or large noise variances.

3 Analysis of real data

The analysis of the prices of Gazprom's stocks for 2016 is carried out on the basis of the first-order autoregression. In this case, it is natural to take as the model the following modification of (11):

$$Y_t = f(Y_{t-1}) + \xi_t, \tag{13}$$

where $t = 2, ..., n$, Y_t is the stock price at the time moment t. We take the parametric function in the form $\varphi(Y_{t-1}; \theta^{(1)}, \theta^{(2)}) = \theta^{(1)} + \theta^{(2)} Y_{t-1}$, where for simplicity we set $\theta^{(1)} = 0$, $\theta^{(2)} = 1$, i.e. $\varphi(Y_{t-1}; 0, 1) = Y_{t-1}$.

As the nonparametric estimate of the interpolation forecast for Y_t, we take the following modification of estimate (2):

$$\hat{Y}_t = \hat{r}(Y_{t-1}) = \frac{\sum\limits_{j \geq 2, j \neq t} Y_t K\left(\dfrac{Y_{t-1} - Y_{j-1}}{h_n}\right)}{\sum\limits_{j \geq 2, j \neq t} K\left(\dfrac{Y_{t-1} - Y_{j-1}}{h_n}\right)}. \tag{14}$$

The combined estimate, for which $\varphi(Y_{t-1}; 0, 1) = Y_{t-1}$, takes the form

$$\bar{Y}_t = \hat{R}_{\hat{\lambda}_B}(Y_{t-1}) = (1 - \hat{\lambda}_B(Y_{t-1}))\hat{r}(Y_{t-1}) + \hat{\lambda}_B(Y_{t-1}) Y_{t-1}, \tag{15}$$

where

$$\hat{\lambda}_B(Y_{t-1}) =$$

$$= \frac{\sum_{j=1}^{B}(\hat{r}(Y_{t-1};Y^*_{j,1},...,Y^*_{j,n-1}) - \hat{r}(Y_{t-1}))(\hat{r}(Y_{t-1};Y^*_{j,1},...,Y^*_{j,n-1}) - Y_{t-1})}{\sum_{j=1}^{B}[\hat{r}(Y_{t-1};Y^*_{j,1},...,Y^*_{j,n-1}) - Y_{t-1},)]^2}. \quad (16)$$

Based on the prices of stocks $Y_1,...,Y_n$, formulas (14) and (15), the estimates of forecasts by one step for price Y_{n+1} are defined as follows:

$$\hat{Y}_{n+1} = \hat{r}(Y_n;Y_1,...,Y_{n-1}) = \frac{\sum_{i=2}^{n} Y_i K\left(\frac{Y_n - Y_{i-1}}{h_n}\right)}{\sum_{i=2}^{n} K\left(\frac{Y_n - Y_{i-1}}{h_n}\right)}. \quad (17)$$

$$\bar{Y}_{n+1} = \hat{R}_{\hat{\lambda}_B}(Y_n) = (1 - \hat{\lambda}_B(Y_n))\hat{r}(Y_n) + \hat{\lambda}_B(Y_n)Y_n, \quad (18)$$

where

$$\hat{\lambda}_B(Y_n) =$$

$$= \frac{\sum_{j=1}^{B}(\hat{r}(Y_n;Y^*_{j,1},...,Y^*_{j,n-1}) - \hat{r}(Y_n;Y_1,...,Y_{n-1})(\hat{r}(Y_n;Y^*_{j,1},...,Y^*_{j,n-1}) - Y_n)}{\sum_{j=1}^{B}[\hat{r}(Y_n;Y^*_{j,1},...,Y^*_{j,n-1}) - Y_{n-1},)]^2}. \quad (19)$$

Let there be $n+L$ stock prices. Estimates of forecasts $\hat{Y}_{n+2}$ and $\bar{Y}_{n+2}$ will be constructed at n prices $Y_3,...,Y_{n+1}$ by formulas (17) and (18). Similarly, at n prices, shifting by the required number of steps, make forecasts $\hat{Y}_{n+3},...,\hat{Y}_{n+L}$ and $\bar{Y}_{n+3},...,\bar{Y}_{n+L}$.

The quality of identification and forecasting will be characterized by means of the average relative errors $\delta_{real}(\hat{r})$ and $\eta_{real}(\hat{r})$:

$$\delta_{real}(\hat{r}) = \frac{1}{n-1}\sum_{i=2}^{n}\frac{|Y_i - \hat{r}(Y_{i-1})|}{Y_i}100\%, \quad \eta_{real}(\hat{r}) = \frac{1}{L}\sum_{i=n+1}^{n+L}\frac{|Y_i - \hat{r}(Y_{i-1})|}{Y_i}100\%.$$

Consider the case $n = 100$. In Fig. 3 there are presented the results of identification and prediction for the combined model using estimates (15) and (18), and the behavior of the weight coefficients (16) and (19) are shown in Fig. 4.

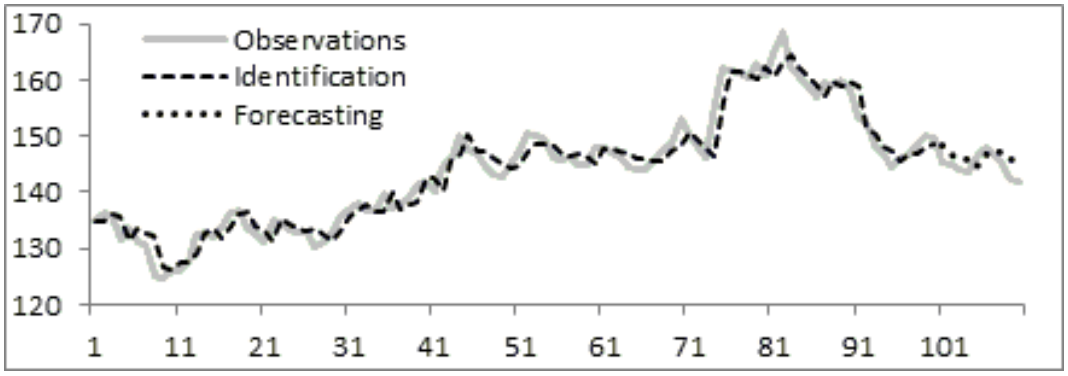


Figure 3: Identification and forecasting using combined estimates (15) and (18)

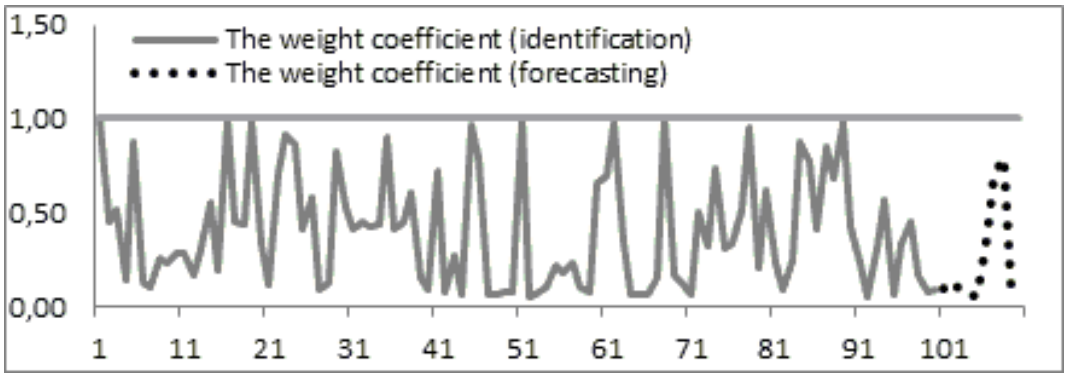


Figure 4: Plot of the dependence of the estimates of the weight coefficients (16) (identification) and (19) (forecasting)

For different volumes of observations, Table 2 gives the average relative errors of identification and prediction.

Table 2: Average relative errors of identification and prediction

n	10	50	100
$\delta_{real}(\hat{r})$	1.87	1.40	1.34
$\delta_{real}(\hat{R})$	1.67	1.29	1.31
$\eta_{real}(\hat{r})$	1.90	2.80	1.16
$\eta_{real}(\hat{R})$	1.41	1.20	1.01

From the results obtained, in practice it is preferable using the combined evaluation in comparison with the nonparametric estimate, especially in the case of small sample sizes.

Conclusions

In this paper, the problem of identification of a stochastic object by means of a combined estimate is considered, which is a weighted sum of the nonparametric estimate of the regression and some function given by the researcher. Adaptive combined estimates are constructed on the basis of which algorithms for predicting static and dynamic objects are proposed.

Based on the results of numerical simulation, the advantage of adaptive combined estimates is shown in comparison with nonparametric regression estimates for small samples sizes and a large noise level. The expediency of applying the proposed approach in practice is illustrated in the analysis of the prices of Gazprom's stocks for 2016.

Acknowledgements

This work was supported by Tomsk State University Competitiveness Improvement Program for 2013-2020.

References

[1] Dmitriev Yu.G., Koshkin G.M. (1987). On the Use of a Priori Information in Nonparametric Regression Estimation. *IFAC Proceedings Series*. Vol. **2**, pp. 223–228.

[2] Dmitriev Yu.G., Koshkin G.M. (1987). Using Additional Information in Nonparametric Estimation of Density Functionals. *Automat. and Remote Control*. Vol. **48**, No. **10**, pp. 1307–1316.

[3] Dmitriev Yu.G., Skripin S.V. (2012). On the Combined Estimate of the Probability of Failure-free Operation for a Full Sample *Journal of Control and Computer Science. Tomsk State University*. No. **4(21)**, pp. 32–38.

[4] Kitaeva A.V., Koshkin G.M. (2009). Recurrent Nonparametric Estimation of Functions from Multidimensional Functional Density and their Derivatives. *Automat. and Remote Control*. Vol. **70**, No. **3**, pp. 389–407.

[5] Kitaeva A.V., Koshkin G.M. (2009). Kernel Estimates of Basic Functionals for Dependent Observations. *Bulletin of the Tomsk Polytechnic University*. Vol. **314**, No. **2**, pp. 26–31.

[6] Koshkin G.M., Lukov V.Yu., Piven I.G. (2016). Nonparametric Algorithms of Identification and Prediction in the ARX-Models. *Proceedings. The Second International Symposium on Stochastic Models, in Reliability Engineering, Life Science, and Operations Management (SMRLO 2016)*. Beer Sheva, Israel. Conference Publishing Services The Institute of Electrical and Electronics Engineers, pp. 620–623.

[7] Koshkin G.M., Tarasenko F.P. (1988). Nonparametric Algorithms for Identifying and Control of Continuous-discrete Stochastic Objects. *8-th IFAC-IFORS Symposium on Identification and System Parameter Estimation*. Beijing. Pargamon Press, No. **2**, pp. 882–887.

[8] Nadaraya E. (1964). On Estimating of Regression. *Theory Probab. Appl.* Vol. **9**, No. **1**, pp. 141–142.

[9] Skripin S.V. (2008). Properties of a Combined Regression Estimate for Finite Sample Sizes. *Bulletin of the Tomsk Polytechnic University*. Vol. **313**, No. **5**, pp. 10–14.

[10] Vasiliev V.A., Koshkin G.M. (1998). Nonparametric Identification of Autoregressions. *Theory Probab. Appl.* Vol. **43**, No. **3**, pp. 507–517.

[11] Watson G.S. (1964). Smooth Regression Analysis. *Sankhya. Indian J. Statist.* Vol. **A26**, pp. 359–372.

Nonparametric Prediction Model of Real Estate Value

Maksim A. Denisov [1], Anna A. Korneeva [1], Oleg A. Ikonnikov [2]
[1] *Siberian Federal University, Krasnoyarsk, Russian Federation*
[2] *Siberian State University of Science and Technology named after M.F.Reshetnev, Krasnoyarsk, Russian Federation*
e-mail: max_denisov00@mail.ru, korneeva_ikit@mail.ru, ik_ol@mail.ru

Abstract

The paper considers the issue of estimating the value of real estate through a sample of observations using nonparametric estimation of the regression function. The problem is solved on the base of the sample containing data on one-room apartments in Krasnoyarsk city. Computational experiments shows the high accuracy of the obtained estimate. The information system for estimating the value of real estate is developed.

Keywords: a priori information, nonparametric identification, forecasting.

Introduction

Real estate has a significant influence on human life and acts as a valuable economic resource. In this regard, the methods of estimating the value of real estate are being actively studied. For example, in papers [1, 2] consider the question of constructing a parametric model for estimating the cost of housing. Afterwards, regression and factor analysis are applied to the models in order to identify significant or insignificant factors affecting the final model output value. This article centres around alternate approach to estimating the cost of housing - nonparametric approach. Advantages and disadvantages of both methods considers below.

The cost of an apartment is stochastic and depends on a large number of factors including the economic situation in general. Therefore, it is quite difficult to estimate the final cost of habitation at the moment. When a person faces the problem of buying an apartment, first of all, the budget plays a significant role.

Taking into account the problem of complex forecasting of prices, a person who is far from understanding the peculiarities of the behaviour of the housing market will not immediately be able to navigate and choose the best option for him. For such cases the cost forecasting of an apartment and providing information in an accessible form inclusive of the client personal needs and desires are topical tasks.

To solve various kinds of problems associated with forecasting prices or predicting the variables of a process or phenomenon regression analysis methods are used. Such methods involve studying the dependence of output variables on input variables. An important problem in this case is the choice of a parametric, or nonparametric approach to estimating regression dependence. The first case is applicable if the model structure of the investigated process or phenomenon can be given. In other words, it is represented as a known functional expression, described by a finite set of

parameters. The second case is used when there is a lack of priori information when the object is represented as a "black box" and there is no way to define its structure for sure.

Besides that, nonparametric analysis allows to predict new observations. Often, we have already had a some kinds of data set, but, unfortunately, not always they represent a complete picture of the process or phenomenon. Anyway, it is quite possible to predict the value we need based on the priori information we have even if it describes the process or phenomenon noncompletely. The results of parametric methods in this matter are too restrictive to obtain reasonable explanations for the observed phenomena, while nonparametric methods give better results [3].

In addition to this, the parametric systems associated with forecasting are most often difficult to implement, especially in the conditions of insufficient a priori information. Their disadvantage consists in a cumbersome mathematical description. This complicates the process of mathematical modelling and design of an adequate model of the investigated process or phenomenon. Therefore, the more optimal solution is to apply a nonparametric estimate, where there is no need to know the exact structure.

1 The Problem Statement

The paper centres around the problems of constructing nonparametric model and forecasting of real estate value.

A sample, which is used in the paper, contains information about 1358 numbers of the cost of one-room apartments in Krasnoyarsk. The sample does not contain omissions and overshoots. The sample consists of the following characteristics description of the apartments: total apartment space – u_1, kitchen size – u_2, location (district) – u_3, floor – u_4, walling material of the apartment – u_5 and layout – u_6. The output variable is the apartment price x. It should be mentioned that the kitchen size, the living space and the price are quantitative characteristics; the remaining characteristics are nominal.

2 The Method of Solution

As the regression estimator was taken nonparametric the Nadaraya-Watson estimation, which is presented below:

$$\hat{x}(u) = \frac{\sum\limits_{i=1}^{s} x_i \Phi\left(\dfrac{u - u_i}{c_s}\right)}{\sum\limits_{i=1}^{s} \Phi\left(\dfrac{u - u_i}{c_s}\right)}, \tag{1}$$

where $\{x_i, u_i, i = \overline{1, s}\}$ is initial sample of observations, c_s is estimator bandwidth of a kernel, $\Phi(\bullet)$ is kernel function and s is sample size.

A kernel $\Phi(\bullet)$ has the parabolic form as it gives more accurate results relative to the standard rectangular and triangular ones [4]. The form of the parabolic kernel is presented below:

$$\Phi^{(q)}(z) = \begin{cases} 0.75(1-|z|)^2, & |z| \le 0, \\ 0, & |z| < 0, \end{cases} \tag{2}$$

where $z = \dfrac{u - u_i}{c_s}$.

Kernel (2) is used only for input quantitative characteristics. For nominal characteristics the kernel looks as follows:

$$\Phi^{(n)}(z) = \begin{cases} 1, & u = u_i, \\ 0, & u \ne u_i. \end{cases} \tag{3}$$

On the basis of the equations (1)-(3), the estimator is used in this paper takes the form:

$$\hat{x}(u) = \frac{\sum\limits_{i=1}^{s} x_i \prod\limits_{j=1}^{2} \Phi^{(q)}\left(\dfrac{u_j - u_{ji}}{c_s}\right) \prod\limits_{j=3}^{6} \Phi^{(n)}(u_j - u_{ji})}{\sum\limits_{i=1}^{s} \prod\limits_{j=1}^{2} \Phi^{(q)}\left(\dfrac{u_j - u_{ji}}{c_s}\right) \prod\limits_{j=3}^{6} \Phi^{(n)}(u_j - u_{ji})}. \tag{4}$$

The accuracy of the model is evaluated in accordance with the following expression:

$$W = \frac{1}{s} \sum_{i=1}^{s} |x_i - \hat{x_i}|, \tag{5}$$

where x_i is output of the object, $\hat{x_i}$ is output of the model.

3 Computational Experiments

At the first stage, it is necessary to find the optimal bandwidth c_s for nonparametric estimation in the light of all the factors affect the cost of housing. Bandwidth founds using cross-validation method. Below the Fig. 1 illustrates dependence of the error (5) on estimator bandwidth c_s.

It can be seen that the minimum of the mistake is reached in the interval from 13 to 13.5. To define an exact value of estimator bandwidth reduce part of the results of the calculations to a table 1.

Based on the results given in table 1, it can be confirmed that the optimal value $c_s = 13.1$, as its the modelling error is minimal.

Upon the above results, construct a nonparametric estimate (4) for the optimal bandwidth $c_s = 13.1$, taking into account all characteristics that affect the price. The simulation results are presented in Fig. 2.

Based on the results shown in Fig. 2, it can be seen that in some cases a nonparametric estimate cannot be calculated. In particular, the value is unknown for 37 residential objects. That is because the initial sample does not have enough

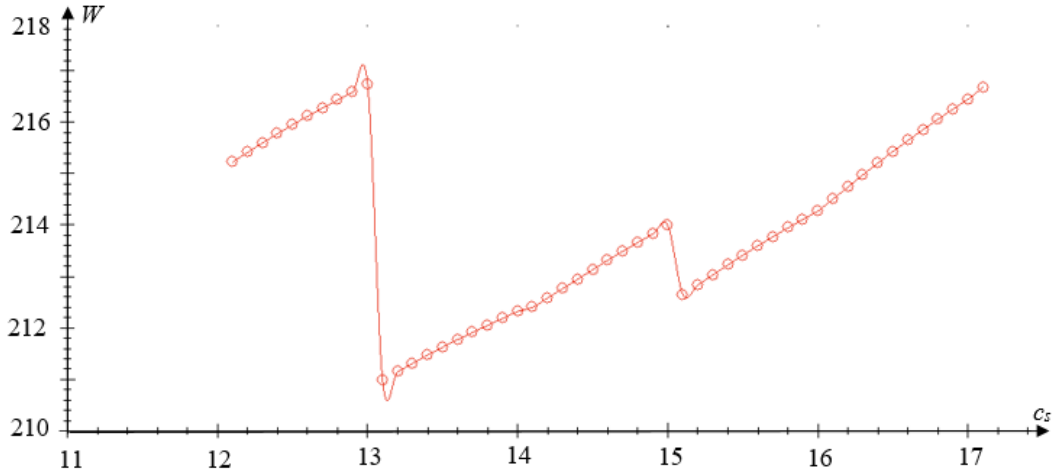


Figure 1: Dependence of the error (5) on the value of estimator bandwidth c_s

Table 1: Numerical modelling error values for each estimator bandwidth

c_s value	Modelling error W value
13	216.73
13.1	210.98
13.2	211.15
13.3	211.31
13.4	211.47
13.5	211.62

information for forecasting, and also because the factor imposes a limitation on the points incoming under the bell of the kernel function.

The problem, which is encountered above, initiated for further study of the estimate (4). Optimal bandwidth c_s will be found similar to the method described earlier but without including characteristic u_3, which, as it turned out, strongly affects the final results of the model's output.

Fig. 3 shows the dependence of the bandwidth value c_s from the error (5). On this figure it is shown that the optimal bandwidth value is $c_s = 7.1$ in compliance with modelling error (5) $W = 194.87$.

Following the results above, construct a nonparametric estimate for the optimal bandwidth parameter. Below, in Fig. 4, the simulation results are demonstrated.

Therefore, based on the results presented in Fig. 4, it can be seen that the nonparametric model constructed without including characteristic u_3 describes the object more accurately. In this case, the value is unknown only for 6 residential objects.

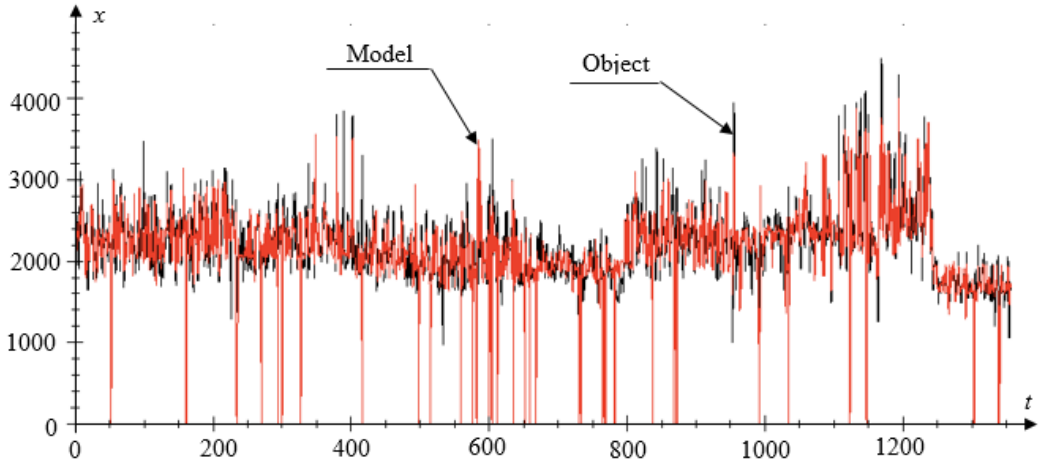


Figure 2: Dependence of the output of the model and the object on the time bar

The presented estimate (4) and the optimal bandwidth c_s, calculated above, in the light of the factor u_3, formed the basis to create the information system in which it is possible to set the parameters of the apartment at its own discretion. As a result, it is available to get the cost of housing relative to your preferences based on the training sample.

For example, take the 2nd district, the total area of the apartment is 42 m^2, kitchen area is 8 m^2, 1st floor, walling material of the apartment is 2nd, layout is 4th. As a result, the final cost is calculated equal to 2 million 269 thousand rubles.

Unfortunately, there are some cases in which the value of the price cannot be determined. As in the previous paragraph they are justified by the insufficient amount of information.

The client who chooses an apartment does not always have the opportunity to know the parameters of his future housing for sure. There are situations in which the parameters of the apartment must be changeable. That means, the parameters should be set interval, depending on the wishes of the client. It is necessary in order not to be limited, for example, only to 9 squares meters of the kitchen, but to have an approximate price of apartments with kitchens from 9 m^2 to 13 m^2. And this possibility is taken into account in the system.

For instance, from the previous example, we are satisfied with 2nd district, walling material of the apartment is also 2nd and layout is 4th, but it would be acceptable to know the cost of housing with total area of the apartment from 42 m^2 to 56 m^2, kitchen area from 8 m^2 to 15 m^2 and consider the apartments not only on the 1st floor, but also on the 2nd. Thus, considering all our desires, the total cost turned out to be lower and became equal to 2 million 150 thousand rubles.

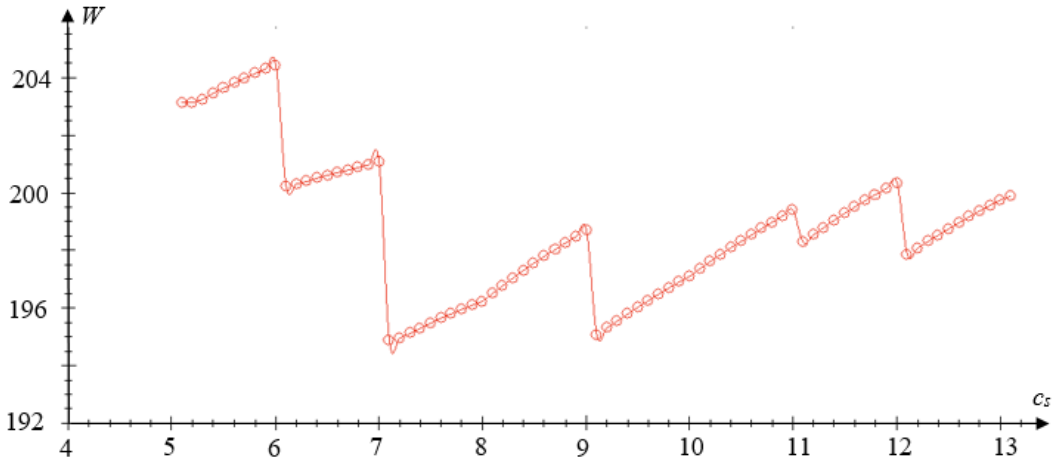


Figure 3: Dependence of the error (5) on the value of estimator bandwidth c_s

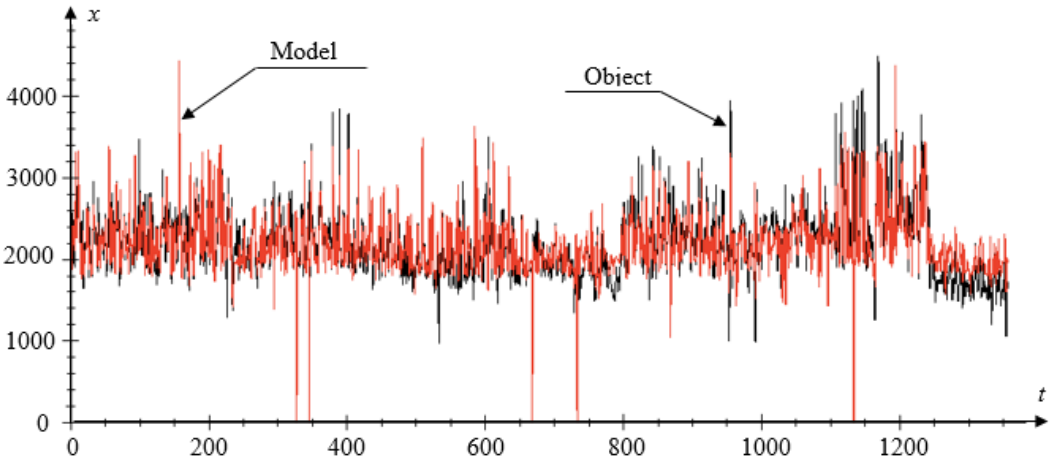


Figure 4: Dependence of the output of the model and the object on the time bar

Conclusions

The information system has been developed using the nonparametric algorithm to predict the real estate value. In this system, it is possible to adjust the parameters of the apartment to be point and interval.

It turned out that the location characteristic, in which the apartment is located, affects the nonparametric estimator used in this paper. Without this factor a more accurate estimate of real estate value has been built.

In the future, we plan to compare nonparametric and parametric estimation methods on the basis of the sample presented in the paper.

References

[1] Renner A.G., Stebunova O.I. Modelirovanie stoimosti zhil'ja na vtorichnom rynke zhil'ja [Modelling the cost of housing in the secondary market of housing]. *Vestnik Orenburgskogo gosudarstvennogo universiteta,* 2005, no. 10-1, pp. 179-182.

[2] Juferova N.Ju., Senashov S.I., Surnina E.V. Informacionnaja sistema ocenki stoimosti kvartir na vtorichnom rynke zhil'ja kak instrument upravlenija investicijami [Cost estimation of apartment's information system at secondary housing markets as a management tool for investments]. *Vestnik Sibirskogo gosudarstvennogo ajerokosmicheskogo universiteta im. akademika MF Reshetneva,* 2009, no.4, pp. 219-222.

[3] Härdle W. (1989). *Applied nonparametric regression.* Cambridge university press, Cambridge, p. 17.

[4] Ruban A.I. *Metody analiza dannyh* [Methods of data analysis]. Krasnojarsk, IPC KGTU Publ., 2004. pp. 121-124.

Nonparametric Model of Linear Dynamical Systems of High Orders

O. A. Ikonnikov
Siberian State University named after academician M. F. Reshetnev, Krasnoyarsk, Russia
e-mail: ik_ol@mail.ru

Abstract

The article considers linear dynamic systems of high order, as the object of modeling. The urgency of solving such problems is reasoned. The work presents as mathematical model and results of numerical research of the constructed model.

Keywords: Nonparametric model, linear dynamical systems, high order, nonparametric approximation

Introduction

Linear dynamical systems (LDS) are the dominant link in the range of research related to the tasks of regulation and control. Recently, due to the increasing need for the establishment of various control systems, the problem of mathematical modelling is becoming essential. Nowadays there are many ways of constructing mathematical models, the effectiveness of each of which depends on the specific task. Therefore, one of the key problems in this area is the creation of a model with a special versatility and simplicity. In this paper the nonparametric approach to constructing the mathematical model of the LDS based on the Duhamel integral using stochastic approximations is considered [3].

There is a lot of literature devoted to the problem of modeling and identification in general. However, the research performed by the authors, as a rule, oriented on processes whose order is low. The reason of it is the lack of the necessary requirements of the research such processes due to the rather narrow field of practical application, as well as a significant increase in complexity to use for the calculation algorithms.

In the study of such processes is inevitable there are many problems associated with the specifics of the problem being solved. One of such problems is the problem of rational choice of the numerical method for solving differential equations of high orders. So, experimentally in the process, it was found that the numerical finite difference method, which is successfully applied for the solution of low order differential equations, is not suitable in applying it to high order equations due to the various errors are occurred. In this regard, has been proposed another numerical method as more suitable for such tasks, from the (m, k) – type Rosenbrock methods [4].

1 Nonparametric model

In control theory, to describe the DS work, at the entrance of which there are signals of arbitrary shape, is very often used the so-called Duhamel integral (convolution)[1]:

$$x(t) = k(0)u(t) + \int_0^t k^{'}(t-\tau)u(\tau)d\tau = k(0)u(t) + \int_0^t h(t-\tau)u(\tau)d\tau, \qquad (1)$$

where k(0),k(t -τ) is the transition function of the system (the signal produced at the output of the system when applied to its input a unit step 1(t)), the weigh function of the system (the signal produced at the output of the system when applied to its input a single pulse),τ is a variable of integration.

A mathematical model constructed on the basis of this integral using a nonparametric approximation regression curve has the following form [3]:

$$x_s(t) = k(0) \cdot u(t) + \frac{1}{s \cdot c_s} \cdot \sum_{i=1}^{s} \sum_{j=1}^{t\Delta\tau} k_i \cdot H^{'}(\frac{t - \tau_j - t_i}{c_s}) \cdot u(\tau_j)\Delta\tau, \qquad (2)$$

where s is the sample volume, Cs is the fuzzifying parameter, satisfying the following conditions:

$$c_s > 0, s = 1, 2, ...; \lim_{s\to\infty} c_s = 0; \lim_{s\to\infty} (s \cdot c_s) = \infty. \qquad (3)$$

H(.) - the function from the class of bell-shaped, satisfies the following conditions for convergence:

$$\int_{\Omega(x)} H^{'}(u)du = 0, , c_s \int_{\Omega(x)} uH^{'}(u)du = -1 \qquad (4)$$

$\Delta\tau$ - step numerical integration, the magnitude of which depends on the accuracy of taking the integral in the model.

2 Numerical research

The numerical research of linear dynamic processes of the tenth, fifteenth, and nineteenth orders of the methods described above are performed. The results of the performed numerical research are displayed in the following figures. Here is the process in the object shown in solid line, and the output of the model – dotted (these graphic images do not differ from each other). Tc is the machine computation of the model. The counting process was stopped when the amplitude of the process in the current time was approximately 1% of the maximum.

Let's see what happens if we observe the process five times more often, that is $h = h/5$.

As can we see from fig.2, process loses its stability when the reduction sampling. The situation is complicated in the study of systems of higher order, which in turn is a consequence of the occurrence of large errors of the method for such cases.

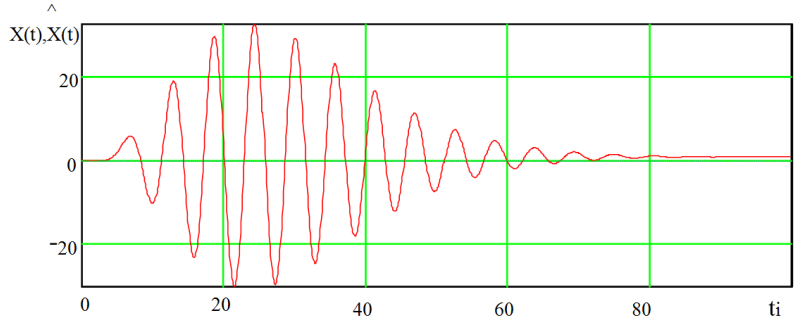


Figure 1: Transition function of the 10 - th order object (numerical finite difference method, s = 1000, h = 0.1 s., Tc = 8 min)

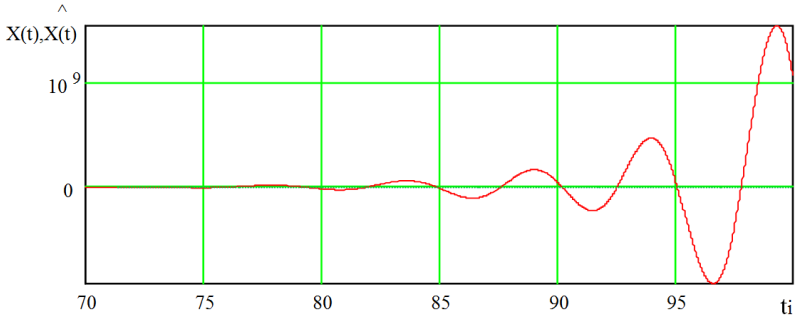


Figure 2: Transition function of the 10 - th order object (numerical finite difference method, s = 5000, h = 0.02 s)

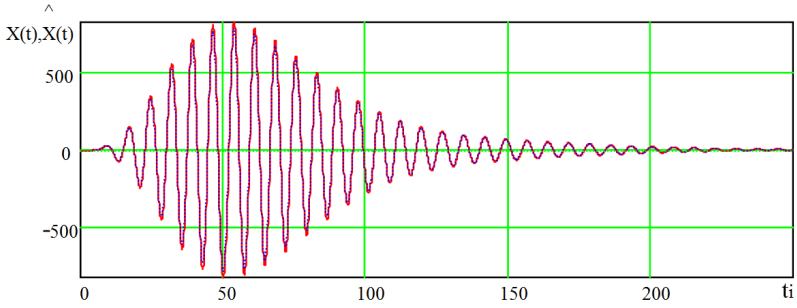


Figure 3: Transition function of the 10 - th order object (numerical finite difference method, s = 2500, h = 0.1 s., Tc=65 min)

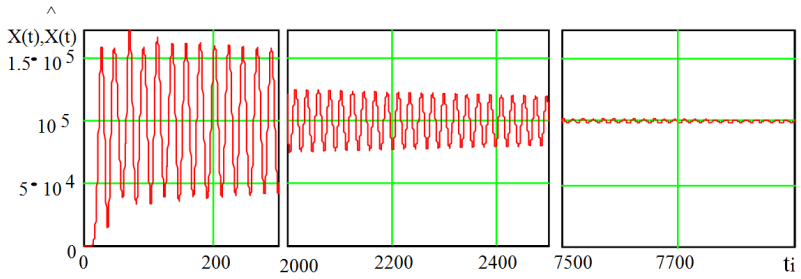


Figure 4: Fragmented image of 15-th order object transition function (method of (m, k) - type methods Rosenbrock series, s = 80000, h = 0.1 s., Tc = 4355 min)

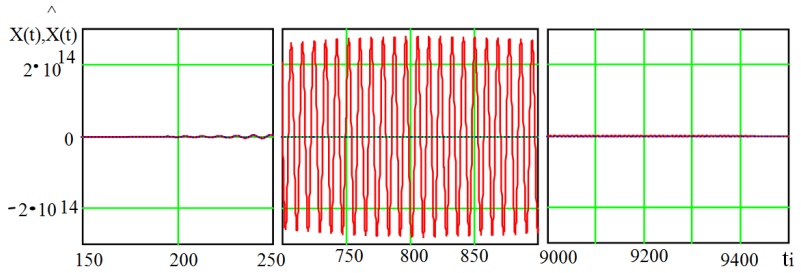


Figure 5: Fragmented image of 19-th order object transition function (method of (m, k) - type methods Rosenbrock series, s = 95000, h = 0.1 s., Tc = 5569 min)

The following figures will present the results of numerical simulation by the method of (m, k) - type methods Rosenbrock series.

It should be noted that in Fig.1,2 and 3 shows the transition functions of objects that have in their description exactly the same differential equation (same coefficients). Further, when reducing the sampling interval (even significantly), the process (Fig.3) practically does not change its shape, and in the course of significant growth h, the graph gradually takes on the appearance of some broken line, which is a natural for this kind of task. The transitional processes of objects of higher order are shown below.

The responses to various inputs were received.

The results of these research are convinced that the nonparametric theory is universal in the sense of ordinal lack of certainty in the differential equations describing the object (curves are practically the same in the graphs).

But despite the positive aspects of noparametric, we observe and acknowledge the fact that with increasing order differential equations of the investigated object are significantly growing as the volume of samples, and adjusting processes time, which leads to even more substantial increase in computer time spent on the implementation of mathematical calculations. Below shows the graphical dependences of the above parameters from the order (n) of solved differential equation, which allow to visually see what difficulties arise when we deal with such processes.

As can we see from fig. 6, 7, when increasing the order from 10 to 9, the sample

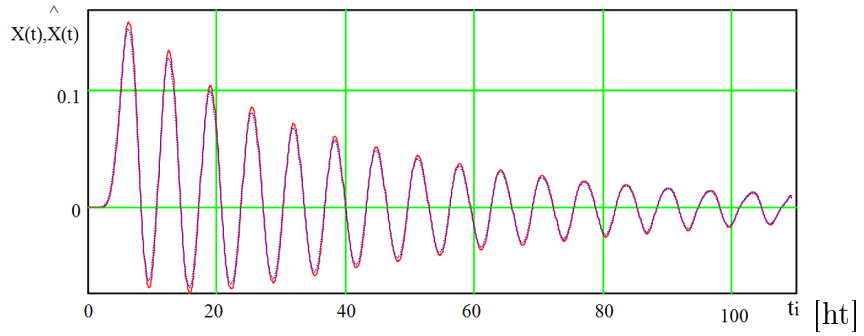


Figure 6: The 10-th order object and its model responses to sin(t)*cos(t), (method of (m, k) - type methods Rosenbrock series, s = 1200, h = 0.1 s., Tc = 15 min)

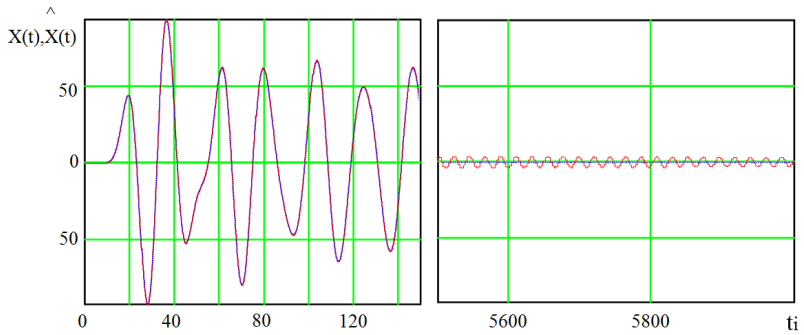


Figure 7: The fragments of 10-th order object and its model responses to sin(t) to (method of (m, k) - type methods Rosenbrock series, s = 80000, h = 0.1 s., Tc = 5866 min)

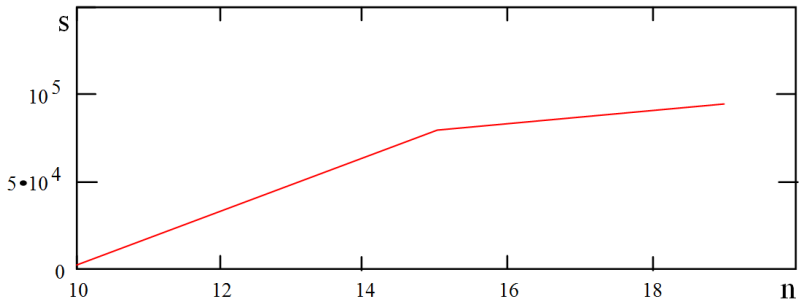


Figure 8: The calculated function of the sample volume from the differential equation order

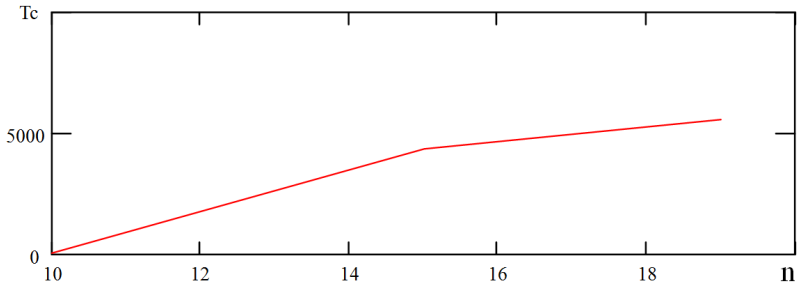


Figure 9: The calculated function of computing time from the differential equation order

volumes, and, respectively, control time increased almost 80 times, computing time increases approximately 170 times. All of this creates some inconvenience in the research of such processes and, in this regard, the organization of the account some other way to reduce TS is planned.

References

[1] Feldbaum A. A., Dadykin A. D., Manovtsev P. A., Mirolubov N.N. Theoretical foundations of communication and control. Moscow. *state publishing house of physical-Mat. literature*, 932p.

[2] Grop D. (1989). Methods of identification systems. Moscow. *Publishing house Mir*, 304p.

[3] Medvedev A. V. (1975). *Identification and Control for Linear Dynamic Systems of Unknown Order. Optimation Techniques*. -In: IFIP Technical Conference. Berlin-Heidelberg-New York, Springer-Verlang.

[4] Novikov E. A., Shitov Yu. A. (1997). Research (m, k) – methods of solving stiff systems with one and two calculations of the right side. Krasnoyarsk. *Preprint No. 15 VTS so an USSR*.

[5] Pirumov U. G. (2008). *Numerical methods*. Moscow. *Publishing house MAI*, 188p.

Robust Polytomous Logistic Regression Based on Bianco and Yohai Estimator

Yurii V. Chernikov and Daniil V. Lisitsin
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: jskar93@mail.ru, lisitsin@ami.nstu.ru

Abstract

This paper is devoted to robust parameter estimation for polytomous (multinomial) logistic regression. Several versions of the Bianco and Yohai estimator are considered. The estimator provides robustness against deviation of observations distribution from an postulated one. In paper the formulas of estimator versions are described, results of a Monte Carlo study are presented. Also the joint use of the Bianco and Yohai estimator and the estimator optimal with respect to weighed L_2-norm of Hampel's influence function is studied.

Keywords: parameter estimation, robustness, polytomous logistic regression, Bianco and Yohai estimator, influence function, Monte Carlo method.

Introduction

The classical statistical 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. But this problem can be solved by using robust estimators [10].

Generally robustness theory is used for the quantitative continuous random variables modeling. Much less attention is paid to the modeling of nominal variables and existing approaches often are of semi-heuristic nature. The last statement is confirmed by a few papers devoted to robust estimation of parameters of the regression model with a polytomous (multinomial) response (see, for example [1, 8, 12, 19, 20]).

Recently we develop methods for parameter estimation of the polytomous logistic regression [13] which are based on the general theory of asymptotically optimal estimation of unknown model parameters from multivariate nonhomogeneous incomplete data [4, 14]. At the bottom of this theory we find synthesis of approach by F. Hampel [10] which is associated with the influence function and approach by A.M. Shurygin [18] which is associated with the Bayesian point-mass contamination model distribution. The resulting methods are optimal with respect to weighed L_2-norm of the influence function and robust against the deviation of the actual distribution of observations from a postulated (or ideal) one. Previously, this theory is applied to cases with nonhomogeneous quantitative (including count), qualitative, mixed data, and also in the presence of missing data [4, 5, 13, 15].

Asymptotically optimal estimator is defined by a system of nonlinear equations which is often having many solutions for the polytomous logistic regression. Therefore, the choosing an estimate as one of solutions can be problematic. A frequently used approach in this case is the choice of an appropriate initial approximation with a

computing of estimate by some local solution method. Such an initial approximation usually is a good (in some sense) robust estimator.

As such an estimator for logistic regression model, the Bianco and Yohai estimator [1] can be used; it has a definition through optimization, what is convenient for calculations. In addition to [1], there are papers [3, 6, 19], where propose variants of the Bianco and Yohai estimator. Another option is proposed in this paper.

Since the Bianco and Yohai estimator is of interest in itself, in this paper, all variants are researched by the Monte Carlo method for polytomous logistic regression. There are no similar researches of the corresponding variants in [1, 3, 6, 19].

In paper also the joint use of the Bianco and Yohai estimate (in a role of the initial approximation) and asymptotically optimal estimate is studied.

1 The Bianco and Yohai Estimator

Let the distribution of a discrete random variable ζ_i under the ith observation be given by a set of model probabilities

$$\mathsf{P}\{\zeta_i = j | x_i, \alpha\} = p_j(x_i, \alpha), \quad i = 1, \ldots, N, \quad j = 1, \ldots, J,$$

where x_i is a vector of deterministic input variables, α is a vector of parameters.

For modeling dependence of the nominal output variable (response) from input variables the polytomous (multinomial) logistic regression is often used. Corresponding probabilities are of the form

$$p_j(x, \alpha) = \frac{\exp\left[f^{\mathsf{T}}(x)\alpha_j\right]}{1 + \sum_{k=1}^{J-1} \exp\left[f^{\mathsf{T}}(x)\alpha_k\right]},$$

where $f(x)$ is a vector of regressors (functions of input variables), α_j is the jth subvector of the vector α (subvectors $\alpha_j, j = 1, 2, \ldots, J-1$, do not intersect), $\alpha_J = 0$, 0 is a null vector, namely $\alpha^{\mathsf{T}} = (\alpha_1^{\mathsf{T}}, \ldots, \alpha_{J-1}^{\mathsf{T}}, 0^{\mathsf{T}})^{\mathsf{T}}$.

The maximum likelihood (ML) method is the classical method of estimating the parameter vector α. The ML estimate is obtained by maximizing the log-likelihood function, namely $\hat{\alpha} = \arg\max_a L(a)$, where $\hat{\alpha}$ is an estimate of α, $L(a) = \sum_{i=1}^{N} \ln p_{z_i}(x_i, a)$, z_i is the observation of ζ_i. The ML estimator is a particular case of M-estimators which are defined by minimization, namely $\hat{\alpha} = \arg\min_a Q(a)$, where $Q(a) = \sum_{i=1}^{N} \rho(z_i | x_i, a)$, ρ is a loss function. The loss function of the ML estimator for the ith observation defined as $\rho(z_i | x_i, a) = -\ln p_{z_i}(x_i, a)$.

The ML estimator is often unstable, when the model assumptions are violated. In such cases, using robust estimation methods is more advisable.

One approach to constructing robust estimates is proposed by Bianco and Yohai in [1]. This method consists in modifying the loss function; new loss function has the form $\rho(z_i | x_i, a) = \mathrm{R}(-\ln p_{z_i}(x_i, a))$, where $\mathrm{R}(t)$ is some bounded, differentiable, and nondecreasing function. Such modification leads to a violation of an asymptotic unbiasedness of an estimator (for the ML estimator this condition is satisfied

automatically). Therefore this modification must be accompanied by an additional correction that ensures this condition. The correction is carried out by including an additional term which has the form $\sum_{j=1}^{J} G(p_j(x,\alpha))$, where $G(p) = \int_0^p \mathrm{R}'(-\ln u)\,du$. As a result we receive minimization function

$$Q(a) = \sum_{i=1}^{N} \Big[\mathrm{R}(-\ln p_{z_i}(x_i,a)) + \sum_{j=1}^{J} G(p_j(x_i,a))\Big].$$

We represent a number of known and one new variants of the Bianco and Yohai estimator in the form of formulas for corresponding functions $\mathrm{R}(t)$, $G(p)$ (sometimes — accurate to inessential terms). For some variants, we presented also simpler dependence $\tilde{\mathrm{R}}(p) = \mathrm{R}(-\ln p)$. Each variant has a name in which the two first letters (BY) symbolize the names of the authors of the estimator.

1. BY [1] (original Bianco and Yohai estimator):

$$\mathrm{R}(t) = \begin{cases} t - t^2/2c, & t \leqslant c, \\ c/2, & t > c, \end{cases} \quad G(p) = \begin{cases} p + (p\ln p - p + \mathrm{e}^{-c})/c, & -\ln p \leqslant c, \\ 0, & -\ln p > c. \end{cases}$$

2. BYR [3] (from "radical", due to the presence of an argument under the radical sign; for more details, see [3]):

$$\mathrm{R}(t) = \begin{cases} t\mathrm{e}^{-\sqrt{d}}, & t \leqslant d, \\ -2\mathrm{e}^{-\sqrt{t}}(1+\sqrt{t}) + \mathrm{e}^{-\sqrt{d}}(2(1+\sqrt{d})+d), & t > d, \end{cases}$$

$$G(p) = \begin{cases} p\mathrm{e}^{-\sqrt{-\ln p}} + \mathrm{e}^{1/4}\sqrt{\pi}\Phi(\sqrt{2}(\frac{1}{2}+\sqrt{-\ln p})) - \mathrm{e}^{-1/4}\sqrt{\pi}, & p \leqslant \mathrm{e}^{-d}, \\ p\mathrm{e}^{-\sqrt{d}} + \mathrm{e}^{1/4}\sqrt{\pi}\Phi(\sqrt{2}(\frac{1}{2}+\sqrt{d})) - \mathrm{e}^{-1/4}\sqrt{\pi}, & p > \mathrm{e}^{-d}, \end{cases}$$

where $\Phi(t) = (\sqrt{2\pi})^{-1}\int_{-\infty}^{t} \mathrm{e}^{-u^2/2}\,du$ is c.d.f. of the standard Gaussian distribution.

3. BYP [6] (from "power", due to the presence of raised probability):

$$\mathrm{R}(t) = \frac{1-\mathrm{e}^{-\beta t}}{\beta}, \quad \tilde{\mathrm{R}}(p) = \frac{1-p^{\beta}}{\beta}, \quad G(p) = \frac{p^{\beta+1}}{\beta+1}.$$

4. BYLL [6] (from "log-logistic"):

$$\mathrm{R}(t) = -\ln(\eta + \mathrm{e}^{-t}), \qquad \tilde{\mathrm{R}}(p) = -\ln(\eta+p), \quad G(p) = p - \eta\ln(\eta+p).$$

5. BYG (c.d.f. of the Gaussian distribution is used):

$$\mathrm{R}(t) = \sqrt{2\pi}\left[\Phi(\gamma t) - 1/2\right]/\gamma, \quad G(p) = \sqrt{2\pi}\mathrm{e}^{(2\gamma^2)^{-1}}\left[1 - \Phi(\gamma(-\ln p) + 1/\gamma)\right]/\gamma,$$

6. BYS [19] (from "sech"):

$$\mathrm{R}(t) = 1 - \mathrm{sech}(\omega t), \quad \tilde{\mathrm{R}}(p) = 1 - 2p^{\omega}/(1+p^{2\omega}),$$

$$G(p) = 2p^{1+\omega}\left\{\frac{1}{p^{2\omega}+1} - \frac{1}{1+\omega}\,{}_2F_1\left[1, \frac{1+\omega}{2\omega}; \frac{1}{2}\left(3+\frac{1}{\omega}\right); -p^{2\omega}\right]\right\},$$

where ${}_2F_1(a, b; c; z)$ is a hypergeometric function. For two particular cases of BYS, function $G(p)$ has a simpler form. For $\omega = 1$ occurs $G(p) = 2p^2/(1+p^2) - \ln(1+p^2)$, for $\omega = 1/2$ occurs $G(p) = -4\sqrt{p} + 4\arctan(\sqrt{p}) + 2p\sqrt{p}/(p+1)$. Note that we simplified the original variant of BYS due to the fact that we assume the input variables to be deterministic, when in [19] they are random.

Each of the variant have positive parameter: c, d, β, η, γ, ω. For the first five variants, the ML estimator is limit as $c \to \infty$, $d \to \infty$, $\beta \to 0$, $\eta \to 0$, $\gamma \to 0$. The authors of [1, 3] used the values of the parameters $c = -\ln(0.03) \approx 3.51$, $d = 1/2$; the authors of [19] recommend $\omega = 1/2$. However, it is unlikely that in all cases these values are best.

2 Studying of the Bianco and Yohai estimator

A Monte Carlo study was conducted to compare the properties of variants of the Bianco and Yohai estimator. The study is based on generating samples with the further computation of the parameter estimates.

To solve the optimization problem, the Broyden—Fletcher—Goldfarb—Shanno method was used; the relative accuracy of the calculations was chosen 0.001. The optimization process started sequentially from a true value of parameters of model, from a ML estimate and from hundred random vectors in which each element has been uniformly distributed in $[-100, 100]$; the best of the found solutions was selected as an estimate. In a case when the solution did not exist (L_2-norm of the current approximation was not less than 200), a random point was regenerated.

A ideal polytomous logistic model with a number of categories $J = 3$, a vector of regressors $f(x) = (1, x)^\mathsf{T}$, $x_i \in [-1, 1]$, $i = 1, \ldots, N$, and vector of parameters $\alpha = (-10, 20, -8, -20)^\mathsf{T}$ was used.

The real distribution of observations was determined by the contamination model $\tilde{p}_j(x, \alpha) = (1 - \varepsilon)p_j(x, \alpha) + \varepsilon h_j(x)$, $j = 1, 2, 3$, where $0 \leqslant \varepsilon < 1$ is a contamination rate, $h_j(x)$, $j = 1, 2, 3$, are the probabilities of the contamination distribution. The contamination distribution was uniform. The sample size N was equal 300. The number of samples in the Monte Carlo method was equal 5500.

The estimation accuracy was determined by average value of the variable

$$s_m = (NJ)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{J} (p_j(x_i, \alpha) - \hat{p}_{jm}(x_i))^2,$$

where $\hat{p}_{jm}(x_i)$ is an estimate of $p_j(x_i, \alpha)$ based on the mth sample.

The accuracy of the ML estimate is 0.00109 (0.00001) for $\varepsilon = 0\%$, 0.0109 (0.0001) for $\varepsilon = 5\%$, 0.0242 (0.0001) for $\varepsilon = 10\%$, 0.0512 (0.0001) for $\varepsilon = 20\%$, 0.0943 (0.0001) for $\varepsilon = 40\%$ (estimated standard deviation of the accuracy is indicated in parentheses).

The accuracy values for the variants of the Bianco and Yohai estimator in Tables 1–3 are given. In a cell the upper number is an accuracy value, the lower number is an estimated standard deviation. Three values of the parameter of each variant were studied; the parameter value is indicated in parentheses after variant name.

Table 1: The Accuracy of BY and BYP

ε	BY(1)	BY(3.51)	BY(5)	BYP(0.5)	BYP(1)	BYP(1.5)
0%	0.00202	0.00211	0.00140	0.00132	0.00198	0.00216
	0.00003	0.00003	0.00002	0.00002	0.00003	0.00003
5%	0.00215	0.00217	0.00159	0.00161	0.00207	0.00218
	0.00003	0.00003	0.00002	0.00002	0.00003	0.00003
10%	0.00378	0.00241	0.00216	0.00278	0.00257	0.00293
	0.00004	0.00003	0.00003	0.00003	0.00003	0.00004
20%	0.0140	0.00472	0.00726	0.0105	0.00754	0.00932
	0.0001	0.00005	0.00007	0.0001	0.00007	0.00007
40%	0.0425	0.0287	0.0497	0.0535	0.0347	0.0355
	0.0001	0.0001	0.0003	0.0002	0.0001	0.0001

Table 2: The Accuracy of BYR and BYG

ε	BYR(0.2)	BYR(0.5)	BYR(1)	BYG(0.5)	BYG(0.75)	BYG(1)
0%	0.00122	0.00122	0.00120	0.00151	0.00239	0.00306
	0.00001	0.00001	0.00001	0.00002	0.00003	0.00004
5%	0.00218	0.00226	0.00229	0.00167	0.00240	0.00296
	0.00002	0.00002	0.00002	0.00002	0.00003	0.00004
10%	0.00592	0.00615	0.00638	0.00218	0.00260	0.00307
	0.00004	0.00005	0.00004	0.00003	0.00003	0.00004
20%	0.0200	0.0203	0.0213	0.00651	0.00498	0.00588
	0.0001	0.0001	0.0001	0.00006	0.00005	0.00006
40%	0.0706	0.0706	0.0744	0.0408	0.0275	0.0277
	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001

From Tables we see the advantage of all variants of the Bianco and Yohai estimator before the ML estimator in the presence of contamination, only BYS(0.5) is equivalent to the ML estimator for $\varepsilon = 40\%$. With increasing contamination rates the accuracy gain is ensured with decrease of the parameter value for BY, and with increase of the parameter value for BYP, BYG, BYS, BYLL. The accuracy of BYR weakly depends on the parameter value.

The best are the ML estimate and BYS(0.5) for $\varepsilon = 0\%$, BY(5), BYP(0.5), BYG(0.5), BYS(0.5) for $\varepsilon = 5\%$, BY(5), BYG(0.5), BYS(1), BYLL(0.3), BYLL(1), BY(3.51) for $\varepsilon = 10\%$, BYS(1) for $\varepsilon = 20\%$, BYS(2) for $\varepsilon = 40\%$.

Further we will show the usefulness of the Bianco and Yohai estimate in a role of the initial approximation for calculation of the estimate optimal with respect to weighed L_2-norm of influence function.

Table 3: The Accuracy of BYS and BYLL

ε	BYS(0.5)	BYS(1)	BYS(2)	BYLL(0.3)	BYLL(1)	BYLL(3)
0%	0.00108	0.00191	0.00509	0.00168	0.00186	0.00207
	0.00001	0.00003	0.00006	0.00002	0.00003	0.00003
5%	0.00168	0.00204	0.00497	0.00184	0.00195	0.00215
	0.00002	0.00003	0.00006	0.00002	0.00003	0.00003
10%	0.00395	0.00225	0.00482	0.00231	0.00237	0.00255
	0.00004	0.00003	0.00005	0.00003	0.00003	0.00003
20%	0.0175	0.00410	0.00547	0.00647	0.00652	0.00708
	0.0001	0.00005	0.00006	0.00006	0.00006	0.00007
40%	0.0943	0.0260	0.0232	0.0363	0.0337	0.0354
	0.0002	0.0002	0.0002	0.0002	0.0001	0.0002

Earlier we studied the generalized radical estimator (GRE) optimal under the weighting function $p_j^{1-\lambda}(x,\alpha)/\sum_{l=1}^{J} p_l^{1-\lambda}(x,\alpha)$, $j=1,\ldots,J$, where $\lambda \geqslant 0$ is a parameter [13]. The GRE of subvector α_j is defined by the system of equations

$$\sum_{i=1}^{N}\left[\delta_{jz_i} - p_j^{1+\lambda}(x_i,\hat{\alpha})/\sum_{k=1}^{J} p_k^{1+\lambda}(x_i,\hat{\alpha})\right] p_{z_i}^{\lambda}(x_i,\hat{\alpha}) \sum_{l=1}^{J} p_l^{1-\lambda}(x_i,\hat{\alpha}) f(x_i) = 0,$$

where δ_{jk} is the Kronecker symbol. The analogues of this estimator for the responses of various types also were studied in [2, 4, 5, 9, 11, 15, 16, 21].

Note that a weighting function defines the probability mass function of the random contamination point under Bayesian point-mass contamination model [4, 14, 15, 18]. It is remarkable that this distribution optimizes the Shannon's entropy (under $0 \leqslant \lambda \leqslant 1$) or the Kerridge's inaccuracy between this distribution and ideal distribution (under $\lambda \geqslant 0$), if the Kullback—Leibler divergence between this distribution and ideal distribution is bounded above [7, 16].

We will study the GRE with $\lambda = 1$ that leads to a constant weighting function and corresponds to entropy maximisation without the restriction of divergence; this estimator is analogue of the estimator of minimum variance sensitivity [17].

The BYS(1) has been computed with accuracy equal 0.0001, in the presence of 20% contamination. The estimate is $\hat{\alpha} = (-4.192, 8.669, -5.697, -13.53)^{\mathsf{T}}$; its characteristics: $Q(\hat{\alpha}) = -154.744$; L_2-norm of the residual vector corresponding to the system of equations defining GRE is 0.0008743; the estimation accuracy is $s_1 = 0.004858$. Then, with usage of this estimate in a role of the initial approximation, the GRE has been computed by the Newton's method with same accuracy. The estimate is $\hat{\alpha} = (-4.362, 8.954, -5.557, -13.10)^{\mathsf{T}}$; its characteristics: $Q(\hat{\alpha}) = -154.735$ (the function minimizing the BYS(1)); L_2-norm of the residual vector is 0.00004543, the estimation accuracy is $s_1 = 0.004707$. These two estimates are close so, that graphs of the dependencies of the probabilities estimates visually do not differ almost. We will note that the BYS(1) can be considered the GRE if to determine the accuracy equal 0.001 for the GRE calculation.

Conclusion

In this paper a number of variants of the Bianco and Yohai estimator for polytomous logistic regression are explored. The results of the research showed the advantage of this estimator before the maximum likelihood estimator. The dependence of the quality of estimation from the estimator parameter are explored, the comparison of all variants relative to the estimation accuracy is carried out.

The usefulness of the Binco and Yohai estimate in a role of the initial approximation for calculation of the estimate optimal with respect to weighed L_2-norm of influence function is demonstrated.

References

[1] Bianco A.M., Yohai V.J. (1996). Robust Estimation in the Logistic Regression Model. *Robust Statistics, Data Analysis, and Computer Intensive Methods*. Ed. by H. Rieder. Springer, New York, pp. 17-34.

[2] Broniatowski M., Toma A., Vajda I. (2012). Decomposable Pseudodistances and Applications in Statistical Estimation. *Journal of Statistical Planning and Inference*. Vol. **142**, pp. 2574-2585.

[3] Croux C., Haesbroeck G. (2003). Implementing the Bianco and Yohai Estimator for Logistic Regression. *Computational Statistics & Data Analysis*. Vol. **44**, pp. 273-295.

[4] Dolgovykh E.M., Lisitsin D.V. (2015). Robust Estimation of Multivariate Regression Model in the Presence of Missing Data. *Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Nonparametric Approach"*. Novosibirsk, pp. 64-71.

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

[6] Eguchi S., Kano Y. (2001). Robustifing Maximum Likelihood Estimation by Psi-divergence. *Research Memorandam of the Institute of Statistical Mathematics*. No. **802**.

[7] Farhadi A., Charalambous C.D. (2008). Robust Coding for a Class of Sources: Applications in Control and Reliable Communication over Limited Capacity Channels. *Systems & Control Letters*. Vol. **57**, pp. 1005-1012.

[8] Flores E., Garrido J. (2001). Robust Logistic Regression for Insurance Risk Classification. *Universidad Carlos III de Madrid Working Papers, Business Economics Series 13*. Working Paper **01-64**.

[9] Fujisawa H., Eguchi S. (2008). Robust Parameter Estimation with a Small Bias against Heavy Contamination. *Journal of Multivariate Analysis*. Vol. **99**, pp. 2053-2081.

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

[11] Jones M.C., Hjort N.L., Harris I.R., Basu A. (2001). A Comparison of Related Density-based Minimum Divergence Estimators. *Biometrika*. Vol. **88**, pp. 865-873.

[12] Kalina J. (2013). Highly Robust Methods in Data Mining. *Serbian Journal of Management*. Vol. **8**, pp. 9-24.

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

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

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

[16] Lisitsin D.V., Gavrilov K.V. (2016). Estimation of Distribution Parameters of a Bounded Random Variable Robust to Bound Disturbance. *Science Bulletin of the NSTU*. Vol. **63**, No. 2, pp. 70-89 (in Russian).

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

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

[19] Tabatabai M. A., et al. (2014). Robust Logistic and Probit Methods for Binary and Multinomial Regression. *Journal of Biometrics and Biostatistics*. Vol. **5**, issue 202.

[20] Wang X. (2014). Modified Generalized Method of Moments for a Robust Estimation of Polytomous Logistic Model. *PeerJ*. Vol. **2**, issue e467.

[21] Windham M.P. (1995). Robustifying Model Fitting. *Journal of Royal Statistical Society. Series B (Methodological)*. Vol. **57**, pp. 599-609.

Neuroevolutionary Forecasting of Innovative Development of the Region with Ecological Regime

Bilgaeva L.[1], Sadykova E.[2], Ochirova G.[2], Zhigdorzhiev V.[1]
[1] *East-Siberian State University of Technology and Management, Ulan-Ude, Russia*
[2] *Baikal Institute of Nature Management SB RAS, Ulan-Ude, Russia*
e-mail: bilgaeval@mail.ru, sad_er@mail.ru, ochgal37@yandex.ru, zhigvb@gmail.com

Abstract

This paper is devoted to the problem solution of neuroevolutionary forecasting of the innovative development of the region with the ecological aspect for example of the Buryatia Republic. We suggest using high-performance computing to increase the convergence rate of the genetic algorithm used to train the neural network.

Keywords: Neuroevolutionary forecasting, genetic algorithm, bipolar sigmoid function, selection, crossover, mutation, high-performance computing, OpenCL technology.

Introduction

The Buryatia Republic occupies an exceptional place in Russia due to Lake Baikal and its unique natural features. A prerequisite for successful economic development of the republic is the development and implementation of technological innovations aimed not only at increasing the efficiency of production activities, but primarily to reduce the anthropogenic load on the natural environment. In this paper, we propose the forecasting of the main parameters of the innovative development of the region using the neuroevolutionary approach. Based on the developed software, a medium-term forecast of the innovative development of the Buryatia Republic was constructed.

1 Genetic algorithm of neural network training

In this problem multilayer perceptron is using as the neural network's structure which includes multiple hidden layers besides input and output layers. Perceptron neurons are sigmoid neurons, the activation function of which is continuous and is expressed as a unipolar or bipolar sigmoid function. The unipolar sigmoid function is represented by a logistic function

$$f(x) = \frac{1}{1 + e^{-\beta x}}, \tag{1}$$

and the bipolar is represented by a hyperbolic tangent

$$f(x) = tanh(\beta x). \tag{2}$$

In this paper, the training of a neural network is proposed to be performed using a genetic algorithm. The choice of this training method is caused by the extensive use of the thus obtained neural network in a time series forecasting problem [5].

The work of the genetic algorithm begins with the definition of input and output data. Input data is a table of factors where the columns represent the set $TS = \{TS_1, ..., TS_c\}$, where c is the number of factors, and the output is an indicator that depends on the input factors. Each column of factors is a time series $TS_i = \{x_{i1}, x_{i2}, ..., x_{ik}\}$, where $i = 1, ..., c$, k is the column size. Forecasting uses sliding window of size $k/2$ and its successive shifts by 1. For the output, which depends on many factors TS_i, a sample $DS = \{s_1, ..., s_k\}$ is constructed. Here $s_j = \{x_{j1}, ..., x_{jk}, y_j\}$ is the sample string, and k is the sample size.

For each time series, an initial population of neural networks is constructed, which is the set $P = \{N_1, ..., N_i, ..., N_n\}$, where n is the size of the population. And the neural network N_i has a set of weights $W = \{w_1, ..., w_t\}$, here t is the total number of weights for all neurons in N_i. Networks differ from each other only by the values of the weights between the neurons. Their structure is the same and is set by the user. Weights are ordered in ascending order of the neuron number to which they belong. Thus, the weights of the neural network represent genes that together form the chromosome as a string of real numbers.

After the initial population is created, a cycle of epochs begins with predetermined stop conditions. During one epoch the following operations are performed: calculation of fitness functions for each individual, selection, crossover and mutation [3]. Let's consider each of these operations.

Calculation of fitness functions for each individual. The input of each neural network is supplied with the first $k/2$ values selected from the time series TS_i. The hidden layer takes this vector to compute an intermediate vector that will be fed to the next layer. Each neuron of a given layer receives the current vector X with the values x_i and with the corresponding weights w_i from the vector W calculates the value of the adder function by the formula

$$S(x) = \sum_{i=1}^{l} (w_i x_i) + w_0, \tag{3}$$

where l - number of neuron's weights, w_0 - bias value.

The resulting value of the adder is fed into the activation function $F(x)$, described by formulas (1) or (2). This function is common for neurons of one layer. Thus, an intermediate vector is formed for the next layer.

For the output layer, the value of the objective function is determined

$$E(x) = \sum_{i=1}^{k} (x_i - x_i')^2, \tag{4}$$

which is the mean square error of the output value of the neural network N_i for the sample DS, according to which the fitness function $F_{N_i} = -E_{N_i}$ is determined.

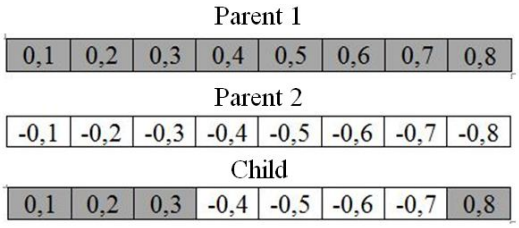


Figure 1: Example of two-point crossover

Selection. For the crossover operation, the elite selection method is used to select the pairs of parents. In this method, a group of individuals P_e is identified, showing the best results. The elitism coefficient E shows the percentage ratio of elite individuals in the population. The size of the elite group is $e = E * n$. Elite individuals are not subject to accidental mutation and replacement by descendants as a result of crossover. This approach allows preserving the best individual in the population until a better one is found. This increases the convergence rate of the genetic algorithm. On the other hand, the probability of convergence in the local minimum is high, since often the same elite individuals will take part in crossover.

This selection method is implemented by sorting in ascending order of the mean square error (or in descending order of the fitness functions of the individual). In this case, the first element indicates the neural network giving the smallest error. A pair of parents is selected as follows: for each non-elite member of the population, a partner from the elite is chosen.

Crossover. Creation of the descendant occurs by two-point crossover of the parents genes. Randomly chosen crossover points c_1 and c_2 determine a child chromosome, t – the number of weights. Genes of non-elite parent are replaced with genes of elite one, and with equal probability, remain either weights $\{w_1, ..., w_{c_1}, w_{c_2+1}, ..., w_t\}$ or $\{w_{c_1+1}, ..., w_{c_2}\}$. Thus, the size of the population remains unchanged. Figure 1 shows an example of two-point crossover. This method shows a sufficiently high convergence in comparison with single-point and multi-point methods of crossover.
Mutation. The number of genes that will be changed is calculated using the formula $m = M * t$, where the mutation ratio M show the percentage ratio of mutating weights among all weights of the neural network. Each of the m genes is randomly selected and a new value w_i is randomly set. Unlike other mutation methods, such as swapping and inverting, this approach introduces more randomness, and therefore reduces the probability of convergence to a local minimum.

The stopping conditions are: to achieve an acceptable value of the error of the output of the neural network or to achieve stagnation, i.e. period of time during which the fitness function of the best individual reached its highest value and does not change over a given number of epochs. This completes the process of training a neural network using a genetic algorithm.

After training, the forecast of the data is calculated for the next few steps, using the best individual (neural network) from the given population. For the time series

TS_i representing the factors, the following K values are calculated. To forecast the sample data DS in it are added K strings composed of the predicted time series TS_i values. Then the sample is fed to the input of the neural network and new values of the indicator are calculated.

1.1 High-performance computing

A specific disadvantage of the genetic algorithm compared with other methods of training the neural network is its low convergence rate. To accelerate the work of the program, OpenCL technology was used, developed and supported by the non-profit consortium Khronos Group. OpenCL includes a set of tools that allow you to use the processing power of the central (CPU) and graphics (GPU) processors for laborious calculations with the possibility of parallelizing the algorithm [1]. To run on a specific device, you need the OpenCL driver installed for it.

To improve performance, has been selected step of calculating the fitness function of each individual genetic algorithm, characterized by high cost in time, unlike other stages. To determine the value of the error of each neural network, you must skip the input data through all the hidden layers, compare the output value with the expected value and repeat this for the entire sample. This gives the average sequential execution time on one processor T_1, which is calculated by the formula: $T_1 = h_{max}^2 h_{count} kn$, where h_{max} is the maximum size of the hidden layer (i.e. the number of neurons in the layer), h_{count} is the number of layers, k is the sample size, and n is the number of neural networks.

The acceleration is based on the parallel execution of calculations for each neural network, since this achieves an efficient distribution of the operations in the streams blocks that do not have information dependencies with each other. Thus, the new execution time (after parallelizing the training process among a certain number of processors p) is $T_1 = h_{max}^2 h_{count} k\frac{n}{p}$, where p is the number of processor units, and the acceleration is $S_p = \frac{T_1}{T_p} = p$, which means high quality parallelization algorithm. The stage begins with the creation of a group of streams equal in size to the number of individuals in the population. Each stream has its own id number and performs calculations only for the neural network N_{id}. Given the output value of out_{id}, the stream determines the err_{id} error and writes it to the string $err = \{err_1, ..., err_n\}$. The step is terminated by returning the string err as the result of the execution.

2 Development of algorithms and software

To solve the problem of forecasting the innovative development of the region with an ecological regime using the approach of implementing a neural network proposed in this work, a program was developed and implemented, the architecture of which is presented in Figure 2. The program consists of four modules: data preparation, genetic algorithm, neural network, visualization of results. Preparing data involves reading the source data from a database or spreadsheet, represented as a floating-point number table, which are displayed on the Source Data tab in the visualization

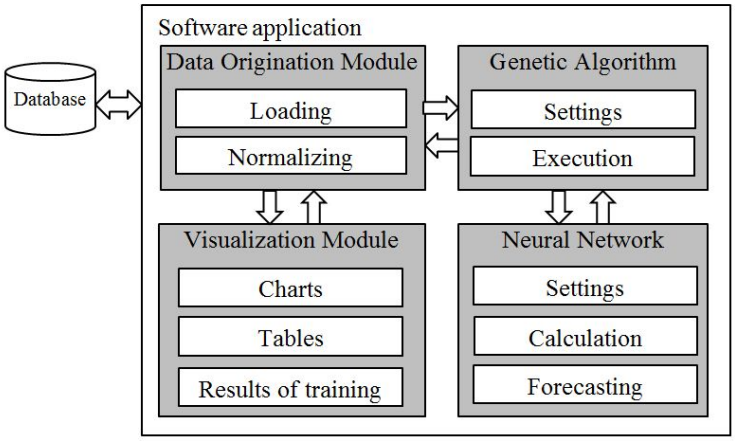


Figure 2: The architecture of the program

module. Here, a selection of factors and an indicator is made for prediction. Next, the data are normalized using the function

$$x' = \frac{x-min}{max-min}(max' - min') + min',$$

performing linear transformation x values in the interval (min, max) in the new interval (min', max').

In the module of the genetic algorithm, the training process takes place. It sets the parameters for stopping the training phase. These may be the threshold of error or the number of epochs in stagnation.

In the neural network module, parameters such as the number of hidden layers, the number of neurons applied to the hidden layers, the type of sigmoid function used for the hidden layers and the output layer, and the number of prediction steps are set.

A special feature of the program is the periodic exchange of data between the modules of the genetic algorithm and the neural network in accordance with the training algorithm. After the training process, the data is forecast and denormalized, then the results are visualized. Ultimately, at the request of the user, it is possible to store the prediction results in a database or a spreadsheet.

3 Computational experiments and analysis of results

During the development of the algorithm for the operation of the neural network and the implementation of software, studies were carried out to select the activation function and increase the speed of the training process of the neural network through the use of high-performance computing.

In the study of the accuracy of the neural network model from the activation function used, four experiments were carried out. It was suggested that the activation function be applied not only to hidden layers, but also to the output layer of the

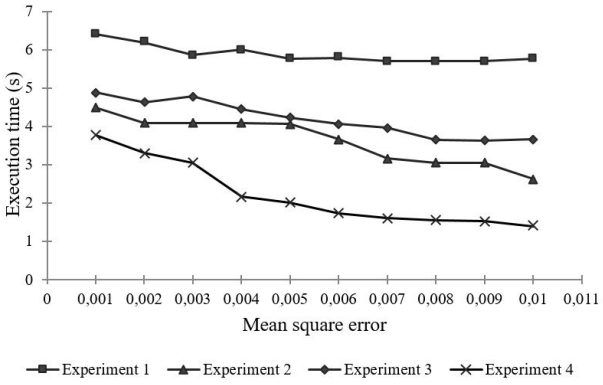


Figure 3: Graph of the time dependence of the algorithm from the given error

perceptron. In the first experiment, a combination of unipolar activation functions on hidden layers and the output layer was used. In the second experiment, a combination of a unipolar activation function on hidden layers and a bipolar activation function on the output layer was used. In the third experiment, a combination of a bipolar activation function on hidden layers and a unipolar activation function on the output layer was used. In the fourth experiment, a combination of bipolar activation functions on the hidden layers and the output layer was used (Fig. 3). The graph shows that the results of the fourth experiment show the best time to reach errors in the range from 0.001 to 0.01. Therefore, in the neural network model, a bipolar activation function was used for the hidden layers and the output layer.

In the course of the experiment, the questions of accelerating the training process of the neural network by applying high-performance OpenCL calculations were investigated. Fig. 4 shows the graph of the time dependence of the genetic algorithm on the mean square error.

The graph shows that parallelization provides a three-fold acceleration algorithm on a central processor with 4 cores. However, the computation on the GPU does not give an advantage over the sequential execution on the CPU. This is due to the frequent operations of copying a large amount of data between the operational and video memory. Next, the results of the medium-term forecast of the innovative development of the Republic of Buryatia, performed in a software environment developed using the neuroevolutionary approach, are presented. The best results were obtained with the following initial parameters of the genetic algorithm and the neural network: number of epochs at which stagnation occurs - 1000; training error for input and output samples - 5%; number of hidden layers - 2; activation function - bipolar sigmoid function; forecast - 5 years.

As a result, the share of innovative goods, works, services in the total volume of goods shipped, works performed and services was adopted. The factor parameters are presented below [2, 4, 6].

Result indicator: Y - the share of innovative goods, works, services in the total volume of shipped goods, works and services performed.

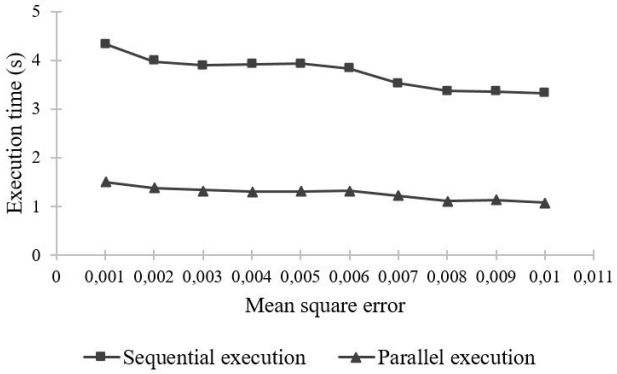


Figure 4: Graph of the time dependence of the algorithm from the given error

Factor indicators: X_1 - internal costs for research and development, as % of GRP; X_2 - the number of patent applications in relation to the number of economically active population; X_3 - specific weight of organizations implementing technological innovations, %; X_4 - number of advanced technologies per 1,000 enterprises, units; X_5 - specific weight of expenses for technological innovations in the total volume of goods shipped, works performed, services, %; X_6 - the norm of environmental investment in fixed assets (the ratio of investment in fixed assets aimed at protecting the environment and rational use of natural resources to GRP), %.

The graph shows (Fig. 5) that the production of innovative goods was at a very low level a fairly long period of time from 2001 to 2010. It is caused by weak innovative activity of organizations, low level of financing of research and development, lack of legal base and state support of innovative organizations. The rise in production of innovative products fell for the period 2011-2014, which is characterized by an increase in the number of innovative-active organizations, an increase in the cost of technological innovation. According to the forecast calculations, the share of innovative goods in the total volume of shipped goods, works, services in 2020 will be 4.8%. Thus, the obtained results of the forecast showed that in the Republic of Buryatia in the period 2016-2018 there will be an increase in the share of innovative goods. The forecast for 2019-2020 testifies that the unstable dynamics observed for a sufficiently long period in the past, due by the state's unequal support in the innovation and investment spheres, will have an impact on the future period of time.

Conclusions

During the development of software to solve the forecasting problem, studies were conducted to determine the parameters of the neural network and the genetic algorithm, which give the most accurate forecasting.

It is known that the use of the evolutionary approach entails a significant increase in the training time, therefore, the application of high-performance computing was proposed. This approach allows you to parallelize the training process and ensure a

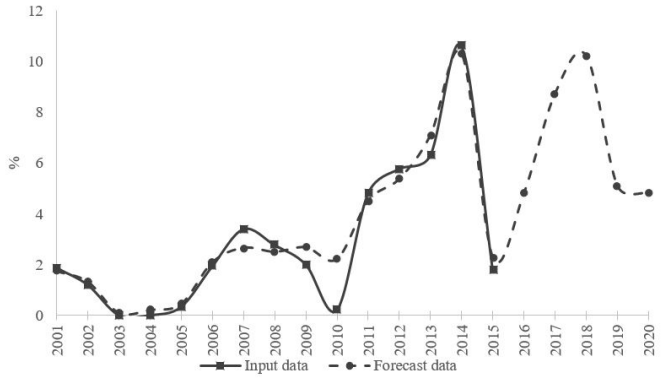


Figure 5: The share of innovative goods, works, services in the total volume of shipped goods, works and services performed, %

three-fold reduction in time.

The results of the forecasting made it possible to establish parameters for the long-term innovative development of the Buryatia Republic. They can be used in executive decision-making on the choice of instruments for regulating the economic activity of a territory with environmental constraints in order to achieve sustainable development.

References

[1] Aaftab Munshi, Benedict R.Gaster, Timothy G.Mattson, James Fung, Dan Ginsburg (2011). *OpenCL Programming Guide*. Addison-Wesley Professional.

[2] *Environmental protection in the Republic of Buryatia: Statistical collection* (2002-2015). Buryatstat, Ulan-Ude.

[3] Goldberg D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional.

[4] *Innovative activity of organizations: Statistical collection* (2010-2015). Buryatstat, Ulan-Ude.

[5] Osovskiy S. (2016). *Neural networks for information processing*. Hotline-Telecom, Moscow.

[6] *Statistical Yearbook: Statistical collection* (2002-2016). Buryatstat, Ulan-Ude.

Study of Adaptive Nonparametric Estimators of the Location Parameter

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

Abstract

Results of study of adaptive robust semi-parametric and nonparametric estimators of the location parameter based on the weighted maximum likelihood method are presented. It is demonstrated that the potential efficiency of the adaptive estimators on local and global supermodels significantly exceeds that of the classical robust estimators.

Keywords: location parameter, robust estimates, nonparametric, weighted maximum liklihood.

Introduction

By the present time a large number of robust estimators synthesized on the basis of various robust criteria and supermodels [1]-[5] have been introduced and used. Unfortunately, investigations have shown that the robust estimators (optimal on classes [3]) can have catastrophically low efficiency for separate distributions from the supermodel. It appears that the robust and efficiency criteria are inconsistent. This is the cause of rightful concern of users of robust statistics methods when choosing an estimation procedure. As a way out from the given situation, adaptive estimators have been suggested in [5] - [12] that are mostly reduced to various procedures of sample truncation. In [9] - [12] the method of construction of robust estimators based on the weighted maximum likelihood method (WMLM) using nonparametric adaptation algorithms was suggested for semi-parametric and nonparametric supermodels.

In the present work, results of investigations of the potential efficiency of the adaptive WMLM estimators of location parameter for the global and local Tukey supermodels and their intersections are considered. New results on the potential efficiency of adaptive estimators, and primarily of the adaptive WMLM nonparametric estimators, on semi-parametric and nonparametric supermodels are presented.

1 Adaptive WMLM estimators

Let there be a random variable X with the distribution function $F(x,\theta)$ and the probability density $f(x,\theta)$ and the aprioristic (ideal) distribution function $G(x,\theta)$ with the probability density $g(x,\theta)$. It is required to estimate the unknown parameter θ from the sample of independent and equally distributed (i.i.d.) random variables $X_N = (x_1, ..., x_N)$ with the distribution function $F(x,\theta)$.

In [9] - [12], the method of constructing robust estimators by the weighted maximum likelihood method (WMLM) was suggested based on the A. M. Shurygin stability criterion [4]. The class of the Shurygin stable estimators [4] was synthesized using the criterion based on variation of the M-estimator variance on the density that obeys the main principle of robust statistics "small deviations in the aprioristic distribution should not lead to large deviations in the characteristics of estimators" [1], [2]. The method of synthesis of robust estimators by the WMLM is based on equations with the estimating function $\phi(x,\theta)$ of the form [9] - [12].

$$\int \phi(x,\theta_n)dF_n(x) = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i,\theta_N) = 0, \tag{1}$$

$$\phi(x,\theta) = \left(\frac{\partial}{\partial\theta}g(x,\theta)+\beta_\theta\right)g^l(x,\theta). \tag{2}$$

where $F_N(x)$ is the empirical distribution function, $0 \le l \le 1$ is the radicalness parameter of the estimator, and β is the parameter determined from the condition of the unbiased estimator

$$E_g(\phi(x,\theta)) = \int \phi(x,\theta)dG(x,\theta). \tag{3}$$

For the location parameter, $\beta_\theta = 0$ [11]. We note that at $l = 0$ we have the maximum likelihood estimator (MLE), at $l = 0.5$ we have the maximum radicalness estimator (heuristic MD, that is, the minimum Hellinger distance estimator), and at $l = 1$ we have the Shurygin maximum stability estimator (SMSE) [4].

The theorem [12] that the random variable $\sqrt{N}(\theta_N - \theta)$ obeys the asymptotically normal distribution with the average

$$b = -\frac{\int\left(\frac{\partial}{\partial\theta}ln\ g(x,\theta)+\beta_\theta\right)g^l(x,\theta)dF(x,\theta)}{\int g^l(x,\theta)\left(\frac{\partial^2}{\partial\theta^2}ln\ g(x,\theta)+l\ \frac{\partial}{\theta\partial}ln\ g(x,\theta)\left(\frac{\partial}{\partial\theta}ln\ g(x,\theta)+\beta_\theta\right)\right)dF(x,\theta)} \tag{4}$$

and the variance

$$V = \frac{\int\left(\frac{\partial}{\partial\theta}g(x,\theta)+\beta_\theta\right)^2 g^{2l}(x,\theta)dF(x,\theta)}{\left(\int g^l(x,\theta)\left(\frac{\partial^2}{\partial\theta^2}ln\ g(x,\theta)+l\ \frac{\partial}{\theta\partial}ln\ g(x,\theta)\left(\frac{\partial}{\partial\theta}ln\ g(x,\theta)+\beta_\theta\right)\right)dF(x,\theta)\right)^2} \tag{5}$$

can be proved. Physically, the role of the radicalness parameter is reduced to "soft" truncation of remote outliers and aprioristic distribution functions $G(x,\theta)$ of the estimator. The radicalness parameter l for local deviations of the distribution $G(x,\theta)$ from $F(x,\theta)$ is used to develop local adaptation algorithms of estimators on the class of stable estimators [9] - [12]. In semi-parametric and nonparametric problems, local deviations of the distribution $G(x,\theta)$ from $F(x,\theta)$ are unknown; therefore, to adjust the parameter l, nonparametric adaptation algorithms are required for estimators. The variance $V(\theta_N, l)$ of the WMLM estimator given by (5) depends on the parameter l, and the optimal l^* value is determined from the condition $l^* = \underset{l}{\operatorname{argmin}}\, V(\theta_N, l)$. In

semi-parametric and nonparametric problems, $V(\theta_N, l)$ is unknown. We find the nonparametric estimator $V_N(\theta_N, l)$ of the variance $V(\theta_N, l)$ using bootstrap procedures [13] and determine the optimal l^* value from the condition $l^* = \underset{l}{\operatorname{argmin}}\, V_N(\theta_N, l)$ – the local nonparametric adaptation algorithm. The bootstrap procedures [13] require large volume of calculations on a computer, which is not critical nowadays. Let us consider a class of nonparametric problems of robust statistics. In this case, the form of the aprioristic distribution $G(x, \theta)$ is unknown, except general information on its continuity and symmetry (global changes) on which local deviations $G(x, \theta)$ from $F(x, \theta)$ are superimposed. The case in point are the semi-nonparametric problems in which the form of $G(x, \theta)$ is unknown, but some aprioristic information on $G(x, \theta)$ is required, for example, on the symmetry of $G(x, \theta)$, which allows one to distinguish $G(x, \theta)$ from $F(x, \theta)$. On the basis of the aprioristic information on $G(x, \theta)$, we construct the nonparametric estimator of the Rosenblatt–Parzen (RP) type $g_N(x, \theta)$ for the distribution density $g(x, \theta)$. Substituting $g_N(x, \theta)$ into (2) instead of $g(x, \theta)$, we obtain the global adaptation algorithm for WMLM estimators.

Let $G(x, \theta)$ be the distribution symmetric about θ, then $g_N(x, \theta)$ is written as [12]

$$g_N(x, \theta, h_N) = \frac{1}{2N\ h_N} \sum_{j=1}^{N} \left(K\left(\frac{x - x_j}{h_n}\right) + K\left(\frac{2\theta - x - x_j}{h_N}\right)\right) \tag{6}$$

where $K(x)$ is the kernel function and h_N is the spread parameter.

Substituting (6) into (2) and (1), instead of $g(x, \theta)$ we obtain the nonparametric estimator θ_N of the WMLM location parameter θ of the symmetric distribution determined by the estimating equation of the form [11]

$$\sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} K'(z_{i,j})\ g_N^{l-1}(x_i, \theta_N, h_N) = 0, \tag{7}$$

where $z_{i,j} = \frac{2\theta_N - x_i - x_j}{h_N}$, $K'(x) = \frac{d}{dx} K(x)$.

2 Formulation of the problem on investigation of the WMLM estimators

We now consider the problem on estimation of the location parameter for a distribution. The given formulation simplifies the problem, but using transformations, many parameters can be reduced to the location parameter of the distribution, which allows us to generalize the results of investigations.

Since the efficiency of estimators depends on the degree of stretching of distribution tails [1]-[5], we here consider the global supermodel S_2 of symmetric distributions with light, intermediate, and strongly stretched tails (D4 < N < L < C):

1. Generalized normal distribution of the fourth degree (D4)

$$g_1(x, \mu, s) = \frac{2}{s\Gamma(0.25)} e^{-\left(\frac{x-\mu}{s}\right)^4}, \tag{8}$$

where $\Gamma(x)$ is the gamma function.

2. Normal distribution (N)

$$g_2(x,\mu,s) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{s}\right)^2} \tag{9}$$

3. Laplace distribution (L)

$$g_3(x,\mu,s) = \frac{1}{2s} e^{-\frac{|x-\mu|}{s}} \tag{10}$$

4. Cauchy distribution (C):

$$g_4(x,\mu,s) = \frac{1}{\pi\left(1+\left(\frac{x-\mu}{s}\right)^2\right)} \tag{11}$$

We define the scale parameters of distributions on the global supermodel so that 0.95 quantiles of distributions coincided with 0.95 quantiles of the standard normal distribution (Table 8).

Table 1: Values of the scale parameters of the distributions

Distribution	D4	N	L	C
Scale parameter s	1.767	1.000	0.7144	0.2605

When choosing the parameters s_i according to Table 1, the distributions differ first of all only by the curve types.

As a local supermodel, we take the Tukey model S_1 with the share of outliers $\epsilon = 0.1$, symmetric outliers (SO) of the form

$$f_i(x) = 0.9g_i(x,0,s_i) + 0.1g_i(x,0,3s_i), i = 1..4 \tag{12}$$

and asymmetric outliers (AO) of the form

$$f_i(x) = 0.9g_i(x,0,s_i) + 0.1g_i(x,7,s_i), i = 1..4 \tag{13}$$

Let us consider the supermodel $S = S_1 \cup S_2$ representing unification of the local S_1 and global S_2 supermodels consisting of 8 distributions (8) - (13).

We will characterize the quality of estimator of the location parameter $\hat{\mu}$ by the potential asymptotic relative efficiency

$$e^*(\hat{\mu},\hat{\mu}_{ML}) = \frac{MSE(\hat{\mu}_{ML})}{MSE(\hat{\mu})} \tag{14}$$

where $\hat{\mu}_{ML}$ is the maximum likelihood estimator with parametrical level of aprioristic information on the form of the distribution function $F(x)$, $MSE = V + b^2$ is the asymptotic mean square error.

For our investigation, we take the following estimators of the location parameter. For each distribution from S, the effective maximum likelihood estimator μ_{ML} and the sample median sm (robust estimator on the nonparametric supermodel of the symmetric distributions) were found. For each distribution from S_2, the maximum likelihood estimator mle ($\hat{\mu}_{ML}$ on S_2), the maximum stability estimator mse, the maximum radicalness estimator mre (heuristic MD, namely, the Hellinger distance estimator), the adaptive semi-parametric estimator ae^* (1) and (2) for distribution * from S_2, and the adaptive nonparametric estimator (7) aen were found.

The estimating equations for distributions from S_2, the bias b, and the variance V of the WMLM estimators were calculated based on (1)–(7) for distributions from S. Omitting all numerous mathematical calculations for finding MSE of the estimators, we present only the final results of investigations.

3 Results

Investigation of estimators was performed on S_1(**AO** and **SO** for $\epsilon = 0$ – without outliers).

Table 2: $e^*(\hat{\mu}, \hat{\mu}_{ML})$ of estimators on **D4**

Estimator	MLE	re	mre	sm	aeD4	aeN	aeL	aeC	ane
$\epsilon = 0$	**1.000**	0.813	0.604	0.300	**1.000**	0.730	0.300	0.254	0.866
AO	0.016	0.814	0.604	0.267	**0.981**	0.593	0.267	0.249	0.860
SO	0.061	0.914	0.706	0.364	**0.956**	0.702	0.364	0.309	0.708

Table 3: $e^*(\hat{\mu}, \hat{\mu}_{ML})$ of estimators on **N**

Estimator	MLE	re	mre	sm	aeD4	aeN	aeL	aeC	ane
$\epsilon = 0$	**1.000**	0.838	0.650	0.637	0.833	**1.000**	0.637	0.599	0.842
AO	0.174	0.839	0.650	0.568	0.834	**0.931**	0.568	0.591	0.893
SO	0.698	0.897	0.705	0.656	0.836	**0.987**	0.696	0.660	0.758

Of considerable interest is investigation of $e^*(\hat{\mu}, \hat{\mu}_{ML})$ (14) of the adaptive nonparametric estimator (ane) on S. For this purpose, the WMLM estimators on the global supermodel of distributions S_2= {**D4**, **N**, **L**, **C**} were investigated without ($\epsilon = 0$) and with asymmetric (**AO**) and symmetric (**SO**) outliers. Here ane was compared with maximum stability estimators (mse*).

Table 4: $e^*(\hat{\mu}, \hat{\mu}_{ML})$ of estimators on **L**

Estimator	MLE	re	mre	sm	aeD4	aeN	aeL	aeC	ane
$\epsilon = 0$	**1.000**	0.889	0.750	**1.000**	0.401	0.755	**1.000**	0.868	0.744
AO	0.892	0.889	0.751	0.892	0.401	0.756	**0.968**	0.859	0.707
SO	0.944	0.859	0.724	0.940	0.381	0.712	**0.950**	0.834	0.705

Table 5: $e^*(\hat{\mu}, \hat{\mu}_{ML})$ of estimators on **C**

Estimator	MLE	re	mre	sm	aeD4	aeN	aeL	aeC	ane
$\epsilon = 0$	**1.000**	0.851	0.619	0.814	0.603	0.943	0.855	**1.000**	0.897
AO	**0.993**	0.853	0.674	0.729	0.602	0.944	0.848	**0.993**	0.798
SO	**1.000**	0.854	0.619	0.807	0.594	0.935	0.860	**1.000**	0.734

Using the Monte Carlo method, for each distribution from S_2 with **AO** and **SO** their MSE_N and $e^*(\hat{\mu}, \hat{\mu}_{ML})$ values of the effective estimators for the parametrically assigned $F(x)$ were found for each estimator μ_N using the bootstrap procedure. On the distributions of the supermodel S, the average efficiency $mean(e^*) = \frac{1}{4}\sum_{i=1}^{4} e^*(\hat{\mu}, \hat{\mu}_{ML}, F_i)$ and the minimum efficiency $min(e^*) = min(e^*(\hat{\mu}, \hat{\mu}_{ML}, F_i))$ for each estimator were calculated (Table 6).

Table 6: Average and minimal efficiencies of estimators on the global supermodels

Estimator	$\epsilon = 0$		AO		SO	
	mean(e*)	min(e*)	mean(e*)	min(e*)	mean(e*)	min(e*)
sm	0.694	0.334	0.198	0.091	0.710	0.389
mreD4	0.590	0.463	0.549	0.437	0.608	0.434
mreN	0.685	0.391	0.664	0.321	0.725	0.446
mreL	0.528	0.174	0.326	0.092	0.513	0.184
mreC	0.541	0.172	0.515	0.139	0.557	0.215
ane	**0.837**	**0.743**	**0.814**	**0.707**	**0.766**	**0.705**

Conclusions

1. Classical robust estimators (mreD4, mreN, mreL, and mreC) lose up to 40% of their efficiency on local supermodels (Tables 2–5).

2. On local supermodels, the adaptive WMLM estimators (aeD4, aeN, aeL, and aeC) are leading with the efficiency close or equal to 1 (Tables 2–5).

3. On global supermodels, the classical robust estimators (mreD4, mreN, mreL, and mreC) catastrophically lose their efficiency ($min(e^*) = 0.09$, Table 6).

4. On global supermodels, the adaptive nonparametric WMLM estimator (ane) is leading with loss in the potential efficiency of (20%–30%) – the payment for high statistical uncertainty of the problem (Table 6).

References

[1] Huber P.J. (1984). *Robustness in statistics*. Mir, Moscow.

[2] Hampel F. R., Ronchetti E. M., Rausseu P. J., Stahel W. A. (1989). *Robust statistics*. Mir, Moscow.

[3] Tsypkin Ya. Z. (1995). *Principles of information theory of identification*. Nauka, Moscow.

[4] Shurygin A. M. (2000). *Applied statistics. Robustness. Estimator. Prediction*. Financy i Statistika, Moscow.

[5] Shulenin V. P. (2016). *Robust methods of mathematical statistics*. Publishing House of Scientific and Technology Literature, Tomsk.

[6] Stone C. J. (1976). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist*. Vol. **3**, pp. 267-284.

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

[8] Hogg R. V. (1982). On adaptive statistical inferences. *Communications in Statistics – Theory and Methods*. Vol. **11**, pp. 2531-2542.

[9] Simakhin V.A. (2004). *Nonparametric statistics. Theory of estimators. Part 2*. Publishing House of Kurgan State University, Kurgan.

[10] Simakhin V.A., Rymar I.V. (2006). Nonparametric robust estimates of the location and scale parameters. *Proc. SPIE*. Vol. **6160**, pp. 230-239.

[11] Simakhin V.A. (2004). *Robust nonparametric estimators*. LAMBERT Academic Publishing, Germany.

[12] Cherepanov O.S. (2015). *Robust parameter estimates based on the weighted maximum likelihood method* (Ph.D. dissertation). Tomsk State University, Tomks, Russian Federation.

[13] Efron B. (1998). *Nonconventional methods of multidimensional statistical analysis*. Financy i Statistika, Moscow.

Nonparametric Algorithms for Recovery Of Mutually Unbeatted Functions on Observations

Anna A. Korneeva, Svetlana S. Chernova and Anastasia V. Shishkina
Siberian Federal University, Krasnoyarsk, Russia
e-mail: korneeva_ikit@mail.ru

Abstract

The problem of reconstructing a function from observations with random errors is considered. Moreover, at the formulation stage of the problem there is no stage associated with the parametric structure of this function. Therefore, the estimate is sought in the class of nonparametric statistics, when the original description of the function is unknown up to a parameter vector. The peculiarity of this problem is that the desired function is described by a mutually ambiguous characteristic and the generally accepted nonparametric estimation proves to be unsuitable. It was necessary to introduce a new class of nonparametric estimators. The results of some computational experiments are presented.

Keywords: A priori information, a nonparametric model, mutually ambiguous characteristics, nonparametric estimates.

Introduction

The problem of reconstructing a function from observations is considered when the process under investigation is described by mutually ambiguous characteristics. This problem reduces to an approximation problem, the main feature of which is the absence of a priori information on the parametric structure of the model of the process under study. We propose a nonparametric estimation of mutually ambiguous characteristics, its some modification and the results of numerical studies.

When regression functions are restored from observations, nonparametric estimates are often used. It is assumed that the nature of its dependence is single-valued in the argument. Below we consider the problem of reconstructing a function from observations in a mutually ambiguous relationship. This required some changes to the well-known estimate of Nadaraya-Watson.

In cybernetic problems, there is often a need to use a priori information. There are the following levels: systems with complete information; Systems with incomplete information; Systems with active accumulation of information; System with parametric uncertainty [1].

In this paper we consider problems with nonparametric information.

1 Nonparametric approach

This approach is based on nonparametric estimates of the probability density $p(x)$ from observations $x_i, i = \overline{1,s}$. Nonparametric estimates of the multidimensional

probability density were considered in detail in [3,4] and have the form:

$$P_s(x) = \frac{1}{s}\sum_{i=1}^{s}\frac{1}{c_s}\prod_{j=1}^{k}\Phi\left(\frac{x^j - x_i^j}{c_s}\right), \tag{1}$$

where $P_s(x)$ – estimation of the distribution density of elements, s – sample size, k – dimension of the vector x, c_s – the blur parameter, which determines the "delta-shape" of the core $\Phi(v)$ [3,4]. Here $\Phi(v)$ – is a compactly bell-shaped square-integrable function satisfying the conditions [2,3,4].

$$\begin{gathered} 0 < \Phi(v) < \infty \forall v \in \cap(v), \frac{1}{c_s}\int \Phi\left(\frac{x - x_i}{c_s}\right) dx = 1, \\ \lim_{s\to\infty}\frac{1}{c_s}\Phi\left(\frac{x - x_i}{c_s}\right) = \delta(x - x_i). \end{gathered} \tag{2}$$

Also c_s satisfies the following conditions:

$$c_s > 0, \lim_{s\to\infty} s(c_s)^k = \infty, \lim_{s\to\infty} c_s = 0. \tag{3}$$

In the computational experiment, the bell-shaped functions $\Phi(v)$ of various types are used, for example, a rectangular, triangular, parabolic kernel.

2 Nonparametric estimation of the regression function from observations

To restore the regression function $M\{y|x\}$ from observations $\{x_i, y_i, i = \overline{1,s}\}$ we use nonparametric estimates of the probability density (1). $M\{y|x\}$ has the form

$$M\{y|x\} = \frac{\int_{\Omega(y)} yp(x,y)dy}{\int_{\Omega(y)} p(x,y)dy}. \tag{4}$$

Replacing in (4) $p(x,y)$ with nonparametric estimates (1) and using the property:

$$\frac{1}{c_s}\int_{\Omega(y)} y\Phi\left(\frac{y - y_i}{c_s}\right) dy = y_i, i = \overline{1,s}. \tag{5}$$

It is easy to obtain a nonparametric estimate of the Nadaraya-Watson regression function, which for the one-dimensional case is as follows:

$$Y_s(x) = \frac{\sum_{i=1}^{s} y_i\Phi\left(\frac{x-x_i}{c_s}\right)}{\sum_{i=1}^{s}\Phi\left(\frac{x-x_i}{c_s}\right)}, \tag{6}$$

and for the case if the k-dimensional vector x is equal to:

$$Y_s(x) = \frac{\sum_{i=1}^{s} y_i \prod_{j=1}^{k} \Phi\left(\frac{x_j - x_j^i}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{k} \Phi\left(\frac{x_j - x_j^i}{c_s}\right)}, \tag{7}$$

where $x_i, y_i, i = \overline{1,s}$ – sample of observations, $\Phi(v)$ – bell-shaped function, v – arbitrary variable, c_s – blur option.

When restoring a mutually ambiguous regression function, the Nadaraya-Watson estimate should be modified as follows:

$$Y_s(x) = \frac{\sum_{i=1}^{s} y_i \Phi\left(\frac{x_t - x_i}{c_s}\right) \Phi\left(\frac{x_{t-1} - x_{i-1}}{c_s}\right) \Phi\left(\frac{y_{t-1} - y_{i-1}}{c_s}\right)}{\sum_{i=1}^{s} \Phi\left(\frac{x_t - x_i}{c_s}\right) \Phi\left(\frac{x_{t-1} - x_{i-1}}{c_s}\right) \Phi\left(\frac{y_{t-1} - y_{i-1}}{c_s}\right)}, \tag{8}$$

where x_{t-1}, y_{t-1} the values of the coordinates of the regression function at the previous step of its estimation [5].

As shown by numerous computational experiments, it is expedient (6) to correct somewhat as follows:

$$Y_s(x) = \frac{\sum_{i=1}^{s} y_i \Phi\left(\frac{x_t - x_i}{c_s}\right) \Phi^0\left(\frac{x_{t-1} - x_{i-1}}{c_s}\right) \Phi^0\left(\frac{y_{t-1} - y_{i-1}}{c_s}\right)}{\sum_{i=1}^{s} \Phi\left(\frac{x_t - x_i}{c_s}\right) \Phi^0\left(\frac{x_{t-1} - x_{i-1}}{c_s}\right) \Phi^0\left(\frac{y_{t-1} - y_{i-1}}{c_s}\right)}, \tag{9}$$

where $\Phi^0(v)$ to within a coefficient, repeats $\Phi(v)$, and $\Phi^0(v) = 1$, if $v < 1$ and 0 in another cases. In this case $\Phi^0(v)$ will not affect the recovery error, but will allow the algorithm to "fix" the previous point of motion when evaluating each subsequent point.

If x is a vector $(x_1 ... x_k) \in R^k$ of dimension k. The training sample in this case has the form: $x_{1i} ... x_{ki}, y_i, i = \overline{1,s}$. When restoring a mutually ambiguous regression function, a nonparametric estimate must be modified as follows:

$$Y_s(x_t) = \frac{\sum_{i=1}^{s} y_i \prod_{j=1}^{k} \Phi\left(\frac{x_t^j - x_i^j}{c_s}\right) \prod_{j=1}^{k} \Phi\left(\frac{x_{t-1}^j - x_{i-1}^j}{c_s}\right) \Phi\left(\frac{y_{t-1} - y_{i-1}}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{k} \Phi\left(\frac{x_t^j - x_i^j}{c_s}\right) \prod_{j=1}^{k} \Phi\left(\frac{x_{t-1}^j - x_{i-1}^j}{c_s}\right) \Phi\left(\frac{y_{t-1} - y_{i-1}}{c_s}\right)}, \tag{10}$$

where x_{t-1}^j, y_{t-1}^j values of the coordinates of the regression function at the previous step of it's estimation.

The nonparametric estimate (11) can be modified as follows:

$$Y_s(x_t) = \frac{\sum_{i=1}^{s} y_i \prod_{j=1}^{k} \Phi\left(\frac{x_t^j - x_i^j}{c_s}\right) \prod_{j=1}^{k} \Phi^0\left(\frac{x_{t-1}^j - x_{i-1}^j}{c_s}\right) \Phi^0\left(\frac{y_{t-1} - y_{i-1}}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{k} \Phi\left(\frac{x_t^j - x_i^j}{c_s}\right) \prod_{j=1}^{k} \Phi^0\left(\frac{x_{t-1}^j - x_{i-1}^j}{c_s}\right) \Phi^0\left(\frac{y_{t-1} - y_{i-1}}{c_s}\right)}, \tag{11}$$

where $\Phi^0(v)$ is the same as above.

3 Computational experiment

When carrying out a computational experiment, the mutually ambiguous characteristics can have different forms: circles, ellipses, and others. Without loss of generality, the mutually ambiguous characteristic of the function $y(x)$ is taken (for reasons of simplicity) in the form of a circle.

$$x^2 + y^2 = r^2, \tag{12}$$

where r – radius of a circle.

In this case, the training sample was formed as follows: the starting point x‘ was arbitrarily set and computed y‘(x) in accordance with (12). As a result, a sample $x_i, y_i, i = \overline{1,s}$ was formed. Let's pay attention to what could be determined as a result of a uniform step Δx or random number generator $x_i \in \Omega(x), i = \overline{1,s}$. In the process of computer research, other mutually ambiguous characteristics of dependence $y(x)$. When reconstructing a mutually ambiguous characteristic from observations that the researcher does not know, it is important to choose the direction of motion, although in principle it can be arbitrary at the initial stage. But all subsequent changes to the current variable x are strictly dependent on the previous one.

Processes characterized by mutually ambiguous dependencies have such a feature that the values $x_t, t = 1, 2...$ appear strictly sequentially in one direction or another. In Fig. 1 shows such a process. Let, for example, in the first step $x = x_1$, then x_2, etc. The value x_t appears only after x_{t-1}, that is, there is a "movement" x_t in an arbitrary chosen direction. The appearance of the values begins at a certain point x_t and moves sequentially, passing the points $t_2, t_3, ...$ In this case, the transition x_1, for example, to x_5 is impossible, until the previous four points have been traversed.

Thus, the essence of the proposed estimates (8,9) is that when evaluating the next point, the "fixing" to the previous point is performed in the corresponding algorithms (8,9).

In Fig. 1 represents a process that is a circle. Motion along the variable x occurs from right to left and from left to right, which characterizes the sequential appearance of sampling values.

In the next step, the random effect of h on the observations of y_i

$$h_i = l y_i \xi, \tag{13}$$

where $\xi \in [-1; 1]$, interference level $l = 0\%, 5\%, 10\%$.

As a criterion for the accuracy of nonparametric estimation, the following relation was used:

$$w = \frac{\sum_{i=1}^{s} |y_i - y_s(x_i)}{\sum_{i=1}^{s} |y_i - \overline{y}|}, \tag{14}$$

where $\overline{y} = \frac{1}{s} \sum_{i=1}^{s} y_i$ – average, $y_s(x_i)$ – nonparametric estimate, y_i – the true sample obtained by formula (12).

We present the results of a numerical study illustrating the effectiveness of the algorithm. As a bell-shaped finite function, a triangular core was used. The algorithm

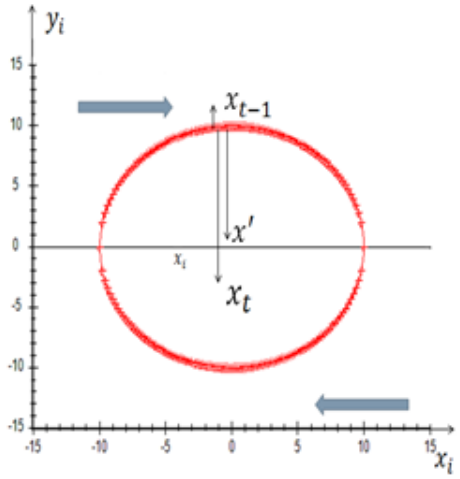


Figure 1: The sample

was tested on training samples of various sizes, with a subsequent increase in the sample size by adding new elements to the existing ones: $s = 50, 100, 500$.

In all figures, let's designate the figure (1) as the training sample, (2) the non-parametric estimate.

The work of the algorithm (8) in Figures 2, 3, 4 is demonstrated in different conditions: when the sample size is 50, 100, 500 elements; the interference level is 0%; the experiment was conducted in the mode of the sliding examination.

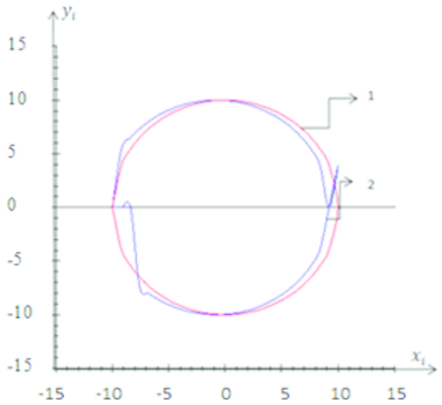


Figure 2: S=50, W=0,1098

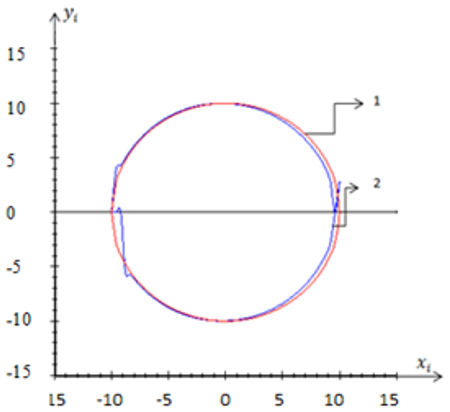


Figure 3: S=100, W=0,0469

Figure 5 shows the dependence of the recovery error on the volume at different levels of interference.

In computational experiments, other mutually ambiguous characteristics were also used. Some fragments of the study are given below. The experiment was carried out under different conditions: the sample size is 200 elements; The interference level is 0%; In the sliding examination mode, see Fig. 6. Comparing Figures 6 and 7 (the

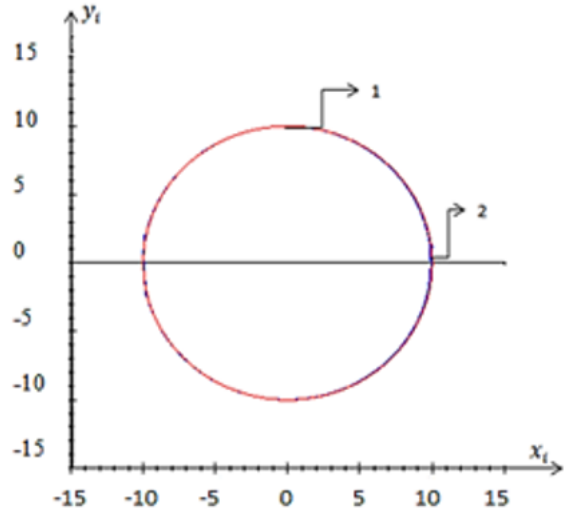


Figure 4: S=500, W=0,0068

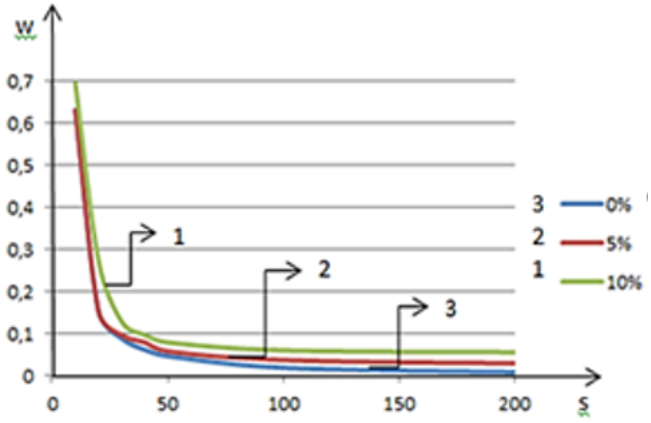


Figure 5: Result of the experiment

level of interference is equal to 10

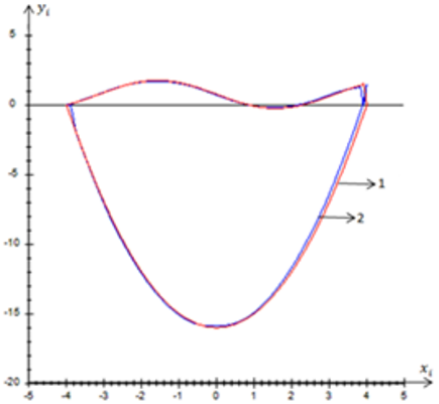


Figure 6: S=200, W=0,018

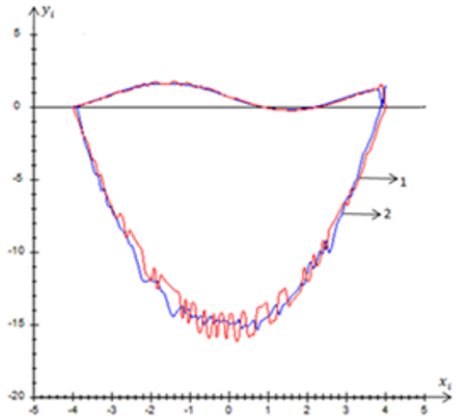


Figure 7: S=200, l=10%, W=0,0011

Let's demonstrate the operation of the modified algorithm (9), under the following conditions: at an interference level of 5% and 10%; With a sample size of 100 elements; In the mode of sliding examination. Comparing the recovery errors, see Figures 8, 9 and see a slight improvement. Hence, the non-parametric estimation became more precise.

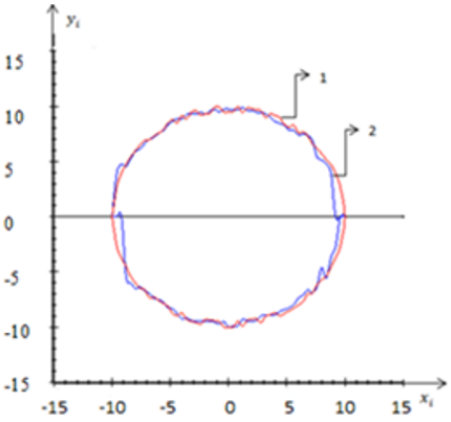


Figure 8: S=100, l=5%, W=0,057

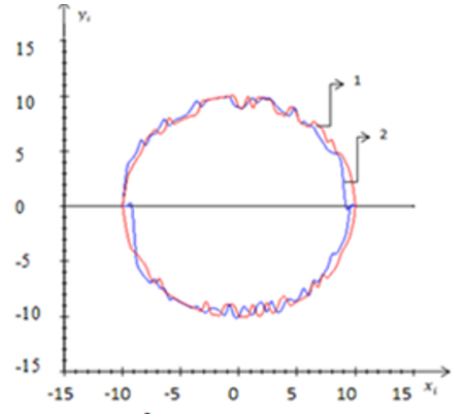


Figure 9: S=100, l=10%, W=0,0817

It should be noted that: with a decrease in the recovery error (w), the accuracy of the estimate increases; With the increase in the sample size (s), the recovery error (w) decreases; The error value increases with increasing interference level (l).

Perhaps, the question arises: "Why was the circle used to check the operation of the algorithm?", Because there are so many more complex figures, and the answer is simple - the feature of this algorithm is its universality. This means that the algorithm

does not fundamentally what function to restore, whether it's a circle, an ellipse, an Archimedes spiral or a curtain of Aniezi. "Fixing" at the preceding point x_t, that is, at the point x_{t-1} and preserving the bypass direction, it is always possible to obtain a nonparametric estimation of mutually ambiguous functions.

Conclusions

The main result of this paper is the introduction of a new class of nonparametric estimation of mutually ambiguous functions from observations with errors. This distinguishes the tasks of nonparametric estimation from known nonparametric estimates of the Nadaraya-Watson regression function. Some modifications of nonparametric estimations are given, under such conditions attention is drawn to the technique of bypassing the introduced nonparametric estimations along the trajectory determined by the elements of the training sample.

For simplicity of numerical investigation, a function described by a circle was taken, although this is not essential for the proposed algorithm. In other words, algorithms are proposed that are suitable for reconstructing ambiguous dependencies described by more complex curves, the nature of which is unknown a priori, is known only: the sample of observations of the process being studied.

References

[1] Feldbaum A.A. (1963). *Fundamentals of the theory of optimal automatic systems.* Fizmatgiz, Moscow.

[2] Medvedev A.V. (2015). *Fundamentals of the theory of adaptive systems.* SibGAU, Krasnoyarsk.

[3] Nadaraya E.A. (1983). *Nonparametric estimation of probability density and regression curve.* Publishing house of the Tbilisi University, Tbilisi.

[4] Vasilyev V.A., Dobrovidov A.V., Koshkin G.M (2004). *Nonparametric estimation of functionals on the distributions of stationary sequences.* Nauka, Moskow.

[5] Zhivoglyadov V.P., Medvedev A.V., Tishina E.V (1973). *Reconstruction of ambiguous static characteristics from experimental data.* Automation of industrial experiment, Frunze.

Some Remarks on the Theory of Non-Parametric Systems

A.V. Medvedev
Siberian Aerospace University, Krasnoyarsk, Russia
e-mail: saor_medvedev@sibsau.ru

Abstract

The report deals with new classes of processes of discrete-continuous type, namely, H-, T-, K-processes. Such processes are widespread in reality. The construction of models of similar processes is associated with certain features that distinguish the latter from those generally accepted in the theory of identification. Some new models for the processes under consideration are presented.
Key words:
Keywords: a priori information, parametric models, nonparametric models.

Introduction

In the presence of small amount of a priori information about the process under investigation, a situation often arises when the parametric structure of the model is unknown up to parameters. This leads to the need for the development of non-parametric methods of identification and control. The situation is further complicated by the fact that some stochastic dependencies may exist between the components of the input and output variables, which, of course, are not known to the researcher, as well as other features often arising in the construction of the model, in particular, the different discreteness of control of the process variables. There is also a certain feature associated with the need for modeling dynamic processes as inertial-free objects with a net delay.

This is due to the fact that some output variables of the object are measured at significantly longer time intervals than the input ones, and these intervals significantly exceed the time constant of the object. For example, a number of variables are measured electrically (in this case, the discreteness of the control Δt can be quite small), and other variables are monitored as a result of chemical analysis or physical and mechanical tests (in this case the discreteness of control ΔT is large, i.e. $\Delta T >> \Delta t$). Then the object under study can be regarded as static with delay. Such a process can be presented in the form:

$$x(t) = f(u(t-\tau), \xi(t)), \tag{1}$$

where $x(t)$ is an output variable of the object, $u(t-\tau)$ is an input variable, τ is delay, $\xi(t)$ is random perturbation acting on the object, (t) is continuous time.

Let's consider some issues related to modeling of static systems. Let $u = (u_1, u_2, ..., u_k) \in \Omega(u) \subset R^k$, $x \in \Omega(x) \subset R^1$, each component of $u_i \in [a_i, b_i], i = 1, 2, ..., k$, and

$x \in [c, d]$. When studying real processes, the values of the coefficients $\{a_i, b_i, c, d\}, i = 1, 2, ..., k$ are always known. From this point on, without loss of generality, we take these intervals to be unit intervals. Then, $\Omega(u)$ is a unit hypercube, $\Omega_k(u) = [0; 1]$, i.e. $u \in [0; 1]$, $\Omega_{k+1}(u, x) = [0; 1]$, $(u, x) \in \Omega_{k+1}$.

The task of identification is often reduced to a parametric one, the solution of which consists of two main stages: the first stage is the selection (definition) of the parametric model in the form:

$$\hat{x} = \hat{f}(u, \alpha), \tag{2}$$

where α is a vector of parameters; the second stage is the evaluation of the parameters α based on the incoming sample elements $(u_1, x_1), (u_2, x_2), ..., (u_s, x_s)$, that is, parameters assessment α_s.

This is the general scheme for solving parametric identification problems. We only note that the weakest point here is the choice of the parametric structure of the model. If at the first stage a blunder is admitted, then the resulting model is unlikely to be satisfactory.

1 H-models

The process under investigation proceeds without loss of generality in the unit cube $\Omega(u, x) \in \Omega(u_1, u_2, x) \subset R^3$. If we omit the influence of random perturbations $\xi(t)$ and measurement errors of u_1, u_2, x for simplicity reasons, the process will proceed along the surface $\Omega^H(u, x) \in \Omega(u, x)$, as follows from model (2), which is a surface.

If the process under investigation has a tubular structure, then its model could have the form [1]:

$$x_s(u) = F(u, \alpha, \vec{u}_s, \vec{x}_s), \tag{3}$$

where $F(\cdot)$ is some functional including both the parametric component and the corresponding nonparametric estimates; $\vec{u}_s, \vec{x}_s$ are time vectors $\vec{u}_s = (u_1, ..., u_s)$,$\vec{x}_s = (x_1, ..., x_s)$. Model (3) represents the genetic formation of the methods of parametric and nonparametric identification (at the present time [2] such methods abroad are called data-based methods), in other words it is a "child of parents" of parametric and local approximation methods.

In the case of the H-process, the model (2) can be corrected as follows (H-model):

$$\hat{x} = \hat{f}(u, \alpha_s) I(u), \tag{4}$$

where an indicator $I(u)$ has the form

$$I(u) = \begin{cases} 1, \text{if } u \in \Omega^H(u), \\ 0, \text{if } u \notin \Omega^H(u). \end{cases} \tag{5}$$

We only note that the region $\Omega^H(u) \subset \Omega(u)$ is not known to us, but only the sample $\{u_i, x_i, i = \overline{1, s}\}$ is known. If the indicator $I(u)$ is equal to zero, then the

estimate $\hat{x}(u)$, $\hat{x}_s(u)$ can not be computed, i.e. with such values of the components of the vector $u \in \Omega(u)$ process can not proceed. If the indicator $I(u)$ for any value $u \in \Omega(u)$ is equal to one, then the model (4) coincides with (2). As an estimation of the indicator, we can define the following approximation:

$$I_s(u) = sgn \sum_{i=1}^{s} \Phi\left(c_s^{-1}(x_s(u) - x_i)\right) \prod_{j=1}^{k} \Phi\left(c_s^{-1}(u^j - u_i^j)\right), \tag{6}$$

where

$$x_s(u) = \sum_{i=1}^{s} x_i \prod_{j=1}^{k} \Phi(c_s^{-1}(u^j - u_i^j)) / \sum_{i=1}^{s} \prod_{j=1}^{k} \Phi(c_s^{-1}(u^j - u_i^j)), \tag{7}$$

where c_s is a smoothing parameter, $\Phi(\cdot)$ is a kernel function.

We give the following example relating to the identification of the inertial-free system. Let the object be described by the equation:

$$x(u) = f(u_1, u_2, u_3), \tag{8}$$

where three-dimensional vector $u_1, u_2, u_3 \in R^3$ is an input variable, $x \in R^1$ is an output variable.

Let the components of the vector of input variables $u = (u_1, u_2, u_3)$ be stochastically independent. In this case, it is natural to use the usual path described above.

Further, suppose that objectively the components of the vector of input variables are functionally connected, for example, as follows:

$$u_2 = \phi_1(u_1), u_3 = \phi_2(u_2) = \phi_2(\phi_1(u_1)). \tag{9}$$

Naturally, the researcher does not know about the existence of dependencies (9). Otherwise, we could make the substitution (9) in (8) and obtain the dependence of x on one variable u_1 of the form:

$$x(u) = f(u_1, \phi_1(u_1), \phi_2(\phi_1(u_1))), \tag{10}$$

Thus, the dependence (10) under the conditions indicated above can be reduced to a one-dimensional dependence of x on u_1.

In case of the dependence u_3 on u_2 is absent, equation (8) is easily reduced to the form:

$$x(u) = f(u_1, \phi_1(u_1), u_3), \tag{11}$$

i.e. bivariate dependence x on u_1, u_3.

Hence we can conclude that in the presence of a functional dependence between the components of the vector u, we obtain the dependence x on u, in this case we have one-, two-, three-dimensional dependence. And now we analyze more interesting case, which is directly related to H-processes. Let u_3 and u_2, although in an unknown way, but stochastically connected. First, if the components of the vector u are independent,

then the process under investigation is described by a function of three variables. If two components of the vector of input variables u are connected by a functional dependence, then the process is described by a function of two variables. Secondly, if two variables are stochastically connected, then it turns out that the process is described by a function of more than two variables, but less than three! Here we come to dependence on a fractional number of variables and, consequently, to a space of fractional dimension. This fact was already known in mathematics, although its origins lie in the field of geometric studies of natural objects and are described in B. Mandelbrot's book "Fractal Geometry of Nature". Here is a small quotation from this book: "Liquid, gas, solid - three habitual physical states of matter existing in the three-dimensional world. But what is the dimension of the club of smoke, clouds, or rather their boundaries, continuously eroded by turbulent air movement? It turned out that it is more than two, but less than three. Fractional value! Similarly, you can calculate the dimension of other real natural objects - for example, a shoreline washed by a surf, or the crown of a tree rustling in the wind. The circulatory system of human is pulsating, alive it has a dimension of 2.7"[3]. Previously, the fractional dimension of space was known as the dimension of the Hausdorff-Besicovitch space. The experience of developing some computer systems for modeling and controlling discrete-continuous processes leads us to the conclusion that many really existing processes can be referred to the class of H-processes, and their models to the class of H-models.

Let consider the following situation. For simplicity of considerations, let the process of interest be described by a function $x(u) = f(u_1, u_2, u_3)$.

In the case of stochastic dependence between the variables, according to the available training samples, it is possible to calculate the quadratic approximation errors $u_{2s}(u_1)$, $u_{3s}(u_1)$ where are $u_{2s}(u_1)$, $u_{3s}(u_1)$ are the non-parametric estimates of the Nadaraya-Watson class [4, 5, 6]. We introduce the approximation errors in the form:

$$\delta_{21} = \sigma_{u_2}^{-2} \sum_{i=1}^{s} (u_2 - u_{2s}(u_1))^2, \delta_{31} = \sigma_{u_3}^{-2} \sum_{i=1}^{s} (u_3 - u_{3s}(u_1))^2, \tag{12}$$

where δ_{21} and δ_{31} are the quadratic errors obtained by non-parametric reconstruction of the defined dependences. In the presence of a function of many variables, other variants of the dependencies of some components of the input vector from others can be defined. So, in the previous example, the force of the stochastic relation λ between two arbitrary variables can be calculated, for example, from formula:

$$\lambda_{21} = 1 - \delta_{21}, \lambda_{31} = 1 - \delta_{31}. \tag{13}$$

Thus, the strongest stochastic connection (the functional one) is equal to 1, and if $\lambda \approx 0$, then there is no connection, and in case of stochastic dependence between the input variables it is $0 < \lambda < 1$.

If we interpret H-processes in a more general case as a function of several variables, then the variability of this function in time can be shown on the following chain of relations acting in time:

$$
\begin{aligned}
x &= f(t, \boldsymbol{u_1}, \boldsymbol{u_2}, \boldsymbol{u_3}, \boldsymbol{u_4}, \boldsymbol{u_5}, \quad , \boldsymbol{u_7}, \boldsymbol{u_8}, \boldsymbol{u_9}, \boldsymbol{u_{10}}), t \in T1 \\
x &= f(t, u_1, \ \boldsymbol{u_2}, \boldsymbol{u_3}, u_4, \ \boldsymbol{u_5}, \quad , \boldsymbol{u_7}, \boldsymbol{u_8}, u_9, \ \boldsymbol{u_{10}}), t \in T2 \\
x &= f(t, u_1 \ \boldsymbol{u_2}, \boldsymbol{u_3}, u_4, \ \boldsymbol{u_5}, \quad , \boldsymbol{u_7}, u_8, \ u_9, \ \boldsymbol{u_{10}}), t \in T3 \\
x &= f(t, u_1, \ \boldsymbol{u_2}, \boldsymbol{u_3}, u_4, \ \boldsymbol{u_5}, \quad , \boldsymbol{u_7}, u_8, \ u_9, \ \boldsymbol{u_{10}}), t \in T4 \\
x &= f(t, \quad , \boldsymbol{u_2}, \boldsymbol{u_3}, u_4, \ \boldsymbol{u_5}, u_6, \ \boldsymbol{u_7}, u_8, \quad , \boldsymbol{u_{10}}), t \in T5 \\
x &= f(t, \quad , \boldsymbol{u_2}, \boldsymbol{u_3}, u_4, \ \boldsymbol{u_5}, u_6, \ \boldsymbol{u_7}, u_8, \quad , u_{10}), \ t \in T6 \\
x &= f(t, \quad , \boldsymbol{u_2}, u_3, \ u_4, \ \boldsymbol{u_5}, u_6, \ \boldsymbol{u_7}, \quad , \quad , u_{10}), \ t \in T7 \\
x &= f(t, \quad , \boldsymbol{u_2}, u_3, \quad , \boldsymbol{u_5}, u_6, \ u_7, \ u_8, \quad , u_{10}), \ t \in T8 \\
x &= f(t, u_1, \ \boldsymbol{u_2}, u_3, \ u_4, \ \boldsymbol{u_5}, \boldsymbol{u_6}, u_7, \ \boldsymbol{u_8}, \ , u_{10}), \ t \in T9
\end{aligned}
\tag{14}
$$

Thus, in the actual H-processes, the influence of the value of variables changes significantly: some variables may lose their influence on x, some variables may lose the value first and then restore it, and some variables may appear for the first time, such as u_6, u_7.

If we preserve the mathematical appearance of the interpretation of a function of several variables as a point of a multidimensional space, then in the presence of an H-process we arrive at a space of fractional dimension F^λ.

The calculation of the dimension F^λ can be done, for example, as follows:

$$
\dim F^\lambda = (n+1) - \sum_{i=1}^{n-1} \lambda_{i,i+1}, \tag{15}
$$

where n is dimensionality of vector u, $\lambda_{i,i+1}$ is force of the stochastic relation between two variables u_i and u_{i+1}.

Other schemes for computing the dimensionality of space can also be proposed, for example:

$$
\dim F_1^\lambda = (n+1) - \sum_{i=1}^{n-1} \lambda_{1,i+1}, \tag{16}
$$

where $\lambda_{1,i+1}$ is the dependence of all components of the vector u on one component u_1.

A careful analysis of the chain of relations (14) leads to the following reflections. Functions can be represented as a series or parametrized in some way, and these parameterized relationships can change in time in some way due to the variability of variables, in other words, lose (or acquire) their influence on the variable x. The latter is dictated by the properties of the real process under investigation.

2 T-models

Consider processes whose output variables have unknown stochastic constraints, called T-processes, and their models, respectively, T-models. Processes of this type are described by a system of implicit functions of the form (T-model):

$$\hat{F}_j(u^{<j>}, x^{<j>}, \vec{x}_s, \vec{u}_s) = 0, j = \overline{1,m}, \tag{17}$$

where $u^{<j>}, x^{<j>}$ are composite vectors, $\vec{x}_s, \vec{u}_s$ are time vectors (a set of observations arriving at the s-th time point), in particular $(x_1, ..., x_s) = (x_{11}, x_{12}, ..., x_{21}, x_{22}, ..., x_{2s}, ..., x_{m1}, x_{m2}, x_{ms})$.

In this case, the estimation of the components of the output variable vector x for known values u, as already noted above, leads to the need to solve the system of equations (17).

The problem reduces to the fact that for a given value of the vector of input variables $u = u'$, it is necessary to solve system (17) with respect to the vector of output variables x. The general scheme for solving such a system is:

- first, the current value of the input variables $u_t = (u_{t1}, ..., u_{tn})$ is substituted into equation (17). This places the subdomain in the observation space $\vec{x}_s, \vec{u}_s$. And then the errors are computed:

$$\varepsilon_{i,j} = \hat{F}_j(u^{<j>}, x^{<j>}(i), \vec{x}_s, \vec{u}_s), j = \overline{1,m}, \tag{18}$$

where $\hat{F}_j(u^{<j>}, x^{<j>})$ has the following form:

$$\varepsilon_j(i) = \hat{F}_j(u^{<j>}, x_j(i)) = x_j(i) - \frac{\sum_{i=1}^{s} x_j[i] \prod_{k=1}^{<n>} \Phi\left(\frac{u_k - u_k[i]}{c_{su_k}}\right)}{\prod_{k=1}^{<n>} \Phi\left(\frac{u_k - u_k[i]}{c_{su_k}}\right)}, \tag{19}$$

where $< n >$ is the dimension of the composite vector u_k, $< n > \leq n$. In the following, this notation is also used for other variables.

- The next step is to estimate the conditional mathematical expectation:

$$x_j = M\{x|u^j, \varepsilon = 0\}, j = \overline{1,m}. \tag{20}$$

As an estimation (20), we take a non-parametric estimation of the Nadaraya-Watson regression [9]:

$$\hat{x}_j = \frac{\sum_{i=1}^{s} x_j[i] \prod_{k_1=1}^{<n>} \Phi\left(\frac{u_{k_1} - u_{k_1}[i]}{c_{su}}\right) \prod_{k_2=1}^{<m>} \Phi\left(\frac{\varepsilon_{k_2}[i]}{c_{s\varepsilon}}\right)}{\prod_{k_1=1}^{<n>} \Phi\left(\frac{u_{k_1} - u_{k_1}[i]}{c_{su}}\right) \prod_{k_2=1}^{<m>} \Phi\left(\frac{\varepsilon_{k_2}[i]}{c_{s\varepsilon}}\right)}, j = \overline{1,m} \tag{21}$$

The representation (18) in the form (19) is due to the possibility of approximation the corresponding component x by the local properties of the non-parametric estimators.

3 K-models

When modeling multidimensional processes, there is often a situation in which the description of the corresponding models differs in different channels due to various a priori information on different channels. In this case, the model of the process under investigation is a synthesis of fundamental laws, well-tested parametric models and non-parametric models.

In other words, K-models fundamentally differ from generally accepted in that they take into account all the variables that determine the multidimensional process, depending on the level of a priori information about the various channels of the process. Thus, K-models are an organic synthesis that describes the process or system of interrelated objects in all their diversity.

4 Control algorithms of H-processes

Many processes occurring in natural phenomena in living organisms, as well as many technological processes, has tubular structures in the space of measured variables. The researcher most often does not know about the presence of a tubular structure. And in this case, when constructing a model of similar processes or control algorithms, a researcher starts from the fact that on the basis of the available a priori information the parametric structure of the model is somehow determined or simply postulated. The next step is usually reduced to estimating the parameters which are included into the corresponding models or algorithms. In fact, the process under investigation proceeds in some of its subregions. In this case, the usual methods of constructing models must differ from those generally accepted in the theory of identification. It seems appropriate to recall one, which has already become a historical fact. In [7] R.E. Kalman remembers one phrase, expressed by L.S. Pontryagin during his stay at Stanford: "Mathematicians do not believe in probability." He also drew attention to the difficulties of working with real data on the basis of the classical (Kolmogorov) theory of probability. Further he said [7]: "I agree with Kolmogorov that with the statistics is something wrong".

Below we show a possible version of the H-process. From considerations of simplicity of visualization, we give such a process for the case of three-dimensional space.

In Figure 1, the following designations are used: $u_1, u_{\in}\Omega(u) \subset R^2$ is input actions, and $x \in \Omega(x) \subset R^1$ is output action, $\Omega(u,x)$ is a single hypercube, $\Omega^H(u,x) \subset \Omega(u,x)$ is subdomain in which the process under investigation proceeds. Moreover, $\Omega(u,x)$ is always known, but $\Omega^H(u,x)$ is almost never known.

When controlling multidimensional inertial-free systems with delay, where H-processes take place through different channels of the object, certain peculiarities arise. The main one is that the defining values of the output variables can not be determined arbitrarily, as it is most often accepted in control theory. We show this for an object that has two input and two output variables.

Figure 2 shows two H-processes along the channels "$x_1 - u$", "$x_2 - u$". The input variable μ is a measured but uncontrolled vector. As we see above, these H-processes

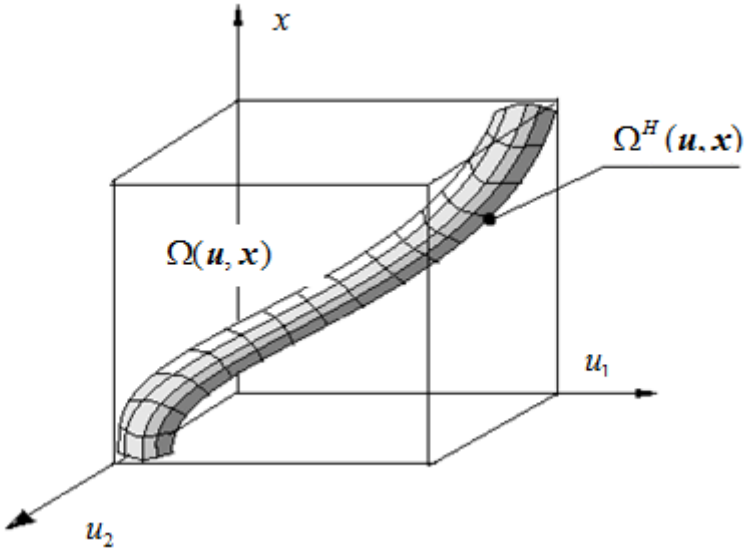


Figure 1: H-process

do not intersect in the space of input - output variables. From this it follows that both driving influences x_1^* and x_2^* can not be achieved simultaneously while controlling this system by u. Below are shown the intersecting H-processes along the same channels [8, 9].

From this figure it is clear that two H-processes have a joint subdomain $G^H(u,x) = \Omega^{H^1}(u,x) \cap \Omega^{H^1}(u,x)$. It is in this H-domain we should determine the values of setting actions x_1^* and x_2^*. To do this, we form from the initial training sample $\{u_i, x_i, i = \overline{1,s}\}$ the elements belonging to the region $G^H(u,x)$. For example, as follows:

$$\text{if } \sum_{i=1}^{s}\prod_{j=1}^{k}\Phi\left\{\frac{u^j-u_i^j}{c_s^u}\right\}\prod_{j=1}^{n}\Phi\left\{\frac{\mu^j-\mu_i^j}{c_s^\mu}\right\}\prod_{j=1}^{m}\Phi\left\{\frac{x^j-x_i^j}{c_s^x}\right\} > 0, \text{ then } (u_i,\mu_i,x_i) \in G^H(u,x),$$

$$\text{if } \sum_{i=1}^{s}\prod_{j=1}^{k}\Phi\left\{\frac{u^j-u_i^j}{c_s^u}\right\}\prod_{j=1}^{n}\Phi\left\{\frac{\mu^j-\mu_i^j}{c_s^\mu}\right\}\prod_{j=1}^{m}\Phi\left\{\frac{x^j-x_i^j}{c_s^x}\right\} = 0, \text{ then } (u_i,\mu_i,x_i) \notin G^H(u,x),.$$

where $G^H(u,x)$ is the subdomain of the intersection of all H-processes (in this case two). Non-parametric control algorithms, taking into account the foregoing, are nonparametric Nadaraya-Watson chains that compute from the sample belongigng to $G^H(u,x,\mu)$ the mathematical expectations of control for fixed setting actions $x^*(t) \in G^H(x)$ and $\mu \in G^H(\mu)$ and the preceding components of the vector u which are already found.

Conclusion

The result of the above is an analysis of the peculiarities that arise when modeling processes of the "tubular" structure, which always takes place if the components of

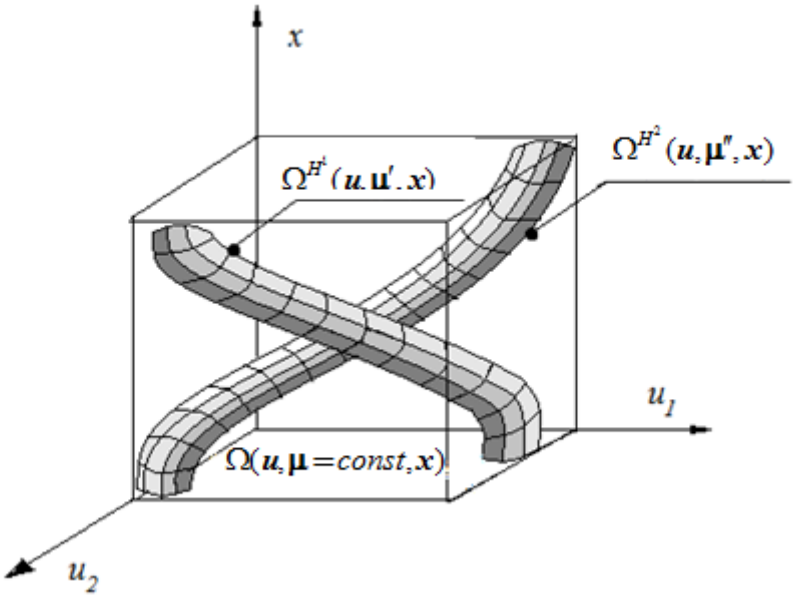


Figure 2: Nonintersecting H-processes

the vector of the input variables of the process are stochastically dependent. In this case, the traditional models of static systems with delay are not applicable or can lead to significant errors. The most interesting is that we come to a space of fractional dimension. Of course, the fact of the disappearance and appearance of the role of certain input variables in different time periods in the values of the output variables of the process is important. This fact is closely related not only to the space of fractional dimension, but also to the space of varying dimension.

When controlling multidimensional H-processes, first of all, it is necessary to determine the corresponding setpoints for the output variables, and then use non-parametric control algorithms, which are called the matrix $\mathscr{U}^m$-regulator. This is an essential feature of the construction of the control system, in contrast to the traditional control algorithms for multidimensional inertial-free systems. In the present paper, all the H-processes shown in Figures 1 - 3 are considered in three-dimensional space and in a "frozen" form for reasons of visualization simplicity. In fact, the reduced H-models and control algorithms function when the real processes are in motion.

References

[1] Medvedev A.V. (1995). Data Analysis in Identification Problems. *Computer data analysis and modeling*. Vol. **2**. pp. 201-206.

[2] Block G., Hartman A. M., Dresser C. M., Carroll M. D., Gannon J., Gardner L. (1986). A data-based approach to diet questionnaire design and testing. *American journal of epidemiology*. Vol. **124**. pp. 453-469.

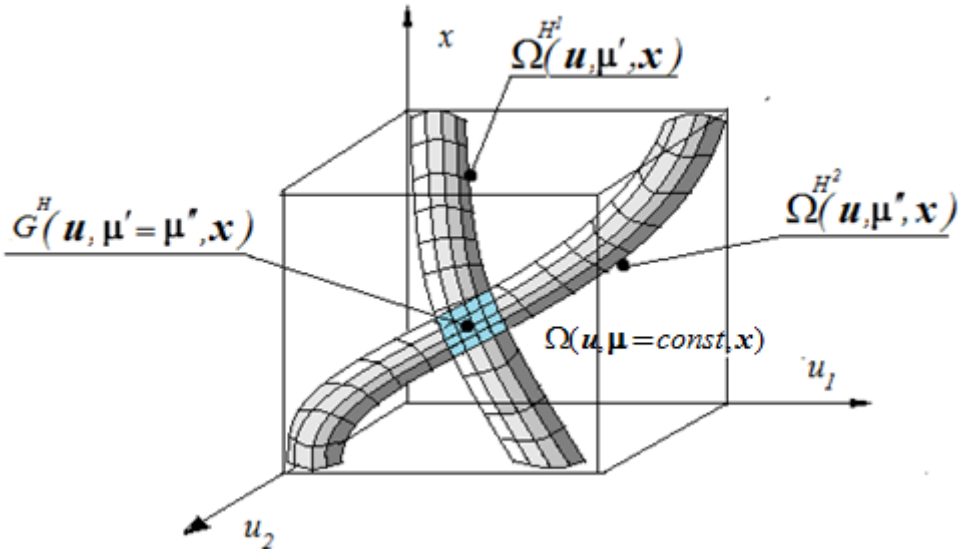


Figure 3: Intersecting H-processes

[3] Mandelbrot B. B., Pignoni R. (1983). The fractal geometry of nature. W. H. Freeman and Company, New York.

[4] Nadaraya E. A. (1965). Nonparametric estimates of the regression curve. *Trudy VTS AN GSSR*. Vol. **5**. pp. 56-68.

[5] Vasiliev V.A., Dobrovidov A.V., Koshkin G.M. (2004). Nonparametric estimation of functionals from distributions of stationary sequences. Nauka, Moscow.

[6] Kotkinik V. Ya. (1985). Nonparametric identification and data smoothing. Nauka, Moscow.

[7] Kalman R. (1985). Identification of systems with noises. *Uspekhi mat. nauk*. Vol. **40**. pp. 27-41.

[8] Medvedev A. V. (2011). Nonparametric Approximation in Adaptive Systems Theory.*Works of Applied Methods of Statical Analysis. Simulation and Statistical Inference*. Novosibirsk, pp. 195-212.

[9] Medvedev A.V. (2014) Some remarks on H-models of inertia-free processes with delay. *Vestnik of Siberian State University*. Vol. **54**. pp. 50-54.

Non-parametric Dual Control Algorithms of Discrete-continuous Processes with Dependent Input Variables

Ekaterina A. Chzhan
Siberian Federal University, Krasnoyarsk, Russia
e-mail: ekach@list.ru

Abstract

In the report we propose a non-parametric dual control algorithm for multidimensional processes with stochastic dependence between input variables. The algorithm does not depend on the equation of the described object, it is based on nonparametric estimates of the Nadaraya-Watson class. The object of investigation has a tubular structure, i.e. the input variables are related by stochastic dependence. Series of computational experiments are carried out for both the multidimensional inertialess process with delay and for the tubular process, the results of which shows the effectiveness of the proposed methods.

Keywords: non-parametric statistics, dual control, static system, inertialess process.

Introduction

This article focuses on the problem of the non-parametric dual control which was firstly suggested by Feldbaum [1]. The major idea is to combine the processes of control and learning the object in order to get new information. However, it is necessary to construct such model that the disturbance of the object does not contradict with the goal of control purpose. This theory was extensively developed by Wittenmark [2]. In [3], authors discuss two situations when it is appropriate to apply dual control algorithms: a short time horizon and rapidly changing object parameters. The dual control theory was used in case of investigation the linear stochastic systems with unknown parameters [4]. The paper [5] contains a very broad review of dual control algorithms for systems which structure is specified within the accuracy of parameter vector. There are some examples of dual control theory in different spheres such as paper coating [6] or diabetes investigation [7]. Dual control approach is widely used in the development of adaptive model predictive control [8, 9]. In practice, in most cases the mathematical structure of proceeding processes are unknown. So, in this situation, it is advisable to use the non-parametric dual control algorithms [10]. In this paper we suggest dual control algorithm for multidimensional objects with dependent input variables.

The rest of the paper is organized as follows. In section 2, we present the statement of the dual control problem. In section 3, we propose non-parametric dual control algorithm. In section 4, the results of the numerical experiments of modeling multidimensional objects are described. We conclude our work in Section 5.

1 The Statement of the Problem

Consider a control system, its general scheme is shown in Fig. 1. The notation is as follows: $x(t) = (x_1(t), x_2(t), ..., x_n(t)) \in \Omega(x) \subset R^n$ is an output variable of the process, $x^*(t) = (x_1^*(t), x_2^*(t), ..., x_n^*(t)) \in \Omega(x) \subset R^n$ is a vector of desired output (set point), $u(t) = (u_1(t), u_2(t), ..., u_m(t)) \in \Omega(u) \subset R^m$ is a control input vector, $\mu(t) = (\mu_1(t), \mu_2(t), ..., \mu_k(t)) \in \Omega(\mu) \subset R^k$ is uncontrolled input vector, $\xi(t)$ is a vector random disturbances, (t) is continuous time, $G^{\mu_1}, G^{\mu_2}, ..., G^{\mu_k}$, $G^{x_1}, G^{x_2}, ..., G^{x_n}$ are the system response channels corresponding to different variables and including control tools, $g^{\mu}(t) = (g^{\mu_1}(t), g^{\mu_2}(t), ..., g^{\mu_k}(t)) \in \Omega(g^{\mu}(t)) \subset R^k$, $g^x(t) = (g^{x_1}(t), g^{x_2}(t), ..., g^{x_n}(t)) \in \Omega(g^x(t)) \subset R^n$ are random inaccuracy of measurements of variables of the process with zero mathematic expectation and limited dispersion. The peculiarity of the process is that the input variables are related with some stochastic dependence, the form of which is unknown to the researcher. In Fig. 1 an arc-shaped lines show a variant of possible dependencies.

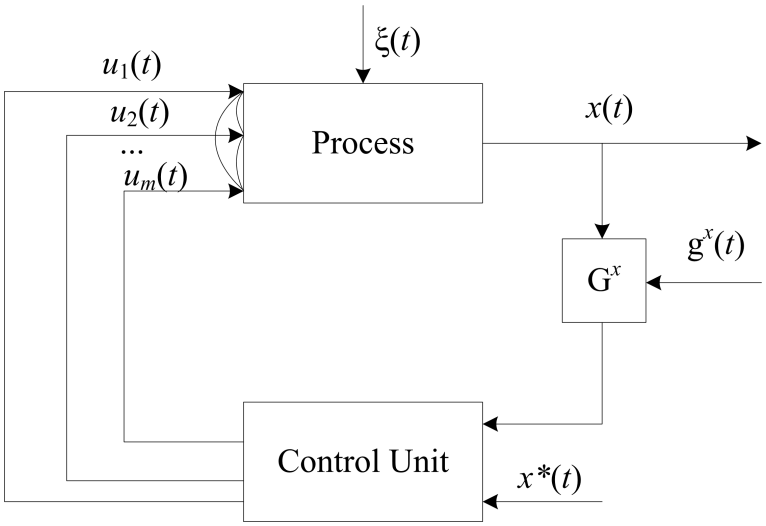


Figure 1: The general scheme of closed loop system

The aim of the control unit algorithm to set a control action $u(t)$ in order to lead the object output value $x(t)$ to the desired value $x^*(t)$. For this purpose, the non-parametric dual algorithm is suggested [10].

2 The control algorithm for multidimensional static process with delay

We investigate the dual control algorithms which were first proposed by Feldbaum [1]. The control aim of such algorithms has dual nature: caution and probing [3]. Feldbaum considered a situation when the structure of the model and the laws of the

distribution of the random noises are known. In [10], the idea of applying the non-parametric estimation of regression function in control systems was firstly suggested for the object with one input and one output variable. The method is robust to non-parametric uncertainty: the mathematical description of the object is unknown.

We propose the control algorithm for a multidimensional object with dependent input variables. The non-parametric dual control algorithm can be represented as follows:

$$u_{j,s+1} = u^*_{j,s} + \Delta u_{j,s+1}, j = 1, 2, \ldots, m, \tag{1}$$

where the component $u^*_{j,s}$ accumulates the knowledge about the object, the component $\Delta u_{j,s+1}$ is the "learners" search step.

In multidimensional case when the output variable $x(t)$ is a vector the search step could have the form:

$$\Delta u_{j,s+1} = \sum_{i=1}^{n} \Theta_i \left(x^*_{i,s+1} - x_{i,s} \right), j = 1, 2, \ldots, m, \tag{2}$$

where $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_n)$ could be found as a minimum of quadratic criterion:

$$R(\Theta_1, \Theta_2, \ldots, \Theta_n) = \left[\sum_{i=1}^{s-1} \sum_{p=1}^{n} \left(x^*_{p,i+1} - \frac{\sum\limits_{j=1}^{s-1} x_{p,j+1} \prod\limits_{q=1}^{m} \Phi \left(\frac{u^*_{q,i} + \sum\limits_{w=1}^{n} \Theta_{w(a)} - u_{q,j+1}}{c^u_s} \right)}{\sum\limits_{j=1}^{s-1} \prod\limits_{q=1}^{m} \Phi \left(\frac{u^*_{q,i} + \sum\limits_{w=1}^{n} \Theta_{w(a)} - u_{q,j+1}}{c^u_s} \right)} \right) \right]^2 \to \min_{(\Theta_1, \Theta_2, \ldots, \Theta_n)}, \tag{3}$$

where $a = x_{w,i+1} - x_{w,i}$.

For the first component $u_1(t)$ of control variable $u(t)$ the addend $u^*_{1,s}$ from the equation (1) can be calculated as a non-parametric estimation of the inverse regression function for discrete observations $\{u_i, x_i, i = 1, 2, \ldots, s\}$ in the following form which was suggested in [10]:

$$u^*_{1,s} = \frac{\sum\limits_{i=1}^{s} u_{1,i} \prod\limits_{v=1}^{n} \Phi \left(\frac{x^*_{v,s+1} - x_{v,i}}{c^x_s} \right)}{\sum\limits_{i=1}^{s} \prod\limits_{v=1}^{n} \Phi \left(\frac{x^*_{v,s+1} - x_{v,i}}{c^x_s} \right)}, \tag{4}$$

where $\Phi \left(\frac{x^*_{v,s+1} - x_{v,i}}{c^x_s} \right)$ is a kernel function, c^x_s is a smoothing parameter. Kernel function and smoothing parameter satisfy convergence conditions [11, 12]. For example, for kernel function $\Phi(z)$, $z = \frac{x^*_{v,s+1} - x_{v,i}}{c^x_s}$ and the smoothing parameter c^x_s the conditions are as following:

$$\begin{array}{ll} c_s^x > 0; & 0 \le \Phi(z) < \infty; \\ \lim\limits_{s\to\infty} c_s^x = 0; & \int\limits_{\Omega(z)} \Phi(z)\, dz = 1; \\ \lim\limits_{s\to\infty} s(c_s^x)^m = \infty; & 1/c_s^x \lim\limits_{s\to\infty} \Phi(z) = \delta(c_s^x z). \end{array} \tag{5}$$

The main idea is that control input $u(t)$ is consistently found for every component, each subsequent value $u_i, i = 2, 3, \ldots, m$ depends on the value $u_i, i = 1, 2, \ldots, m-1$ found in the previous step. The estimation of $u_{j,s}^*,\ j = 2, 3, \ldots, m$ is based on a Nadaraya-Watson estimation of inverse regression function which refers to the local approximation methods [12].

So for components $u_j(t), j = 2, 3, \ldots, m$ addend $u_{j,s}^*, j = 2, 3, \ldots, m$ is proposed to calculate due to the formula:

$$u_{j,s}^* = \frac{\sum\limits_{i=1}^{s} u_{j,i} \prod\limits_{w=1}^{j-1} \Phi\left(\frac{u_{w,s+1}-u_{w,i}}{c_s^u}\right) \prod\limits_{l=1}^{k} \Phi\left(\frac{\mu_{l,s+1}-\mu_{l,i}}{c_s^\mu}\right) \prod\limits_{v=1}^{n} \Phi\left(\frac{x_{v,s+1}^*-x_{v,i}}{c_s^x}\right)}{\sum\limits_{i=1}^{s} \prod\limits_{w=1}^{j-1} \Phi\left(\frac{u_{w,s+1}-u_{w,i}}{c_s^u}\right) \prod\limits_{l=1}^{k} \Phi\left(\frac{\mu_{l,s+1}-\mu_{l,i}}{c_s^\mu}\right) \prod\limits_{v=1}^{n} \Phi\left(\frac{x_{v,s+1}^*-x_{v,i}}{c_s^x}\right)}. \tag{6}$$

3 Computer Experiment

The investigated process has three input variables $u(t) = (u_1(t), u_2(t), u_3(t))$ and two output variables $x(t) = (x_1(t), x_2(t))$. Let the object be described by the following equations:

$$\begin{cases} x_1(t) = 2u_1(t) + \sqrt{u_2(t)} + 0.5 \sin u_3(t) + g^{x_1(t)}, \\ x_2(t) = 1.5u_1(t) + 0.3 \exp u_2(t) + 2u_3(t) + g^{x_2(t)}, \\ u_3(t) = 0.3u_1(t) + \sqrt[3]{u_2(t)} + g^{u_3(t)}; \end{cases} \tag{7}$$

For each output variable $x_1(t), x_2(t)$ the control error $\varepsilon_1, \varepsilon_2$ of control is

$$\varepsilon_j = \sqrt{\frac{s^{-1} \sum\limits_{i=1}^{s} (x *_{ji} - x*_{ji})^2}{(s-1)^{-1} \sum\limits_{i=1}^{s} (x_{ji} - \hat{m}_{x_j})^2}}, j = 1, 2, \tag{8}$$

where $\hat{m}_{xj}$ is an estimation of the output variable $x_j, j = 1, 2$.

Also, we calculate the multidimensional control error:

$$W = \sqrt{\frac{s^{-1} \sum\limits_{i=1}^{s} \sum\limits_{j=1}^{2} (x *_{ji} - x*_{ji})^2}{(s-1)^{-1} \sum\limits_{i=1}^{s} \sum\limits_{j=1}^{2} (x_{ji} - \hat{m}_{x_j})^2}}, j = 1, 2, \tag{9}$$

There is a combined way of data accumulation [11]. When the control algorithm starts to work, there are already some observations which were passively accumulated. In this experiment the size of the learning sample is 200 observations: $\{u_i, x_i, i = 1, 2, ..., 200\}$. The control results for $x_1(t), x_2(t)$ of the object (7) are shown in Fig. 2. Errors of control are the following: $\varepsilon_1 = 0.04$, $\varepsilon_2 = 0.08$, $W = 0.07$.

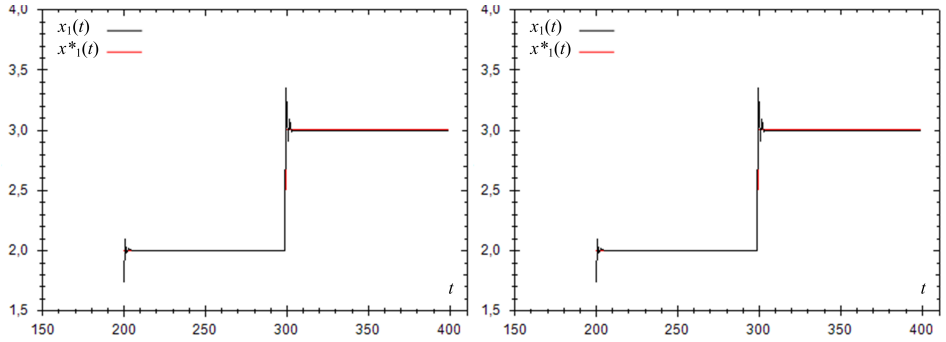


Figure 2: The results of the operation of the control algorithm for object (7)

As it can be seen from the figure 2 and small value of control errors, the control algorithm gives a reasonably accurate result. The graphs of the control actions $u_1(t)$, $u_2(t)$, $u_3(t)$ are shown in figure 3.

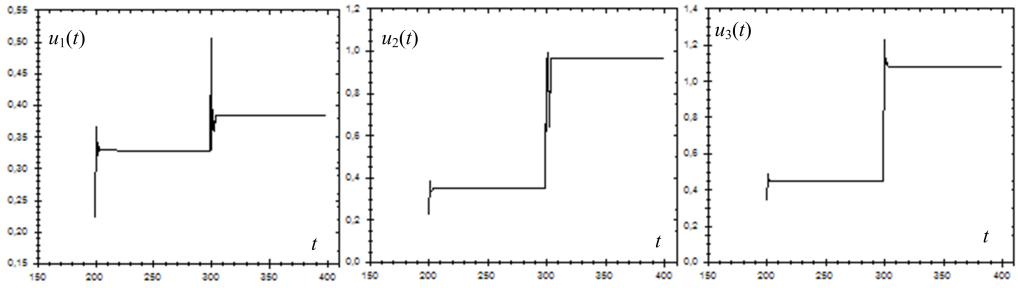


Figure 3: The graph of the control actions $u_1(t)$, $u_2(t)$, $u_3(t)$

Conclusions

In this paper we propose the new non-parametric control algorithm of multidimensional object with dependent input variables. We consider the situation when the mathematical structure of the controlled object is unknown. The combined method of accumulating information is more in line with the objectives of practice as it allows to use all available information. When the full range of possible values of the variables is covered, the algorithm will be adjusted automatically with a new value of the desired output. Computational experiments have shown the effectiveness of the proposed algorithm for controling the nonlinear multi-dimensional object at different levels of interference.

Acknowledgements

I would like to express my gratitude to my scientific supervisor Medvedev A.V. for the valuable advice in the planning of the research and recommendations on the design of the article.

References

[1] A.A. Feldbaum, Fundamentals of the theory of optimal automatic systems, Fizmatgiz Publishing, Moscow, 1963.

[2] B. Wittenmark, Adaptive dual control methods: An overview, in Proc. 5th IFAC Symp. Adaptive Syst. Control Signal Processing. 1995, 67-73.

[3] K. Astrom, B. Wittenmark, Problems of identification and control, J. Math. Anal. Appl. 34 (1971) 90-113.

[4] C.J. Wenk, Y. Bar-Shalom, A multiple model adaptive dual control algorithm for stochastic systems with unknown parameters, Automatic Control. 25 (1980) 703-710.

[5] N. M. Filatov, H. Unbehauen, Survey of adaptive dual control methods, Proc. IEE Control Theory Appl. 1 (2000) 119-128.

[6] A. Ismail, G. A. Dumont, J. Backstrom, Dual adaptive control of paper coating. IEEE Transactions on Control Systems Technology. 11 (2003) 289-309.

[7] A. Bhattacharjee, A. Sutradhar, Data driven nonparametric identification and model based control of glucose-insulin process in type 1 diabetics, Journal of Process Control. 41 (2016) 14-25.

[8] Tor Aksel N. Heirung, Bjarne Foss, B. Erik Ydstie, MPC-based dual control with online experiment design, Journal of Process Control. 32 (2015) 64-76.

[9] G. Marafioti, Enhanced Model Predictive Control: Dual Control Approach and State Estimation Issues (PhD thesis). Norwegian University of Science and Technology (2010).

[10] A.V. Medevev, The theory of non-parametric systems, Control-I, Vestnik SibGAU. 48 (2013) 57-63.

[11] A.V. Medvedev, Fundamentals of Adaptive Systems, SibSAU Publishing, Krasnoyarsk, 2015. (In Russ.).

[12] E. A. Nadaraya, On Non-Parametric Estimates of Density Functions and Regression Curves, Theory of probability & its Applications. 10 (1965) 186-190.

Double Loop Control of Linear Dynamical Systems and an Algorithm for Adjustment of the Typical Controllers Using the Nonparametric Model of a Llinear Dynamical System

Pupkov A., Tsarev R.
Siberian Federal University, Krasnoyarsk, Russia
e-mail: APupkov@mail.sfu-kras.ru

Abstract

This paper considers a double loop control of a linear dynamic system based on a nonparametric model. It describes a two-stage process of constructing a nonparametric controller, which constructs a nonparametric model of a macro-object at the first stage, and then a nonparametric controller at the second stage. The double loop control scheme preserves analog devices of local automatic equipment, which ensures its reliability even in the case when digital devices fail. The authors describe an approach to synthesis of a nonparametric controller for a linear dynamical system of an unknown order when information about a control object is presented in the form of realization of transition functions containing random disturbance. It further proposes an algorithm for realization of assessment of the inverse function for a linear dynamic system, and provides a novel approach to standard controller parameters adjustment using a nonparametric model of a linear dynamical system.

Keywords: nonparametric controller, double loop control, linear dynamical system, transition function.

Introduction

One of the most inescapable links in the modern information technologies is the control technology. In many cases, technological processes and objects belong to the class of linear ones. In this regard, the problem of algorithms development for dynamic processes control remains crucial nowadays. As a rule, P and Pi control laws are applied to these processes, and in this case, the main task is to adjust the parameters of relevant controllers which provide a sufficient level of quality. The most commonly used control laws are associated with natural losses related to the problem of parameter settings. Moreover, the higher the order of an equation describing the process is, the more significant losses we will have in the result.

The typical control laws are currently implemented using control devices. A nonparametric model of a linear dynamical system can be used to adjust the parameters of a controller. In this case, the adjustment is performed in the process of optimization of the generated criterion for the standard deviation of the output of the nonparametric model and the corresponding approximation of the control object's operator.

The modern control theory equips researchers and practitioners with methods for optimal dual control of linear processes. The application of these methods requires exhaustive knowledge about the control object. In practice, a researcher meets incompleteness of a priori information and the influence of disturbance. In these circumstances, it is necessary to develop new methods and control algorithms implementing these methods, so that they take into account incompleteness of information about the control object. The nonparametric approach to the synthesis of a controller allows to cope with the aforesaid problems.

Analog devices of local automation are sufficiently reliable and cope adequately with the control tasks. Development of a better control system implies using up-to-date digital technology. Bearing in mind the possibility of failure of the digital control devices, one may find it reasonable to keep analog devices, since the experience of their use and adjusting is quite extensive. The purpose of this paper is the development of both a double loop control system which application implies the use of previously installed automation, and an algorithm for standard controllers' adjustment using a nonparametric model.

1 Double loop control scheme

The scheme of double loop control is given in Fig. 1.

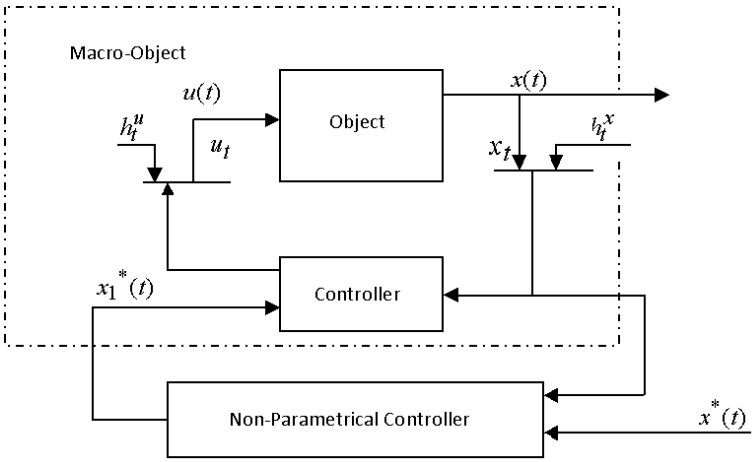


Figure 1: The scheme of double loop control

The scheme of double loop control is given in fig. 1, where: $u(t)$ and $x(t)$ are input and output parameters; $x_1^*(t)$ is a control action for the analog controller, which is generated by a nonparametric controller; $x^*(t)$ is a control action for the nonparametric controller; h_t^u, h_t^x are disturbances in the measurement channels.

Let us describe what the two control loops are. The first control loop is an object with an analog controller, while the second control loop is digital. The first control loop presents the control object for the second control loop, i.e. the first control loop is a macro-object.

The problem is to construct a nonparametric controller of the macro-object using input and output parameters $x_1^*(t)$ and $x(t)$, which are measured from the macro-object at discrete instants of time Δt. We assume that there are disturbances ξ_t, ξ_t^u, ξ_t^x with a zero mean and a limited dispersion in communication channels. The construction of the controller is divided into two stages. The first stage is the construction of a nonparametric model of the macro-object. The second stage is synthesis of a nonparametric controller [3],[4].

The concept of the proposed method for constructing a model of a macro-object is as follows. It is known that the response of a linear dynamic object is described by the Duhamel's integral under zero initial conditions:

$$x(t) = \int_0^t h(t-\tau)u(\tau)\,d\tau, \tag{1}$$

where $h(t)$ is a weight (impulse transition) function of the system.

In our case, when the object is not only a control object but is also an analogue controller, the convolution integral will be:

$$x(t) = \int_0^t h(t-\tau)x_1^*(\tau)\,d\tau, \tag{2}$$

where $h(t)$ is correspondingly transition and weight functions of the macro-object. The above record is valid since both the object and the controller are linear. Then the problem is reduced to estimation of the weight function of the macro-object on the basis of the available measurements.

The weight function $h(t)$ is the first derivative of the transition characteristic with respect to time. Taking into account this fact, it is necessary to obtain an estimation of the transient process based on the sample $(k1, t1), \ldots, (ks, ts)$:

$$k_s(t) = \frac{1}{sC_s}\sum_{i=1}^{s} k_i H\left(\frac{t-t_i}{C_s}\right), \tag{3}$$

where s is the sample size, $H()$ is a bell-shaped functionn, C_s is a parameter of blur, all three satisfy the convergence conditions. The evaluation of the weight function is:

$$h_s(t) = k_s'(t) = \frac{1}{sC_s}\sum_{i=1}^{s} k_i H'\left(\frac{t-t_i}{C_s}\right). \tag{4}$$

Substituting the weight function estimation into the convolution integral, we obtain a nonparametric model of the macro-object:

$$x(t) = \frac{1}{sC_s}\sum_{j=1}^{\frac{t}{\Delta\tau}}\sum_{i=1}^{s} k_i H'\left(\frac{t-\tau-t_i}{C_s}\right)x_1^*(\tau_j)\Delta\tau. \tag{5}$$

The parameter of blur Cs is chosen from the minimum of the root-mean-square criterion [1]:

$$W(C_s) = \sum_{i=1}^{s} (x(t_i) - x_s(t_i, C_s))^2 \to \min_{C_s} \tag{6}$$

The second stage of constructing the macro-object nonparametric controller comes down to estimation of an inverse weight characteristic of the system. It is known that the inverse operator of a linear dynamical system has the same form as a direct operator of this linear dynamical system. The only difference is that the weight and transition functions are defined in the direction of "output-input". Since such realizations can not be obtained on a real object, the inverse characteristics are measured on the linear dynamical system model, solving the equation $x_s(t) = 1(t)$ for the input of the macro-object $x_1^*(t)$. The solution of this equation is an algorithm for calculating points of the inverse transition function $\omega[t]$:

$$\omega[t] = \frac{sC_s - \Delta\tau \sum\limits_{j=1}^{\frac{t-\Delta\tau}{\Delta\tau}} \sum\limits_{i=1}^{s} k_i H'\left(\frac{t-\tau_j-t_i}{C_s}\right)\omega(\tau_j)}{\Delta\tau \sum\limits_{i=1}^{s} k_i H'\left(\frac{-t_i}{C_s}\right)}, \tag{7}$$

where $\omega[0] = 0$.

The resulting implementation $\{(\omega_i, t_i), i = \overline{1,s}\}$ is used to construct the inverse macro-object operator, the evaluation of which is as follows:

$$x_{1,s}^*(t) = \frac{1}{sC_s} \sum_{i=1}^{s} \int_0^t \omega_i H'\left(\frac{t-\tau-t_i}{C_s}\right) x_1^*(\tau) d\tau, \tag{8}$$

where $x_{1,s}^*(t)$ is the evaluation of the inverse operator of the macro-object, ω_i is the implementation of the inverse transition function of the macro-object, $x^*(t)$ is the setting action for the macro-object, the function $H()$ and the parameter of blur C_s satisfy the same convergence conditions.

The transition function is measured on the basis of the object closed by the analog control loop. Thus, the single step action $1(t)$ is the reference for the analog controller but not the input action for the object. Further, when constructing the second control loop, the nonparametric controller generates a control action $x_{1,s}^*(t)$ for the macro-object, which, in its turn, is a reference for the analog controller.

2 Adjustment of parameters of standard controllers using the nonparametric model of a linear dynamic system

Fig. 2 shows a scheme of the control system. $x^*(t)$ is a set value of the controlled variable, $x(t)$ is the controlled variable, $u(t)$ is a control action, $\mu(t)$ is a disturbance,

$W(p)$ and $W_p(p)$ are correspondingly the operators of the linear system and of the controller.

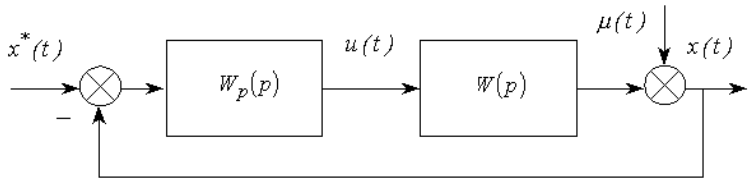


Figure 2: The scheme of the control system

Let us briefly describe the principles of standard controller constructing. Experience shows that the transient characteristics of the vast majority of industrial control objects are usually well approximated by a second-order dynamic (inertia) system with delay. Such system can be described by a differential equation:

$$T_2^2 x''(t) + T_1 x'(t) + x(t) = k \cdot u\,(t-\tau)\,. \tag{9}$$

Or we can describe it as follows:

$$W(p) = \frac{k \cdot e^{-\tau p}}{T_2^2 p^2 + T_1 p + 1}, \tag{10}$$

where k is the transfer factor of the controlled object, τ is a pure delay constant.

In case of a large dynamic control error, it is possible to approximate the control object by a first-order dynamic system with delay (in this case $T2 = 0$) or even simply an inertia-free system with delay ($T2 = T1 = 0$).

The approximate transition function of the optimal controller can be represented as:

$$W_p(p) = \frac{1}{\tau p W_o(p)} \tag{11}$$

where $W_o(p)$ is a transition function of the control object without delay. As a result, the transition functions of the controllers described above are as follows:

$$W_p^{\mathrm{PID}}(p) = \frac{T_1}{k\tau}\left(1 + \frac{1}{T_1 p} + \frac{T_2^2}{T_1} p\right), \tag{12}$$

$$W_p^{\mathrm{PI}}(p) = \frac{T_1}{k\tau}\left(1 + \frac{1}{T_1 p}\right), \tag{13}$$

$$W_p^{\mathrm{I}}(p) = \frac{1}{k\tau\, p}. \tag{14}$$

Such controllers are typical. They implement, respectively, PID, PI and I control laws.

We propose the following approach to adjust parameters of controllers. The coefficients k and τ can be determined experimentally. The transfer factor of the control

object k is equal to the ratio of the object output value to its input value in the steady-state operating mode. The delay constant τ can also be determined empirically. The parameters $T1$ and $T2$ are determined as a result of optimization of the following criterion:

$$W(T_1, T_2) = \frac{1}{s} \sum_{i=1}^{s} \left(x_s \left(t_i, u \right) - x_a \left(T_1, T_2, t_i, u \right) \right)^2 \to \min_{T_1, T_2}, \tag{15}$$

where $x_s(t, u)$ is a response of the nonparametric model of a linear dynamic system to the input action $u(t)$, $x_a(T_1, T_2, t, u)$ is a response of the approximating parametric model of the linear dynamic system corresponding to the selected typical controller to a similar input action.

To construct a nonparametric model of a linear dynamic system, it is necessary to conduct a series of experiments with the object directly. However, in practice, there may be a case when it is impossible to use the existing control scheme without damaging the object. In this case, having the information on the type and current parameters of the existing controller, we determine the transition function of the object. The description of the concept, which is realized in this case, is given below. Suppose that the operator of a closed system has the form:

$$\Phi \left(p \right) = \frac{W \left(p \right) \cdot W_p \left(p \right)}{1 - W \left(p \right) \cdot W_p \left(p \right)}, W \left(p \right) \cdot W_p \left(p \right) = \Phi \left(p \right) - \Phi \left(p \right) \cdot W \left(p \right) \cdot W_p \left(p \right). \tag{16}$$

Applying the inverse Laplace transform to both sides of the equality, we obtain the following integral equation:

$$h_2 \left(t \right) = h_\Phi \left(t \right) - \int_0^t h_\Phi \left(\tau \right) h_2 \left(t - \tau \right) \, d\tau \tag{17}$$

where $h \left(t \right) = L^{-1} \left\{ W \left(p \right) \right\}$, $h_p \left(t \right) = L^{-1} \left\{ W_p \left(p \right) \right\}$, $h_\Phi \left(t \right) = L^{-1} \left\{ \Phi \left(p \right) \right\}$ are impulse transition functions of the nodes of the control scheme and the closed system respectively,

$$h_2 \left(t \right) = L^{-1} \left\{ W \left(p \right) \cdot W_p \left(p \right) \right\} = \int_0^t h \left(\tau \right) h_p \left(t - \tau \right) \, d\tau \tag{18}$$

where $h_p(t)$ is determined by the type and current adjustment of the controller, and is assumed to be known; the estimation of $h_\Phi(t)$ is also known. Thus, replacing the impulse transition functions with their nonparametric estimates, we consequentially search for the numerical solution of the integral equations: equations (17) regarding $h_2(t)$ at the first stage and equation (18) regarding $h(t)$ at the second stage. As a result, we obtain the required estimation of the impulse transition function of the control object, which is used to calculate the criterion function (15). The derivation of the numerical algorithm for solving the problem is similar to the derivation of realization of the transition function for the linear dynamic system's inverse operator:

$$h_2(t) = \frac{h_\Phi(t) - \Delta\tau \sum\limits_{\tau=0}^{\frac{t-\Delta\tau}{\Delta\tau}} h_\Phi(t - \tau\Delta\tau)h_2(\tau\Delta\tau)}{1 + h_\Phi(0)\Delta\tau}, h_2(0) = 0, \tag{19}$$

$$h(t) = \frac{h_2(t) - \Delta\tau \sum\limits_{\tau=0}^{\frac{t-\Delta\tau}{\Delta\tau}} h_p(t - \tau\Delta\tau)h(\tau\Delta\tau)}{h_p(0)\Delta\tau}, g(0) = 0. \tag{20}$$

An important issue that arises during such adjusting of the typical controller parameters is the stability of the system. The loss of stability is not generally observed when we determine the parameters of adjustment from the condition of the minimum of the root-mean-square control error.

However, transient processes in such system often have the form of damped oscillations, the damping intensity of which, as a rule, is insufficient. Therefore, it is necessary to introduce additional restrictions to the criterion of the optimal control system functioning. These restrictions allow to influence the transient processes that arise in the control system.

Thus, optimization of the criterion function (17) with respect to parameters T_1 and T_2 is carried out taking into account the constraints determined from the stability conditions. To isolate the stability region, one can use, for example, the D-decomposition method.

Conclusions

In this article, the authors propose a solution for synthesis of a double loop control scheme with the use of a nonparametric controller of linear dynamical systems as a second loop. The proposed algorithm is based on a nonparametric model of dynamics. The double loop control scheme implies the preservation of the analog devices of local automation, which makes it reliable even when the digital control loop fails, which together with the detailed description of the algorithms for adjustment of the typical analog devices of local automation using a nonparametric model of a linear dynamic system allow the presented approach to solve efficiently the problems mentioned in the article and can be widely used for the development of complex control systems for technological processes in power engineering and various branches of industry.

Acknowledgements

The authors are grateful to Prof. Dr Medvedev Alexandr Vasilievich for his support and encouragement during this scientific research.

References

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

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

[3] Pupkov A.N. (2013). Synthesis and research of non-parametric multi-channel controller linear dynamical systems. *PhD dissertation. Krasnoyarsk State Technical University*, Krasnoyarsk. pp. 132.

[4] Bannikova A., Korneeva A., Kornet M. (2015). About the dual non-parametric control of dynamic systems. *Applied methods of statistical analysis. Nonparametric approach - AMSA'2015*, Novosibirsk, pp. 30-37.

About the Control of a Group of Objects on the Example of Steam Pressure in the CHP Main Line

Nadezhda V. Koplyarova[1], Anatoly V. Chubarov[1], Natalia A. Sergeeva[2]
[1] *Siberian Federal university, Krasnoyarsk, Russia*
[2] *Rd-science, Krasnoyarsk, Russia*
e-mail: koplyarovanv@mail.ru, avchem81@yandex.ru,
n.sergeeva@rd-science.com

Abstract

In the work it is proposed to use an external control device as the main regulator, which is going to control the task for the boiler heat load regulators in the steam pressure control mode, based on the analysis of the operation data of the combined heat and power plant (CHP). In addition, this control device is adaptive, since during the operation of the CHP, the data composing its training sample are accumulated simultaneously with the control.

Keywords: coal thermal power station, combined heat and power plant, steam line, main regulator.

Introduction

The basis of the coal thermal power stations (Combined Heat and Power Plant - CHP) work is to maintain a balance between the generated steam boilers and the power consumed by turbo-blowers and turbo-generators. If this balance is observed, the main focus is on supporting the specified pressure in the main steam line, the deviation from which indicates a violation in the balance of capacity. Accordingly, the main vapor pressure regulator in the main steam line is the main link in the process control system of the CHP during its operation, ensuring proper functioning of this priority. The physical part the desired pressure maintaining is achieved by changing the amount of fuel supplied to the boiler furnaces. For this replies the heat load regulators of each boiler separately, operating in the base mode (maintenance of the set steam flow rate) or peak-regulating (maintenance of the set steam pressure in the main steam line). The mode is determined by the reference from the main controller [1].

To maintain the steam pressure in the steam pipe, a control system is usually used when one of the boilers is operating in a regulating mode and, according to the pressure sensors in the main line, the CHP changes the pressure in the pipeline by changing the capacity. The remaining boilers are in the basic mode, they support the specified steam loads. The disadvantages of this regulator are the limitation of the range of steam pressure regulation by the boiler power in the regulating mode. If we include more boilers for control, there is an effect of mutual influence of boilers with different characteristics, some of which are loaded to maximum, others are loaded to a minimum. With such a scheme, control becomes impossible.

The regulator is designed to maintain the steam pressure in the main steam pipe by changing the boiler output. This regulator is corrective, it works as a PI controller and changes the reference to the heat load controller, if the boiler is in the regulating mode. The main regulator receives a signal on the vapor pressure in the collector and the signal of the setpoint. In this case, the main controller operates according to the PI-regulation principle, correcting the task for the heat load regulator of the regulating boiler (or several, if any).

In this connection, the task is to manage a group of objects, which are the boiler units of the CHP. The paper proposes a control approach similar to that existing at the CHP, but with the additional external control loop. The operation of the external circuit is based on the machine learning. It analyzes the archive data of measurements of the boiler variables, which are not taken into account in the PI regulator, due to which the task for the heat load regulator is issued more quickly.

1 Problem formulation

A system of N boilers of CHP included in a common steam main is considered as a research object. The figure (1) shows the general scheme for the inclusion of N boilers in one steam main [1].

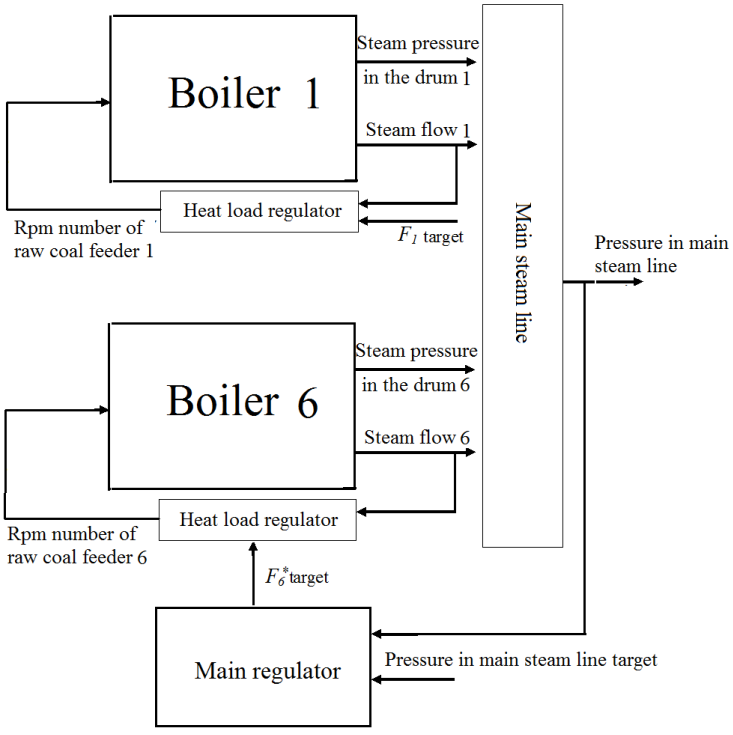


Figure 1: The scheme of N boilers in the CHP

Where the object (boilers or steam generators) structure and parameters are unknown, H are the measuring devices. In the paper we consider the scheme with the

one boiler in a regulating mode. The boiler structures are the same. The boiler has a set of regulators (the level in the boiler drum, vacuum, etc.), which have the task of maintaining or controlling certain variables. We should note, that boiler has the heat load controller (PI controller), that controls its steam capacity.

The disadvantages of the existing control scheme are following:

- The control range is insufficient.

- False operation of the heat load controller in the base mode with perturbations from the main steam line side.

- Changes of the object parameters when changing the composition of operating equipment.

It is required to design the studying process model (N boilers in the steam main) and to propose the steam main pressure controller.

2 Object imitation

To create the object imitation its parameters were estimated with using the real transient process data from one of the Krasnoyarsk coal thermal power station. As a result we obtain the mathematical parametric model that imitates the coal thermal power station processes. The scheme of the model is presented it the figure (2):

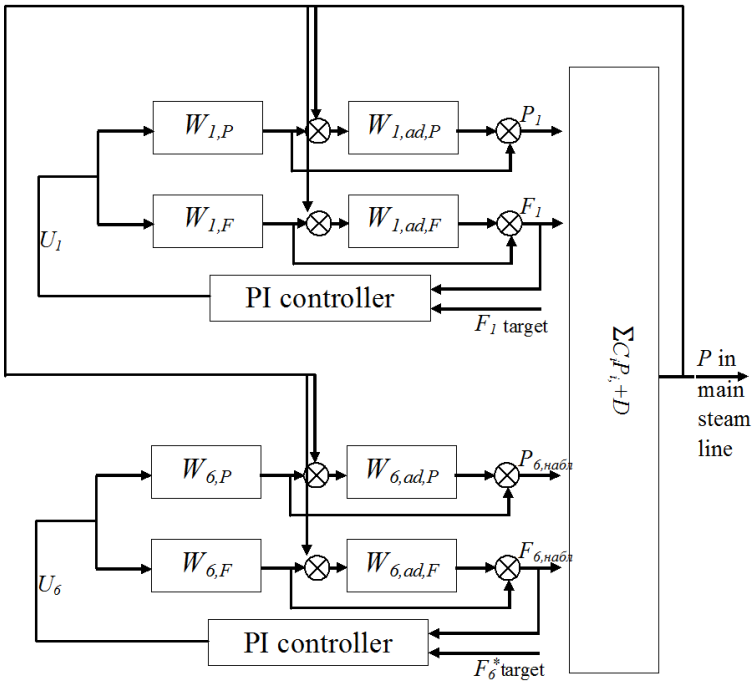


Figure 2: The general structure of the coal thermal power station simulation model

The input variables (rpm number of raw coal feeder) and the system output (the steam pressure and flow) correlation is described with the transfer functions $W_{i,p}$, $W_{i,f}$ with a different coefficients $a_{p,i}$, $b_{p,i}$, $a_{f,i}$, $b_{f,i}$ (for each i boiler):

$$W_{i,p} = \frac{b_{p,i}e^{-st}}{a_{1p,i}s + a_{0p,i}}. \tag{1}$$

$$W_{i,f} = \frac{b_{f,i}e^{-st}}{a_{1f,i}s + a_{0f,i}}. \tag{2}$$

where P is the pressure, F is the steam flow rate, U is a rpm number of raw coal feeder. We also estimate the structure and parameters of the transfer functions W_{ad}, describing the influence of the pressure in the main line - the effect of the pressure difference in the drum and the pressure in the line with some displacement and coefficient. PI controllers on the scheme controls the values of the observed steam flow by changing the input influences (rpm number of raw coal feeder). The figure (3) shows an example of comparing model and object data to confirm the adequacy of the model (using the example of two boilers - 3 and 4):

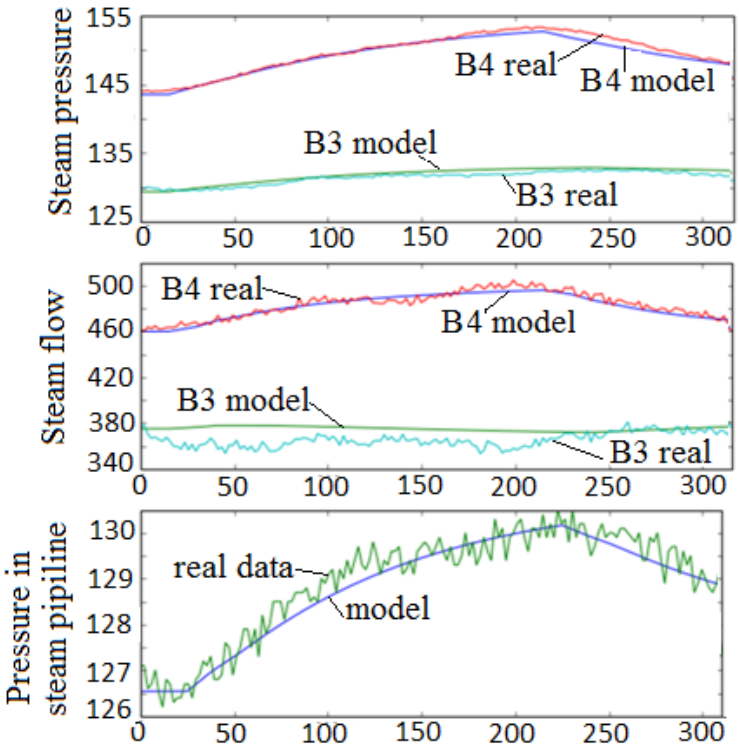


Figure 3: The imitation and real data comparsion

The relative average error of modeling the steam pressure in the drum is $E_{p4} = 0.43\%$ for 4 and $E_{p3} = 1.57\%$ for 3 boilers respectively, the error of the steam flow modeling $E_{f4} = 0.64\%$ for 4 and $E_{f3} = 3.4\%$ for 3 boilers respectively, the error E_{pg} of the steam pressure in the main steam line is 3.61%. The following limits of the variable permissible values for the simulation object are set:

- the steam pressure in the drum$(p1 - p6)$ is in range 133-137 kgf/cm^2;

- the steam pressure in the manifold (main line) (p): 110-130 kgf/cm^2;
- the steam flow from boiler ($f1 - f6$): 350-500 t/h.
- the coal consumption ($u1 - u6$): 110-130 a rpm number of raw coal feeder.

Some computational studies of the imitation object are carried out in the paper.

3 Test experiments on the object simulation

The task of computational experiments on the CHP model is to simulate various operating conditions of boilers for accumulating a test sample (archival data, which then are used for control). As capacity of boilers we will consider the maximum possible value of steam and fuel consumption (a rpm number of raw coal feeder). The scheme of the computation experiments for the test sample data accumulation is the following:

- Operation of the boilers in the basic mode: for each of the N boilers, we set the same target for the steam flow rate of 90% of the maximum possible ($500t$). Then the pressure in the pipeline is $130kgf/cm^2$.

- Changing the load of one boiler with the stable operation of the rest: set the boilers to 1-5 load, which is 90% of the maximum possible. This corresponds to a steam flow rate of $470t/h$. At the same time, we change the load on one of the boilers (the sixth one).

- Adding some variable - simulating the humidity (or quality) of coal. It affects the rpm number of raw coal feeder required to achieve a given boiler output.

The result of the experiments is confirmation of the adequacy of the reaction of the CHP model to various input disturbances, as well as the accumulation of information simulating archive data on the object.

4 Test experiments on the object simulation

The aim of the control in this formulation is to achieve a certain (given) pressure value in the main steam line.

It is proposed to test the possibility of using an adaptive external control loop, based on the analysis of the archive data of the boiler.

We propose to replace the main regulator to the internal control loop which forms the heat load controller targets for one or some boilers. It can form target on the base of the analysis of the real archive data from the coal thermal power station.

The task is to create a control device that allows achieving a given pressure in the pipeline by changing the regulator target of one of the boilers. In this case, the remaining boilers must work in a base mode (without changing the task). All information about the object functioning at each time is fed to the additional external control loop.

The figure (4) shows the scheme of switching on the external control device (replacing the main regulator)[2].

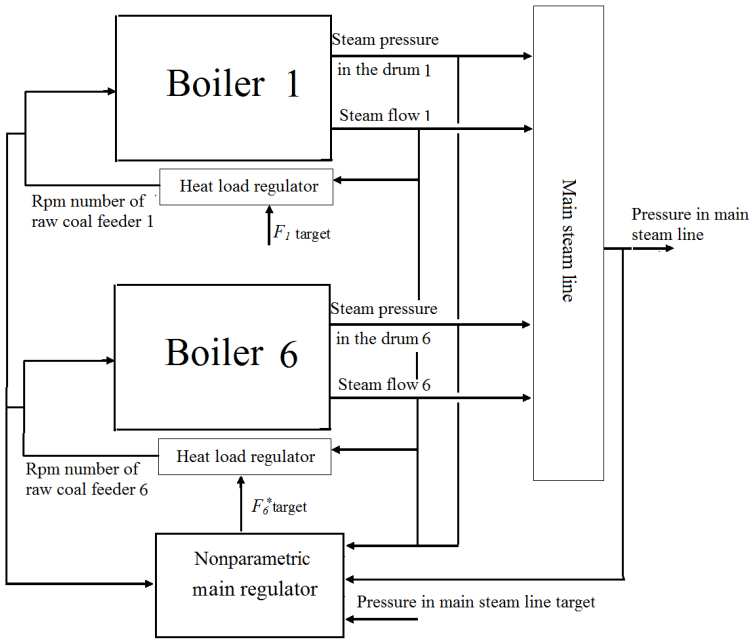


Figure 4: The scheme of the external control device including

At each step i, the researcher is provided with the following information: measuring the values of the input variables (rpm number of raw coal feeder) for each of the boilers $u_{1,i} - u_{6i}$, outputs - steam flow $f_{1,i} - f_{6i}$, steam pressure in the drum for each boiler $p_{1,i} - p_{6i}$, and the pressure value in the main steam line p_i.

In addition, there is accumulated training sample of archival data on the CHP functioning: $\{u_{1,i} - u_{6,i}, f_{1,i} - f_{6,i}, p_{1,i} - p_{6,i}, p_i\}$, $i = 1, s$ data on the object functioning.

Based on the available information, a non-parametric device can be generate to control the setpoint value for one of the boilers (for example, the Nth boiler). We can find the value of the setpoint for the heat load controller N using a nonparametric estimator as follows[2]:

$$f^*_{N,i+1} = \frac{\sum\limits_{j=1}^{i} f_{N,j}\, \Phi\left(\frac{p^*_i - p_j}{c_{s,p}}\right) \prod\limits_{k=1}^{N} \Phi\left(\frac{p_{k,i} - p_{k,j}}{c_{s,p,k}}\right) \prod\limits_{k=1}^{N-1} \Phi\left(\frac{f_{k,i} - f_{k,j}}{c_{s,f,k}}\right) \prod\limits_{k=1}^{N-1} \Phi\left(\frac{u_{k,i} - u_{k,j}}{c_{s,u,k}}\right)}{\sum\limits_{j=1}^{i} \Phi\left(\frac{p^*_i - p_j}{c_{s,p}}\right) \prod\limits_{k=1}^{N} \Phi\left(\frac{p_{k,i} - p_{k,j}}{c_{s,p,k}}\right) \prod\limits_{k=1}^{N-1} \Phi\left(\frac{f_{k,i} - f_{k,j}}{c_{s,f,k}}\right) \prod\limits_{k=1}^{N-1} \Phi\left(\frac{u_{k,i} - u_{k,j}}{c_{s,u,k}}\right)}. \quad (3)$$

where $\Phi(z)$ is a Kernel function satisfied the following conditions[3, 4]:

$$\Phi(z) < \infty, \forall z \in \Omega(z); \int\limits_{\Omega(z)} \Phi(z)dz = 1; \int\limits_{\Omega(z)} \Phi^2(z)dz < \infty; \lim_{s\to\infty} \frac{1}{c_s} \Phi\left(\frac{u}{c_s}\right) = \sigma(u). \quad (4)$$

c_s is a bandwich parameter. It is a certain constant number, the magnitude of which determines the degree of "blurring" of the Kernel function in the predicted

point vicinity, and, accordingly, the degree of smoothness of the estimate obtained. The parameter satisfies the following requirements:

$$c_s > 0, s = 1, 2, ...; \lim_{s\to\infty} c_s = 0; \lim_{s\to\infty} sc_s = \infty; . \tag{5}$$

The choice of the optimal bandwich parameter is carried out on the basis of the condition of the mean-square criterion minimum (the error of the system output observed value and its nonparametric estimation).

At the same time, the principle of non-parametric control devise operation is based on the analysis of archival data and will not be able to offer a control value that is not in the data. In this connection, it is proposed to use a dual control scheme, that is, to include the learning additive. Then the control device algorithm takes the form:

$$\tilde{f}^*_{N,i+1} = f^*_{N,i+1} + k(p_i - p^*_i). \tag{6}$$

With the use of such an algorithm, the object is studied in the course of its control. In addition, a similar circuit can be used to control several boilers in the steam pressure control mode in the steam line. In this case, the formation of tasks for each boiler heat load controller takes place sequentially, taking into account the tasks for the other boilers. We compare the results of steam pressure control in the main line with the use of the PI-regulator and the proposed controller that gives the task for the 6-boiler heat load controller. It is required to bring the system to the setpoint state when changing the target for steam pressure in the pipeline, by controlling the heat load controller of the sixth boiler. The remaining five boilers of the CHP plant are in the base mode. The result is shown in the figure (5).

As we can seen from the figure, the external control device sets the steam flow value of the sixth boiler fast enough to reach the required pressure in the main line, which reduces the time for the transition of a group of objects (six boilers of a CHP plant) to a specified state. It is assumed that there is a training sample describing the various states of the object.

It should also be noted that in this case, the efficiency of the control with the use of the PI controller depends on when and how accurately the controller is set. Since over time some characteristics of the object may change, which will require the setting of the PI controller parameters.

Conclusions

The introduction of an intelligent adaptive control system at the plant allows correcting the target for the heat load regulator of boilers in order to achieve the setpoint (steam pressure in the pipeline). Thus, computational experiments have shown that the use of adaptive non-parametric regulators allows using the capabilities of the main regulator, and eliminating some of the disadvantages of its use existing at the plant. The use of such a circuit allows to accurately control the steam pressure in the main line since it is adaptive. In addition, it gives the task to the heat load controller faster than the PI-regulator.

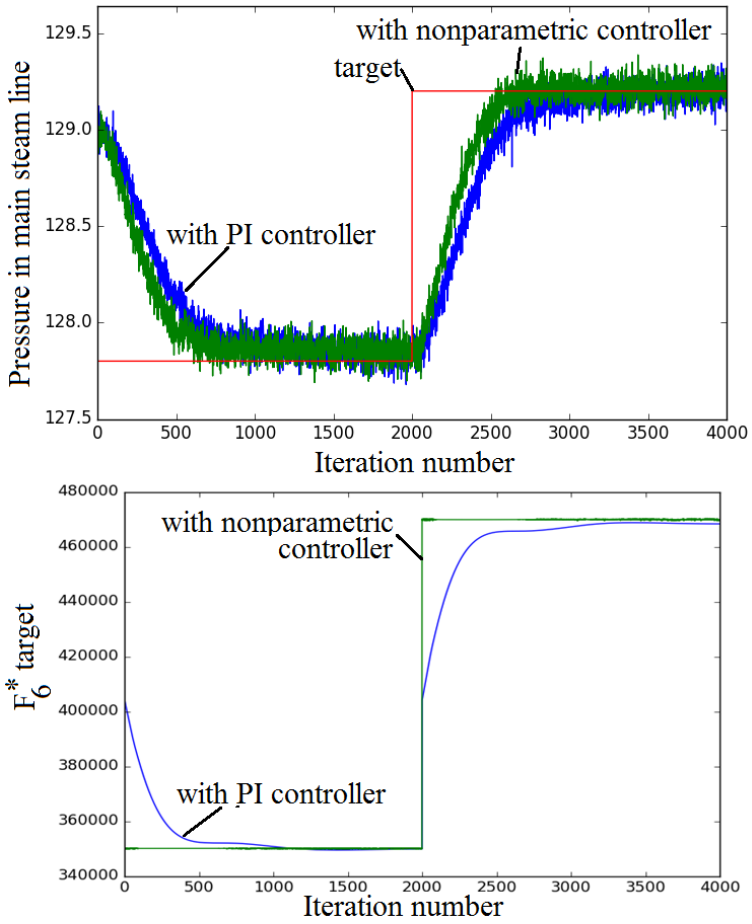


Figure 5: Example of steam pressure control in a line with the use of external controller or PI-regulator

References

[1] Zhalnin D.A., Shorokhov V.A., Evdokimov A.N., Bubnovskii O.A., Churinov A.V.(2010) Experience with the introduction of a pressure regulator system in the main steam line at the Krasnoyarsk TÉTs-2 plant. *Power Technology and Engineering*, Vol. **44(1)**, pp. 52-59.

[2] Medvedev A.V. (1975). Identification and control for linear dynamic System of unknown order *Optimization techniques IFIP Technical Conference*. Berlin-Heidelberg-New-York: Springer-Verlag. pp. 48-56.

[3] Nadaraya E.A. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability*. Vol. **10**, pp. 186–190.

[4] Dobrovidov A.V., Koshkin G.M. (2011). Regularized data-based nonparametric filtration of stochastic signals. *Proceedings of the World Congress on Engineering*. pp. 56–70.

On the Adaptive Control of Group of the Technical Processes under Incomplete Information

Kornet M.E., Raskin A.V., Raskina A.V.
Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia, Krasnoyarsk, Russia
e-mail: araskin654@gmail.com

Abstract

The problem of control the group of technological processes in condition of incomplete information is considered. The group of processes is the series-connected elements of the technological chain, and include the different types of objects for example dynamic or static objects with delay. In this case, the parametric structure up to parameters is unknown. This situation often arises because many dynamic processes are not deeply studied. The factor of unknown distribution random noises causes the complexity of solving the control tasks. The problem of control are investigated in a closed loop, including a system of object - regulator. Consideration of the object in this form allows saving the typical controls, adding an external control loop in case of the technical processes control. In the article details are non-parametric control algorithms for the external loop is provided.

Keywords: adaptive control, nonparametric theory, the sequence of technological objects, dynamic object, the instantaneous object with delay.

Introduction

The task of control of complex multidimensional processes, the technological chain of which can be solved by various production schemes, such as parallel, sequential or their combinations at the moment is relevant for many industries. In this case, the control scheme will contain more than one local object, but a group of objects connected together. It should be borne in mind that when managing a chain of technological impacts, their general sequence should be constructed in such a way that the algorithms for managing local objects are consistent among themselves. To date, this is a set of algorithms for adaptive control of parametric type [11], [2]. Their main feature is knowledge about the parametric structure of the model to within parameters based on a priori information or preliminary studies. In the conditions of a lack of a priori information, one of the ways to solve this problem is to introduce into the control system a non-parametric external circuit that will generate an impact statement for each local regulator. This allows you to better control of the entire sequence. The paper presents adaptive nonparametric algorithms for determining the control actions for an external control loop.

1 Formulation of the problem

Consider the next block-scheme (Fig. 1).

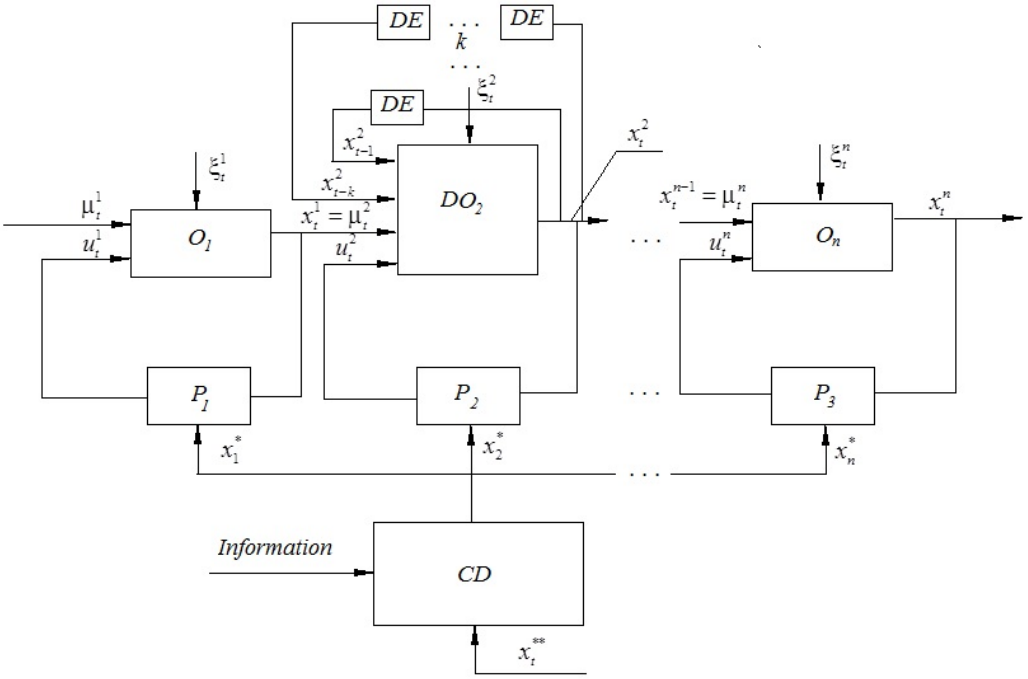


Figure 1: Diagram of a two-circuit sequence control of technological processes

In Figure 1, the following symbols are used: $u_t^1, ..., u_t^n$ - is input controllable actions, $\mu_t^1, ..., \mu_t^n$ - input controlled but not controllable actions, $x_t^1, ..., x_t^n$ - output actions, $O_1, ..., O_n$ - is local inertia-free objects with delay, DO_2 - dynamic object, DE - delay element, $x_{t-1}^2, ..., x_{t-k}^2$ - is output effects of the dynamic object, delayed by the corresponding number of steps, P_1, P_2, P_3 - are a typical regulators as P, PI, PID, $x_1^*, ..., x_n^*$ - is the setting actions for local standard regulators,CD -is control devices of external circuit, Information - are all available measurements of input-output variables of local processes $\{x_i^j, u_i^j, \mu_i^j, i = \overline{1,s}, j = \overline{1,n}\}$, x_t^{**} - is a determining influence on the external control loop, $\xi_t^1, ..., \xi_t^n$ - external noise acting on local processes. In general, all the variables described are vectors.

A non-inertial object with a delay can be represented in the following form:

$$x_i^t = f(u_i^{t-\tau}, \mu_i^{t-\tau} = x_i^{t-\tau}, \xi_i^{t-\tau}), i = \overline{1, n-1}, \tag{1}$$

where f - is an unknown functional, τ - is a delay that can differ by different communication channels, but for reasons of simplicity in the text we have adopted the unified designation of the delay τ. It should be noted that the output variable of the object that affects the next object is essentially an unmanaged input variable.

In the case of a dynamic object, the state of the system at a given time depends on both the input effects and its states in the past $x^{t-1}, ..., x^{t-k}$. In other words, the dynamic system can be described by the following equation:

$$x_i^t = f(x_i^{t-1}, x_i^{t-2}, ..., x_i^{t-k}, u_i^t, \mu_i^t = x_i^t, \xi_i^t), i = \overline{1, n-1} \tag{2}$$

where x_t - is the output variable of the process, u_t - control action, k is the "depth" of the dynamic object memory (in the terminology of AA Feldbaum) [3]. If we draw an analogy with the description of the process under study in continuous time in the form of differential equations, then k is the order of the highest derivative in the corresponding equation. In both cases, it is essential that the form of the functional f is not determined up to parameters.

The control problem is reduced to the development of adaptive nonparametric control algorithms for the outer contour. The proposed control device will generate control actions for local typical controllers, so that the final control actions for the local objects are matched to each other.

2 Nonparametric control algorithm

The following non-parametric estimation of the regression function (Nadaraya-Watson [6]) can be taken as the control action for the regulator of the local instantaneous object O_j with delay from observations $\{x_i, u_i, \mu_i, i = \overline{1,s}$ in a discrete form [3]:

$$x_j^* = \frac{\sum_{i=1}^{s} x_t^j \Phi\left(\frac{x_t^{j-1} - x_i^{j-1}}{c_s^x}\right) \Phi\left(\frac{u_t^j - u_i^j}{c_s^u}\right)}{\sum_{i=1}^{s} \Phi\left(\frac{x_t^{j-1} - x_i^{j-1}}{c_s^x}\right) \Phi\left(\frac{u_t^j - u_i^j}{c_s^u}\right)}, j = \overline{1,n}, \quad (3)$$

where Φ - is the nuclear bell-shaped function, c_s - is the kernel blur coefficient corresponding to each object variable, s - the sample size of the observations. The nuclear function and the kernel blur factor satisfy certain convergence conditions [6].

In the case of a dynamic object, the driving effect is:

$$x_j^* = \frac{\sum_{i=1}^{s} x_t^j \Phi\left(\frac{x_t^{j-1} - x_i^{j-1}}{c_s^x}\right) \Phi\left(\frac{u_t^j - u_i^j}{c_s^u}\right) \Phi\left(\frac{x_{t-1}^j - x_{i-1}^j}{c_s^x}\right) ... \Phi\left(\frac{x_{t-k}^j - x_{i-k}^j}{c_s^x}\right)}{\sum_{i=1}^{s} \Phi\left(\frac{x_t^{j-1} - x_i^{j-1}}{c_s^x}\right) \Phi\left(\frac{u_t^j - u_i^j}{c_s^u}\right) \Phi\left(\frac{x_{t-1}^j - x_{i-1}^j}{c_s^x}\right) ... \Phi\left(\frac{x_{t-k}^j - x_{i-k}^j}{c_s^x}\right)}, j = \overline{1,n}, \quad (4)$$

The blur parameter c_s is determined by solving the problem of minimizing the quadratic exponent of the object's output matching and the driving influence based on the "sliding exam method".

Conclusions

The article considers the task of control of a group of objects in the variant of sequentially distributed elements of the sequence. Nonparametric algorithms of adaptive control are presented. In this case, the use of typical control algorithms without an external control loop can lead to significant errors in the regulation. As a result, when controlling the sequence of processes, it is first necessary to determine the appropriate control actions for the output variables, and then apply the typical control algorithms. This is an essential feature of the construction of the control system, in contrast to the traditional control algorithms.

References

[1] Tsypkin Ja. Z.(1968). *Adaptatsiya i obuchenie v avtomaticheskikh sistemakh [Adaptation and learning in automatic systems]*. Nauka, Moskow.

[2] Eykhoff P.(1975). *Osnovy identifikatsii sistem upravleniya [Identity-based control systems]*. Mir, Moskow.

[3] Fel'dbaum A.A. (1963). *Osnovy teorii optimal'nyh avtomaticheskih system (Fundamentals of the theory of optimal automatic systems)*. Fizmatgiz, Moskow.

[4] Medvedev A.V. (2015). *Osnovy teorii adaptivnyh system (Basic theory of adaptive systems)*. SibGAU, Krasnojarsk.

[5] Nadaraya E. A.(1983). *Neparametricheskie otsenki plotnosti veroyatnosti i krivoy regressii (Non-parametric estimation of the probability density and the regression curve)*. Tbil. Publ, Tbilisi.

Decision Trees Control of Static System under Incomplete Information

Ekaterina Mangalova[1], Olesya Chubarova[2] and Denis Zhalnin[3]
[1] *Siberian State Aerospace University, Krasnoyarsk, Russia*
[2] *Siberian Federal University, Krasnoyarsk, Russia*
[3] *LLC Region-Avtomatika, Krasnoyarsk, Russia*
e-mail: e.s.mangalova@hotmail.com

Abstract

In this paper control problem of static system under incomplete information is discussed. Nonparametric algorithm of control based on decision tree is proposed and tested on real-world data from Thermal Power Station.

Keywords: Decision tree, control, nonparametric estimation.

Introduction

Algorithms of classical control theory belong to the class of parametric algorithms and require assumptions about the object structure (structure of equation describing the object) for solving identification and control tasks. There are cases when we do not have a prior information to choose the parametric structure. In [1] control algorithm based on the Nadaraya-Watson estimator [2] and their sequence was proposed.

However, there are major problems with this approach to multidimensional task. On the one hand, it is connected with observations distribution in high dimensional feature space (especially in case of small number of observations). Suppose now one had $n = 1000$ points uniformly distributed over the ten dimensional unit cube $[0, 1]^{10}$. An average over a neighborhood of diameter 0.25 (in each coordinate) results in a volume of $0.25^{10} \approx 0.00000095$ for the corresponding ten-dimensional cube. Hence, the expected number of observations in this cube will be 0.00095 and any averaging can not be expected. If we fix the count $k = 1$ of observations over which to average, the diameter of the typical neighborhood will be larger than 0.5. It means that the average is calculated over at least one-half of the range along each coordinate [3]. On the other hand, the Nadaraya-Watson estimator has the high computational complexity. The bigger feature space dimension the harder to optimize vector of bandwidths. Decision trees are used for solving such regression tasks [1].

In this paper approach based on the decision trees is proposed to solution of control task.

The paper is organized as follows. In the next section we introduce the statement of the control problem. In the second section we overview the algorithm for Classification and Regression Tree (CaRT) construction. In the third section we propose control algorithm based on Regression Tree and its variants for different control task statements. The fourth part is devoted to algorithm testing. The paper finishes with a conclusion and perspectives for future work.

1 Statement of the control problem

The block scheme of the control process is shown in Figure 1. The following designations are taken: x is vector of output variable, x^* is vector of the desired output x values, u is vector of controlled inputs, μ is vector of observed uncontrolled inputs, ξ is vector of unobserved inputs.

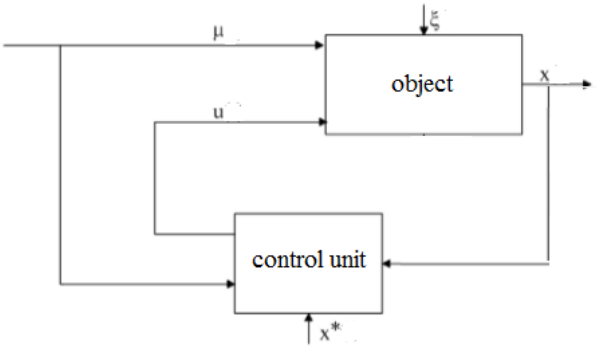


Figure 1: The overall control scheme. x is vector of output variable, x^* is vector of the desired output x values, u is vector of controlled inputs, μ is vector of observed uncontrolled inputs, ξ is vector of unobserved inputs

It is evident from Figure 1 that output object variables x depend on input u, μ, ξ. The control task is to build control unit which generates u such that $E(x, x^*)$ is minimized, where E is some error measurement (MAE, MSE, etc.).

2 Decision tree model

Decision tree is piecewise constant nonparametric model. The most popular decision tree model is Classification and Regression Tree (CaRT) [1]. CaRT is a binary tree where each root node represents a input variable u^{j_r} and a split point b_r. Each leaf node contains values of output variable x which is used to make a prediction.

CaRT fitting involves selecting input variables and split points on those variables until a suitable tree is constructed.

Input variable and split point are chosen using a greedy algorithm to minimize a error function:

$$\min_j \min_b \left(\sum_{i:u_i^j<b} L\left(x_i, \hat{x}^-(\bar{u}_i, j, b)\right) + \sum_{i:u_i^j \geqslant b} L\left(x_i, \hat{x}^+(\bar{u}_i, j, b)\right) \right) \tag{1}$$

where $\hat{x}^-(\bar{u}_i, j, b)$ and $\hat{x}^+(\bar{u}_i, j, b)$ are prediction in point $\bar{u}_i$ based on subspaces after splitting:

$$\hat{x}^-(\bar{u}_i, j, b) = \frac{\sum_{i:u_i^j<b} x_i}{\sum_{i:u_i^j<b} 1}, \tag{2}$$

$$\hat{x}^{+}\left(\bar{u}_i, j, b\right) = \frac{\sum_{i:u_i^j \geqslant b} x_i}{\sum_{i:u_i^j \geqslant b} 1}, \tag{3}$$

where the sum in the denominator is number of points in the considered subspace.

Tree construction ends using a predefined stopping criterion, such as a minimum number of observations assigned to each leaf node of the tree or a maximum depth of the tree.

The splitting procedure is represented by a binary tree shown in Figure 2.

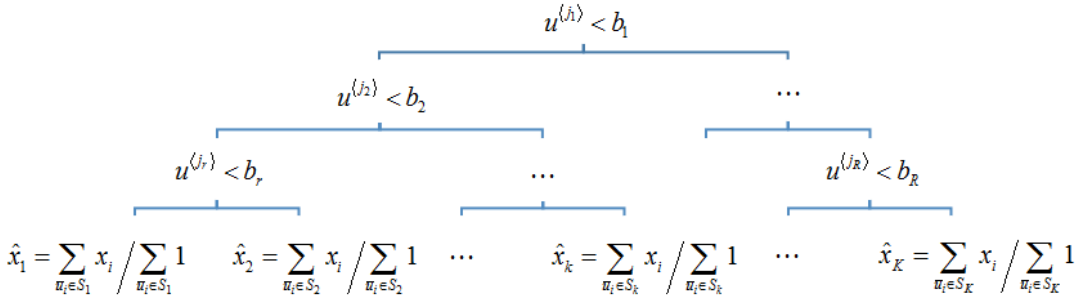


Figure 2: An example of regression tree with K terminal nodes. The predicted value is denoted here as $\hat{x}_k, k = 1, 2, ..., K$

Each node in the tree corresponds to a rectangular region of the predictor space S^k, a subset of the observations lying in the region S^k, a constant $\hat{x}_k$ which is the average response of the observations in k-th rectangular region. Thus, the binary tree model can be formalized as following:

$$\hat{x}\left(\bar{u}\right) = \left\{\hat{x}_k : \bar{u} \in S_k, k = 1, 2, ..., K\right\}. \tag{4}$$

3 Decision tree control algorithm

Control tasks can be solved using the regression trees in different formulations depending on the number of input and output values, presence of observed uncontrolled input variables. Consider the basic formulation of the problem.

3.1 One-dimensional output, controlled inputs

Let there are controlled variables $\bar{u}$ and one output variable x.

Decision tree is fit using training dataset contained simultaneous observations of u and x.

Control algorithm:

1. Set control target x^*.

2. Search such leaf node that

$$k^* = \arg\min_k D^1\left(\hat{x}_k, x^*\right) \tag{5}$$

where D^1 is distance measurement. Let call k^*-th node "target node".

3. Starting from target node we go up to the root of the tree. In each node on the route we cut the subspace according to spilling rule in the current node to get a rectangular region of the predictor space corresponded to target node (Figure 2).

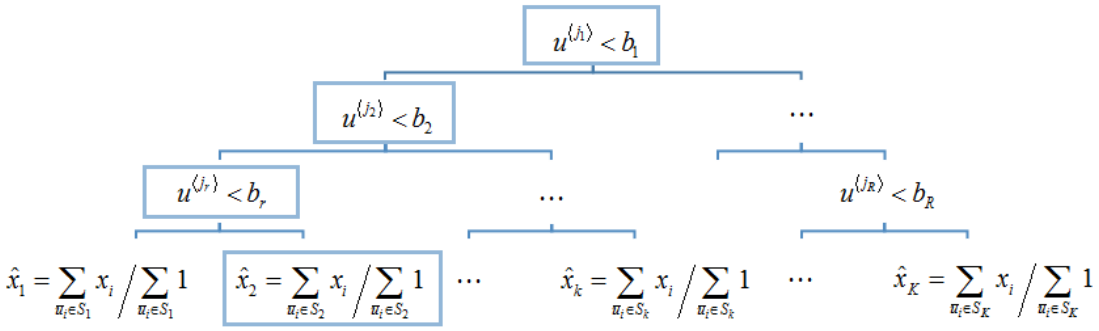


Figure 3: An example of regression tree with K terminal nodes. The predicted value is denoted here as $\hat{x}_k, k = 1, 2, ..., K$. Let target value x^* and distance measurment D^1 are such that $\hat{x}_2$ is the target node. The desired control lies in rectangle $u^{j_r} > b_r \bigcap u^{j_2} < b_2 \bigcap u^{j_1} < b_1$

3.2 One-dimensional output, controlled and observed uncontrolled inputs

Let there are controlled and observed uncontrolled variables $\bar{u}$ and one output variable x. Here we aggregate controlled and observed uncontrolled to make description easier.

Decision tree is fit using training dataset contained simultaneous observations of u and x.

Control algorithm:

1. Set control target x^*.
2. Predict uncontrolled input variables.
3. Cut nodes that can not be achieved with predicted uncontrolled input variables (Figure 4).
4. Search such leaf node that

$$k^* = \arg\min_k D^1\left(\hat{x}_k, x^*\right). \tag{6}$$

5. Starting from target node we go up to the root of the tree. In each node on the route we cut the subspace according to spilling rule in the current node to get a rectangular region of the predictor space corresponded to target node.

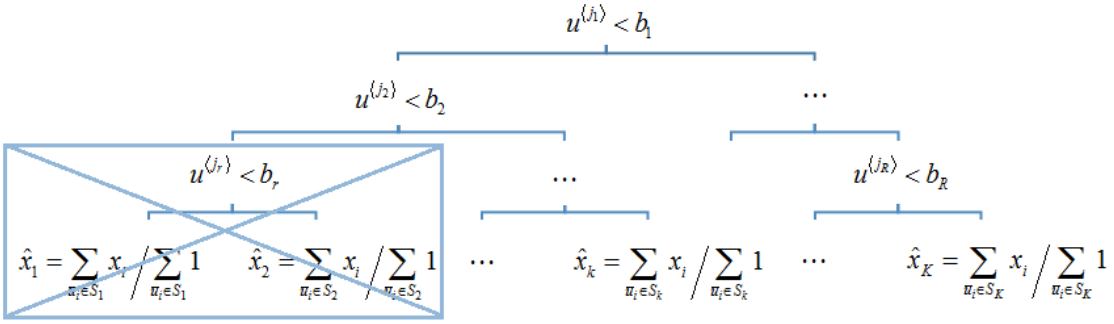


Figure 4: An example of regression tree cutting. Let u^{j_2} is uncontrolled variable and predicted value of u^{j_2} is equals to $b_2 + 1$. It means that left sub-tree from splitting $u^{j_2} < b_2$ can not be achieved with predicted

4 Experimental results

The verification of the proposed control algorithm is conducted on dataset from Thermal Power Station. Observed input variables can be divided into controlled and uncontrolled during the process. In relation to the process of obtaining super-heated steam in a boiler:

controlled variables: fuel consumption, the feed water flow rate (average), flow of secondary air to the burner,

uncontrolled variables: temperature Aero for the Mill-A No. 1 temperature Aero for the Mill-B No. 1 temperature of cold air air pressure for 1 step, air pressure for 2 step, the temperature in the flue of hot air after 2 step, the temperature in the flue of hot air after 1 step, the primary air flow to the Mill-B, C, D, the pressure of the primary air to the Mill-B, C, d, feed water pressure, feed water temperature.

Output variables characterize the quality of the final product: O2 concentration, super-heated steam pressure, the steam flow from the boiler (average), the super-heated steam temperature (average), flue gas temperature.

Dataset was split in training set using for CaRT fitting and testing set using to evaluate proposed control algorithm. Real outputs from testing set were used as target values. In Figure 5 and 6 comparison of real outputs and outputs which get using control algorithm based on CaRT is depicted.

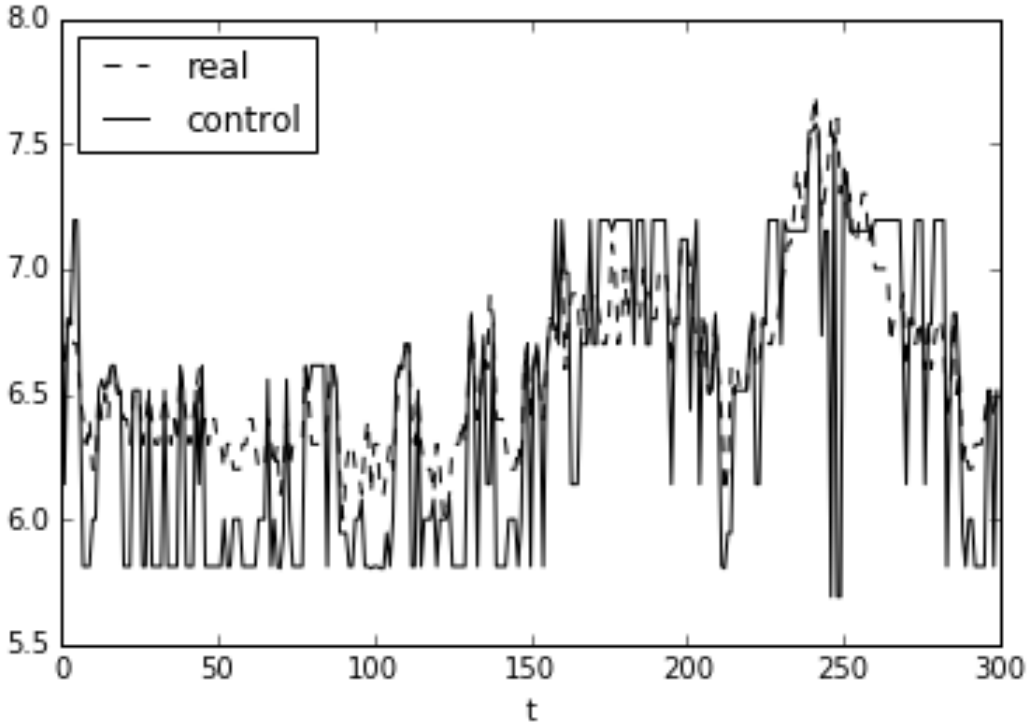


Figure 5: O2 concentration. Testing of control algorithm. Real outputs (dotted line) and outputs which get using control algorithm based on CaRT (solid line)

Conclusions

In this paper decision tree control of multidimensional static system was discussed. Control algorithm based on CaRT was proposed and its variants for different control task statements. In the future we are planning to generalize algorithm for multidimensional task and tree ensembles.

References

[1] Medvedev A.V. (2013). Theory of non-parametric systems. Control-I. *Vestnik SubSAU*. Vol. **2**, pp. 57-63.

[2] Nadaraya E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*. Vol. **9**, pp. 141-142.

[3] Härdle W. (1990). Applied nonparametric regression. Vol. **19**, Cambridge university press.

[4] Breiman L., Friedman J., Stone C. J., Olshen, R. A. (1984). Classification and regression trees. CRC press.

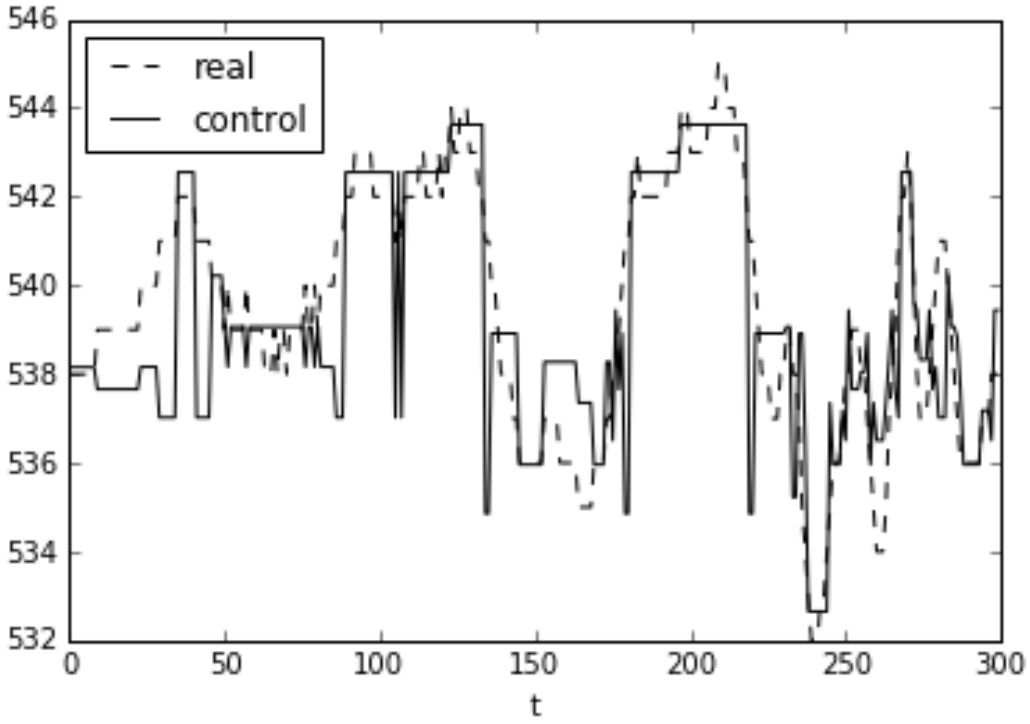


Figure 6: Flue gas temperature. Testing of control algorithm. Real outputs (dotted line) and outputs which get using control algorithm based on CaRT (solid line)

Determination of the Structure of Linear Dynamic Objects in the Condition of Incomplete Information

Anastasia V. Raskina
Siberian Federal University, Krasnoyarsk, Russia
e-mail: raskina.1012@gmail.com

Abstract

The problem of constructing the parametric structure of dynamic object is analyzed. The method of determining the structure of the dynamic differential equation with up to parameters is based on the application of the rule of allocation of significant variables for nonparametric identification. The article deals with the non-parametric model of dynamic objects. The relationship of coefficient blur kernel function and the influence of a particular variable, measured in non-parametric model output object is investigate. The algorithm of identification of the structure of the difference equation of the dynamic object includes the steps of finding the optimal coefficients blur kernel function for each variable sampling rates, elimination of the unimportant variables, modeling and simulation calculation of the relative error.

Keywords: differential equation of dynamic object, the selection of essential variables, object with memory, nonparametric identification.

Introduction

One of the main problems of the modern theory of identification and control is the definition of the structure of the model. At the present moment, most of the article is devoted to the development of parametric identification methods [11] - [2], where the structure of the process is regulated by the a priori information. Then follows the stage of setting parameters using various mathematical approaches, for example, the method of least squares, various recurrence estimates, etc. In the conditions of incompleteness of the a priori, information about the object, this approach is ineffective because of the existing variety of processes, their complexity and little knowledge. In this paper, a new method for determining a linear dynamic object based on a nonparametric identification theory is developed [11] - [2]. The basis of the proposed method lies in the rule of the non-parametric model. This idea was first proposed by Professor A.V. Medvedev in [2].

1 Formulation of the problem

Equation, and is described by the following difference equation:

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

where f - unknown functional , x_t - output variable of the process, u_t - control action, k - "depth" of the dynamic object memory (in the terminology of AA Feldbaum) [3].

If we draw an analogy with the description of the process under study in continuous time in the form of differential equations, then k is the order of the highest derivative in the corresponding equation. Here it is essential that the form of the functional is not defined up to parameters. We introduce the notation:

$$z_t = (z_1, ..., z_{k+1}) = (x_{t-1}, ..., x_{t-k}, u_t), \tag{2}$$

When identifying the dynamical system (1), its parametric model is naturally adopted in the form

$$x_s = f_s(x_{t-1}, ..., x_{t-k}, u_t, \alpha), \tag{3}$$

where α is the parameter vector to be estimated based on the training sample. Thus, in the case of a linear dynamical system, the definition of the structure of a dynamic object (1) reduces to determining the variables that make up the model (3). Taking into account the re-designations (2), the model (3) can be shown in the following scheme (Figure 1), which illustrates the model of a discrete-time dynamical system reduced to a static system model.

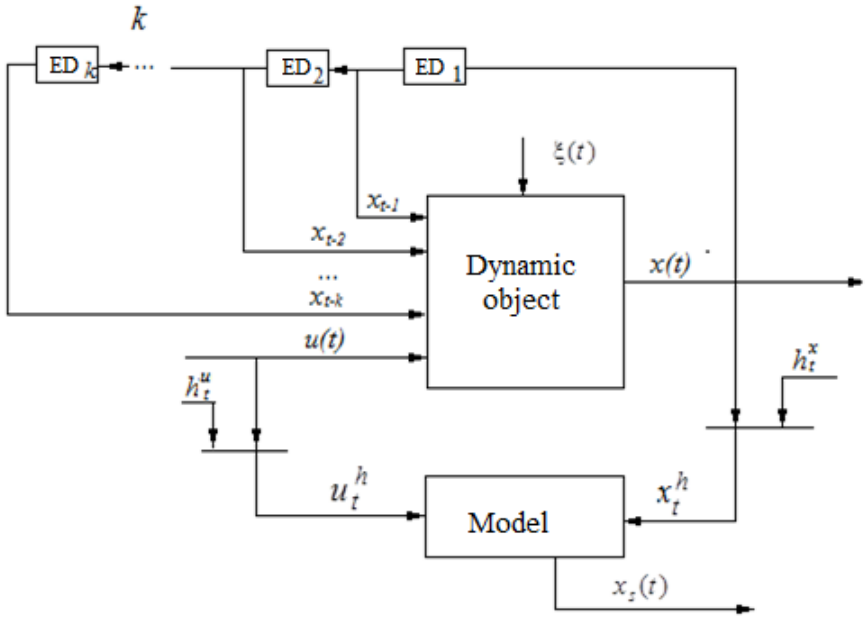


Figure 1: Block diagram of modeling of the dynamic object

In the figure 1: $x_s(t)$ - output model of the object, (t) - is a continuous time, t - discrete time, h_t^u, h_t^x - random interference of measurements of the relevant process variables, the index h for object variables is omitted from consideration of simplicity: $u_t = u_t^h$, $x_t = x_t^h$, $\xi(t)$ - vector random noise.

In this task, the input and output variables are controlled with the same time intervals $\triangle t$. Having made the appropriate measurements, we get the initial sample of variables "input-output" $\{x_i, u_i, i = \overline{1, s}\}$, where s is the sample size. The task

is to determine the output variables delayed by certain cycles, which must be taken into account in model (3), which will allow us to make an assumption about the parametric structure of the model of dynamic objects (3).

2 Highlighting of essential variables

In the classical case depicted in Figure 1, the identification task consists in evaluating the class of operators based on the sample $\{x_i, u_i, i = \overline{1,s}\}$. Thus, as an estimation of the model of an object, a conditional mathematical expectation can be taken of the form:

$$\overline{x}(t) = M\{x(t)/u(t)\}. \tag{4}$$

In the case when a dynamic object is described by a differential equation, with a consecutive sampling, the resulting difference equation will consistently contain all the variables: $x_{t-1}, x_{t-2}, ..., x_{t-k},$. This corresponds to the block diagram of the simulation depicted in Figure 1. Then, as a nonparametric model of the object, we can use a model in which all the coefficients of the difference equation will be taken into account:

$$x_s = \frac{1}{s}\sum_{i=1}^{s} x_i H\left(\frac{u_s - u_i}{c_s^u}\right)\prod_{j=1}^{k}\frac{1}{c_s^{x^j}} H\left(\frac{x_{s-j} - x_{i-j}}{c_s^{x^j}}\right). \tag{5}$$

In the model (5) H - the nuclear bell-shaped function, $c_s^u, c_s^{x^1}, c_s^{x^k}$ - the blurring coefficients of the nuclear function, which satisfy the convergence conditions [11]. As a bell-shaped function H various nuclei can be used. It should be taken into account that model (5) can only be used for equal measurement intervals $\triangle t$. The optimal blur parameters $c_s^{*x^1}, ..., c_s^{*x^k}$ in the presence of a training sample is found from the task of minimizing the indicator of the correspondence between the output of the object x_t and the output of the model x_s^t based on the sliding exam method when in the model (5) the index i excludes the q observation of the variable presented for the exam:

$$R(c_s^{*x^1}, c_s^{*x^k}) = \sum_{q=1}^{s}(x_s^q - x_t^q)^2 = \min_{c_s^{*x^1},...,c_s^{*x^k}}, q \neq i, \tag{6}$$

where the index i appears in the formulas (5).

Essential in the estimate (5) is that each output variable $x_{s-1}, ..., x_{s-k}$ is delayed by some values of its own blur factor $c_s^{*x^1}, ..., c_s^{*x^k}$.

From the models presented above, it can be seen that the degree of contribution of an output variable from the right-hand side of the equations to the final value of the estimate depends on

$$\frac{1}{c_s^{x^j}} H\left(\frac{x_{s-j} - x_{i-j}}{c_s^{x^j}}\right) \tag{7}$$

Expression (7) consists of two parts: $\frac{1}{c_s^{x^j}}$ and $H\left(\frac{x_{s-j}-x_{i-j}}{c_s^{x^j}}\right)$. As for the first factor $\frac{1}{c_s^{x^j}}$, the following dependence is observed: the smaller $c_s^{x^j}$, the greater the contribution the value $\frac{1}{c_s^{x^j}}$ makes to the final estimate. Thus, we can construct the following chain of inequalities:

$$c_s^{x^1} < c_s^{x^2} < ... < c_s^{x^k}, \frac{1}{c_s^{x^1}} > \frac{1}{c_s^{x^2}} > ... > \frac{1}{c_s^{x^k}} \tag{8}$$

Consider the second component of equation (7). The coefficients $c_s^{x^j}$ and the nuclear function must satisfy the following property:

$$\frac{1}{c_s^{\omega}} \int\limits_{\Omega(\omega)} H\left(\frac{\omega-\omega_i}{c_s^{\omega}}\right) d\omega = 1, \tag{9}$$

where ω is some variable. Proceeding from this condition, we can construct the following sequence of inequalities:

$$c_s^{x^1} < c_s^{x^2} < ... < c_s^{x^k}, H\left(\frac{x_{s-1}-x_{i-1}}{c_s^{x^1}}\right) < ... < H\left(\frac{x_{s-k}-x_{i-k}}{c_s^{x^k}}\right) \tag{10}$$

The algorithm for calculating significant variables x_{s-j} is constructed according to the following scheme. First, set the initial value of k. We build a model by the formula (5) and consider the relative modeling error W_0:

$$W_0 = \sqrt{\frac{\frac{1}{s}\sum_{i=1}^{s}(x_i - x_i^s)^2}{\sum_{i=1}^{s}\frac{1}{s-1}(m_x - x_i)^2}} \tag{11}$$

where m_x is the mathematical expectation.

Then, at each i-th iteration, we perform the following set of actions:

1. For each coefficient $c_s^{x^1}, ..., c_s^{x^k}$ there is an optimal value: $c_s^{x^1} = c_s^{*x^1}, ..., c_s^{x^k} = c_s^{*x^k}$.

2. We find from all the obtained values the maximum - $c_{max_s}^{x^j}$.

3. We build a model by formula (5) excluding the factor $H\left(\frac{x_{s-j}-x_{i-j}}{c_s^{x^j}}\right)$ j, taking into account that j is the number for $c_{max_s}^{x^j}$.

4. Consider a relative error W_j.

These actions will be repeated while $W_i > W_{i-1}$.

Conclusions

The algorithm for determining the parametric structure of a linear dynamic object based on the application of the rule for the allocation of essential variables is proposed in the article. This approach is related to the determination of the optimal coefficients of nuclear function using a nonparametric model of the regression function from observations. The article proposes an algorithm whose operation reduces

to determining the order of the differential equation of a dynamic object, which is the order of the highest derivative in the corresponding differential equation when performing an analogy with the description of the process under study in continuous time. This algorithm consists of several stages, including the determination of optimal blurring coefficients, their selection and construction of the final model. One of the main advantages of the proposed algorithm in comparison with today's dominant methods of restoring the parametric structure is the fact that the developed nonparametric algorithm is more applicable to practical problems, since it is able to work in conditions of incomplete a priori information about the object.

References

[1] Medvedev A.V. (2015). *Osnovy teorii adaptivnyh system (Basic theory of adaptive systems)*. SibGAU, Krasnojarsk.

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

[3] Medvedev A.V. (2010). Teorija neparametricheskih sistem. Processy (he theory of non-parametric systems. Processes). *Vestnik SibGAU*. Vol. **3 (29)**, pp. 4-9.

[4] Medvedev A.V. (1977). *Adaptacija i obuchenie v uslovijah neparametricheskoj neopredelennosti (Adaptation and learning in a non-parametric uncertainty*. Nauka, Novosibirsk.

[5] Fel'dbaum A.A. (1963). *Osnovy teorii optimal'nyh avtomaticheskih system (Fundamentals of the theory of optimal automatic systems)*. Fizmatgiz, Moskow.

Goodness of Fit Procedures for Bivariate Failure Time Data Based on a Copula Approach

Hannelore Liero
Institute of Mathematics, University of Potsdam, Germany
e-mail: liero@uni-potsdam.de

keywords: bivariate survival times; censoring; copula; goodness of fit

1 Modeling Bivariate Survival Data by Copulas

The aim of this paper and of the talk is to give a short survey on inference methods for the distribution of bivariate failure time data. And, more important – to show that there are still a lot of interesting open problems on this field.

For one-dimensional complete survival data we have already a well-established and applicable 'package' of methods. Also for censored data, there are good methods for estimation and testing. Copula models are a very useful tool for modeling multivariate complete data. To apply such models we have to combine our 'survival theory knowledge' with the theory of copulas. Methods for estimation and testing copulas for complete data are well-developed. However, for censored data such inference procedures are still a challenge. There are already proposals for test procedures and their realization. But there are questions which require a deeper theoretic consideration.

Let T_1 and T_2 be two survival times with the joint survival function S and marginal survival functions S_1 and S_2, respectively. There are well established methods - for complete and for censored data - to fit S_1 and S_2. However to fit the joint S one has to take into account the dependence structure of the two variables. And this task can be solved by modeling the survival function by a copula.

A copula function $\mathbb{C}$ is a survival function with uniform marginals: $\mathbb{C} : [0,1]^2 \to [0,1]$. We suppose that the marginal survival functions are continuous, then according to Sklar's theorem (1959), there exists a unique copula $\mathbb{C}$, such that

$$S(t_1, t_2) = \mathsf{P}(T_1 > t_1, T_2 > t_2) = \mathbb{C}(S_1(t_1), S_2(t_2)). \tag{1}$$

In this form the problem of describing the dependence structure is separated from the problem of fitting the marginals. In the following we let the marginals be unspecified and consider only the copula function. It is supposed that $\mathbb{C}$ belongs to a certain class of parametric copula functions, say $\mathcal{C} = \{\mathbb{C}_\vartheta \,|\, \vartheta \in \Theta \in \mathbb{R}^k\}$.

1.1 Archimedean Copulas

An important parametric class of copulas is the class of Archimedean copulas, which are generated by a function ψ_ϑ

$$\mathbb{C}_\vartheta(u, v) = \psi_\vartheta^{-1}(\psi_\vartheta(u) + \psi_\vartheta(v)),$$

where ψ_ϑ is a convex, strictly decreasing function defined on $[0, 1]$ with $\psi_\vartheta(1) = 0$.

(Remark: We consider only so-called strict Archimedean copulas, where the generator satisfies $\psi_\vartheta(0) = \infty$, i.e. the inverse exists for all $t \in [0, \infty]$.)

Example 1. *The Gumbel copula is defined by*

$$\mathbb{C}_\vartheta(u_1, u_2) = \exp\big(- [(-\log\ u_1)^\vartheta + (-\log\ u_2)^\vartheta]^{\frac{1}{\vartheta}}\big).$$

Here the generator function is $\psi_\vartheta(u) = (-\log u)^\vartheta$ *and* $\vartheta \geq 1$. *Note, for* $\vartheta = 1$ $\mathbb{C}(u_1, u_2) = u_1 \cdot u_2$, *i.e.,* T_1 *and* T_2 *are independent. The dependence can be characterized by Kendalls tau, which is given by* $\tau_{\text{Gumbel}} = 1 - 1/\vartheta$.

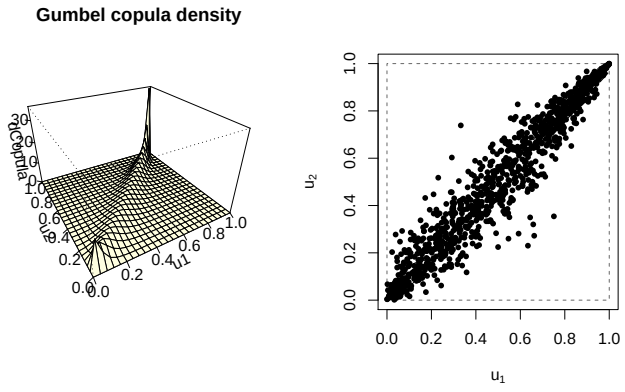


Figure 1: Perspective plot of the density of the Gumbel copula with parameter $\vartheta = 2$ and a scatter plot of 1000 simulated data from this copula.

Archimedean copulas arises in **shared frailty models**. Let us shortly describe this approach: Let Z be a frailty variable with the distribution F_Z and suppose that the conditional hazard function of T_j given $Z = z$ is $\Lambda_j(t|z) = z\Lambda_{j0}(t)$, where Λ_{j0} is the baseline hazard function, $j = 1, 2$. Since $S_j(t|z) = \exp(-\Lambda_j(t|z))$ we obtain under the assumption, that T_1 and T_2 are conditionally independent given the frailty variable, the following:

$$\begin{aligned} \mathsf{P}(T_1 > t_1, T_2 > t_2 | Z = z) &= S_1(t_1|z)S_2(t_2|z) \\ &= \exp(-z(\Lambda_{10}(t_1) + \Lambda_{20}(t_2))). \end{aligned}$$

Let $\mathcal{L}$ be the Laplace transform of the frailty distribution. Taking the expectation with respect to the frailty distribution we obtain

$$\begin{aligned} \mathsf{P}(T_1 > t_1, T_2 > t_2) &= \int \exp(-z(\Lambda_{10}(t_1) + \Lambda_{20}(t)))dF_Z(z) \\ &= \mathsf{E}_Z(\exp(-Z(\Lambda_{10}(t_1) + \Lambda_{20}(t_2)))) \\ &= \mathcal{L}(\Lambda_{10}(t_1) + \Lambda_{20}(t_2)) \\ &= \mathcal{L}(\mathcal{L}^{-1}(S_1(t_1)) + \mathcal{L}^{-1}(S_2(t_2))). \end{aligned}$$

Here it used that $S_j(t) = \mathcal{L}(\Lambda_{j0}(t))$. In other words, the inverse of the Laplace transform of the frailty distribution is a generator of an Archimedean copula. (Note that in this case also the marginal distributions depend on the parameter ϑ.)

Example 2. *Suppose that the frailty variable is gamma distributed with the parameters* $\kappa = \lambda = 1/\vartheta$. *Then* $\mathsf{E}Z = 1$ *and* $\mathsf{Var}Z = \vartheta$. *The Laplace transform of the gamma distribution is* $\mathcal{L}(s) = (\lambda/(\lambda + s))^\kappa = (1 + \vartheta s)^{-1/\vartheta}$. *Taking* $\psi_\vartheta(u) = \mathcal{L}^{-1}(u) = (u^{-\vartheta} - 1)/\vartheta$ *we obtain the Clayton copula*

$$\mathbb{C}(u_1, u_2) = (u_1^{-\vartheta} + u_2^{-\vartheta} - 1)^{-\vartheta} \qquad \vartheta > 0.$$

For this family of copulas Kendall's tau is given by $\tau_{\text{Clayton}} = \vartheta/(\vartheta + 2)$.

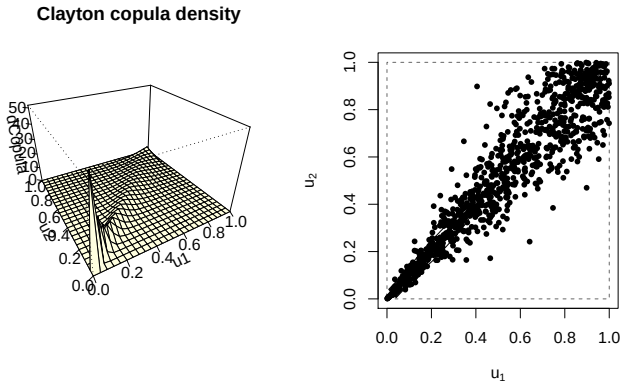


Figure 2: Perspective plot of the density of the Clayton copula with parameter $\vartheta = 8$ and a scatter plot of 1000 simulated data from this copula.

2 Testing the copula function when the observations are complete: A survey of methods

Consider a sample (T_{i1}, T_{i2}) $i = 1, \ldots, n$ of i.i.d. r.v.'s. In this section the most important methods for estimating and testing the copula function in model (1) are summarized:

2.1 Test statistics based on the empirical copula

From $\mathbb{C}(u_1, u_2) = S(S_1^{-1}(u_1), S_2^{-1}(u_2))$ it seems to be natural to estimate $\mathbb{C}$ nonparametrically by

$$\widetilde{C}_n(u_1, u_2) = \widehat{S}_n(S_{n,1}^-(u_1), S_{n,2}^-(u_2)) \tag{2}$$

where

$$S_n(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(T_{i1} > t_1, T_{i2} > t_2)$$

and $S_{n,j}^-$ are the generalized inverse of the empirical marginal survival functions and $\mathbf{1}(\cdot)$ denotes the indicator function. Instead of the empirical copula defined in (2) it is sometimes easier to consider

$$\widehat{C}_n(u_1, u_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(S_{n,1}(T_{i1}) > u_1, S_{n,2}(T_{i2}) > u_2).$$

One can verify that $\widetilde{C}_n$ and $\widehat{C}_n$ differ only by a small quantity n^{-1}.

The asymptotic behavior of the copula process is considered in a lot of papers. We mention here Deheuvels (1979) and Segers (2012). The main point of these papers is to derive the weak convergence of the copula process to a centered Gaussian process under different conditions on the underlying distribution.

Let us consider the following test problems: The simple test problem

$$\mathcal{H} : \mathbb{C} = \mathbb{C}_0 \qquad \text{against} \qquad \mathcal{K} : \mathbb{C} \neq \mathbb{C}_0 \tag{3}$$

or the composite problem

$$\mathcal{H} : \mathbb{C} \in \mathcal{C} = \{\mathbb{C}_\vartheta \,|\, \vartheta \in \Theta \in \mathbb{R}^k\} \qquad \text{against} \qquad \mathcal{K} : \mathbb{C} \notin \mathcal{C}. \tag{4}$$

Goodness-of-fit tests (GOF tests) based on the empirical copula are investigated by Genest et al. (2006), Genest and Rémillard (2008) and Fermanian (2013). Fermanian considers test statistics of the following form:
Kolmogorov-Smirnov type statistics

$$\mathcal{T}_n^{KS} := \sup_{\boldsymbol{u}\in[0,1]^2} \sqrt{n}|\widehat{C}_n(\boldsymbol{u}) - \mathbb{C}_0(\boldsymbol{u})| \qquad \boldsymbol{u} = (u_1, u_2)$$

or the corresponding analogue for testing (4) where $\mathbb{C}_0$ is replaced by $\mathbb{C}_{\widehat{\vartheta}}$ for some $\sqrt{n}$-consistent estimator $\widehat{\vartheta}$. Another type of test statistics are of Anderson-Darling type

$$\mathcal{T}_n^{AD} := n\int (\widehat{C}_n(\boldsymbol{u}) - \mathbb{C}_0(\boldsymbol{u}))^2 w_n(\boldsymbol{u}) d\boldsymbol{u}$$

where w_n is a weight function.

Furthermore, Cramér-von Mises statistics

$$\mathcal{T}_n^{CM} := n\int (\widehat{C}_n(\boldsymbol{u}) - \mathbb{C}_0(\boldsymbol{u}))^2 \widehat{C}_n(d\boldsymbol{u})$$

or chi-squared type statistics can be used to test (3); for testing (4) $\mathbb{C}_0$ is replaced by $C_{\widehat{\vartheta}}$:

$$\mathcal{T}_n^{Chi} := n \sum_{j=1}^{M} w_j (\widehat{C}_n - \mathbb{C}_0)^2 (B_j)$$

where B_j are disjoint rectangles in $[0,1]^2$ and w_j are weights.

To derive a test procedure based on these test statistics it is necessary to know the distribution of the statistics under the null hypothesis or at least, to have an limit distribution. As formulated above, the copula process converges in distribution to a Gaussian process $\mathbb{Z}$, which depends on the hypothetical $\mathbb{C}_0$. Therefore bootstrap methods are necessary to approximate p-values. We will come back to this problem in Section 3.3.

2.2 The Kendall process

Another idea for GOF testing for bivariate r.v.'s is to reduce the problem to a one-dimensional problem. This can be done by considering the Kendall distribution: Let $\mathbb{C}$ be the copula of (T_1, T_2), the Kendall function is defined by

$$K(z) = \mathsf{P}(\mathbb{C}(S_1(T_1), S_2(T_2)) \leq z).$$

The function K depends on $\mathbb{C}$ only, it is a summary of the underlying dependence structure, sometimes it is called Kendall's dependence functions. In general, if we use K instead $\mathbb{C}$ we lose information. However, if we restrict our consideration to Archimedean copula, introduced in (1.1), we have the following statement which was proved by Genest and Rivest (1993):
Let U_1 and U_2 be uniform r.v.'s with copula function $\mathbb{C}(u_1, u_2) = \psi^{-1}(\psi(u_1) + \psi(u_2))$ for some convex decreasing function ψ on $(0, 1]$ with $\psi(1) = 0$. Define

$$X := \frac{\psi(U_1)}{\psi(U_1) + \psi(U_2)} \quad \text{and} \quad V := \mathbb{C}(U_1, U_2)$$

and

$$\lambda(v) = \frac{\psi(v)}{\psi'(v)} \qquad \text{for} \quad 0 < v \leq 1.$$

Then the following holds:
1.) X is uniformly distributed on $(0, 1)$.
2.) V is distributed as

$$K(v) = v - \lambda(v) \qquad \text{Kendall distribution.}$$

3.) X and V are independent.
In fact, the existence of a function ψ for which properties 1.) to 3.) hold implies that $\mathbb{C}(u_1, u_2) = \psi^{-1}(\psi(u_1) + \psi(u_2))$ on its entire domain.

Let us estimate the Kendall function nonparametrically. In a first step "pseudo observations" of $V_i = S(T_{1i}, T_{2i})$ are constructed:

$$\widehat{V}_i = \frac{1}{n-1} \sum_{j=1}^{n} \mathbf{1}(T_{1j} > T_{1i}, T_{2j} > T_{2i}) \qquad i = 1, \ldots, n,$$

then K is estimated by the empirical distribution function of the $\widehat{V}_i$'s:

$$\widehat{K}_n(v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\widehat{V}_i \le v).$$

Example 3. *Figures 3 show the Kendall functions for the copulas considered in Example 1, i.e.* $K_{\text{Gumbel}}(v) = v - v \log v/\vartheta$, *and Example 2, i.e.* $K_{\text{Clayton}}(v) = v + v(1 - v^{\vartheta})/\vartheta$ *and their nonparametric estimates.*

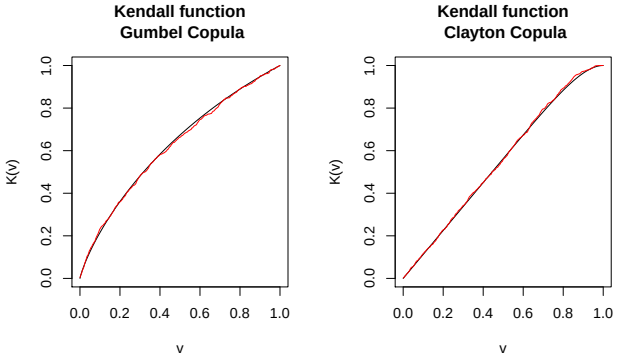


Figure 3: Kendall functions for the Gumbel copula and the Clayton copula considered in Example 1 and Example 2 and their nonparametric estimators. The underlying parameters are $\vartheta = 2$ and $\vartheta = 8$, respectively.

The asymptotic properties of the Kendall process are investigated by Barbe et al. (1996). They derive conditions ensuring that the process converges weakly to a centered Gaussian process and give an explicit formula for the covariance of the limiting process.

The test problem has now the form

$$\mathcal{H} : \mathbb{C} \in \mathcal{C} = \{\mathbb{C}_\vartheta(u_1, u_2) = \psi_\vartheta^{-1}(\psi_\vartheta(u_1) + \psi_\vartheta(u_2)) \,|\, \vartheta \in \Theta \in \mathbb{R}^k\}$$
$$\text{against} \qquad \mathcal{K} : \mathbb{C} \notin \mathcal{C}$$

where the generator ψ_ϑ is known except of the unknown parameter ϑ.

Tests are based mostly on L^2-distances. Let us mention here the statistic proposed by Wang and Wells (2000)

$$\mathcal{T}_n^{Ken} := \int_0^1 (\widehat{K}_n(v) - K_{\widehat{\vartheta}}(v))^2 dv. \tag{5}$$

Here K_ϑ is the Kendall function for the hypothetical model, $\widehat{\vartheta}$ is an appropriate consistent estimator. As usual an estimator for parameters can be obtained by the maximum likelihood method; here the following approach is also appropriate: The correlation coefficient Kendall's tau satisfies the equation

$$\tau = 4 \int_0^1 \lambda_\vartheta(v) dv + 1 = \zeta(\vartheta).$$

If ζ is a one-to-one function we obtain an estimate for ϑ by inverting ζ, i.e. $\widehat{\vartheta} = \zeta^{-1}(\widehat{\tau})$, where $\widehat{\tau}$ is an estimate of τ.
A detailed discussion of this test statistic is given by Yilmaz (2009). Wang and Wells extended this test statistic also for censored data; moreover Chen and Fan (2004) developed a model selection procedure based on test statistics of the form (5).

A test statistic of Cramér-von Mises type is given by

$$\mathcal{T}_n^{KenCM} := n \int (\widehat{K}_n(v) - K_{\widehat{\vartheta}}(v))^2 d\widehat{K}_n(v) \tag{6}$$

The statement of Genest and Rivest (1993) formulated above is the basis for another approach: Wang (2010) considers the r.v.'s X and V. Under the null hypothesis they are independent. Using consistent estimators of X_i and V_i Wang proposes to test the correlation coefficient between X and V. The test statistic is given by Fisher's Z statistic

$$\mathcal{T}_n^{Z} := \frac{1}{2} \log \left(\frac{1 + r_n}{1 - r_n} \right) \tag{7}$$

where r_n is the (Pearson) correlation of the sample $(\widehat{X}_i, \widehat{V}_i)$ with $\widehat{X}_i = \frac{\psi_{\widehat{\vartheta}}(\widehat{S}_1(T_{1i}))}{\psi_{\widehat{\vartheta}}(\widehat{S}(T_{1i}, T_{2i}))}$
Wang states that the distribution of the test statistic under the null hypothesis converges to the standard normal distribution. Moreover, he extends it also to the case of censored data. Chen (2012) implemented this procedure and used it also for a model selection procedure.

2.3 Tests using the copula density

As already written, the limit distribution of the test statistics presented above depend on the underlying distribution of the data. Fermanian (2005) proposed test statistics which are in the limit distribution-free. In this approach it is assumed that there exist a copula density c. A nonparametric kernel estimator of c is defined by

$$\widehat{c}_n(\boldsymbol{u}) = \frac{1}{h^2} \int k \left(\frac{\boldsymbol{u} - \boldsymbol{v}}{h} \right) \widehat{C}_n(d\boldsymbol{u})$$

where k is a 2-dimensional kernel and $h = h_n$ is a bandwidth sequence tending to zero as $n \to \infty$. Fermaninan assumes that the kernel is the product of two compactly

supported kernels k_j, $j = 1, 2$. The first test statistic is simple a chi-square type statistic

$$\mathcal{T}_n^{denChi} = \frac{nh^2}{\int k^2} \sum_{j=1}^{m} \frac{(\widehat{c}_n(\boldsymbol{u}_j) - c(\boldsymbol{u}_j; \widehat{\vartheta}))^2}{c(\boldsymbol{u}_j, \widehat{\vartheta})^2}.$$

where $\boldsymbol{u}_j$ are arbitrarily chosen points in $[0, 1]^2$. Fermanian proved, that under certain smoothness conditions on the underlying distribution, conditions on h_n and on the estimator $\widehat{\vartheta}$ the distribution of $\mathcal{T}_n^{denChi}$ tends to a χ^2-distribution with m degrees of freedom. Applying this test, one has to note that the choice of m, of the points $\boldsymbol{u}_j$ and the choice of the bandwidth is important.

A further test based on nonparametric density estimators is given by the statistic

$$\mathcal{J}_n = \int (\widehat{c}_n(\boldsymbol{u}) - (k * c_{\hat{\vartheta}})(\boldsymbol{u}))^2 w(\boldsymbol{u}) d\boldsymbol{u}$$

where $c_{\hat{\vartheta}}$ is the hypothetical density with the estimated parameter, $(k * c_{\hat{\vartheta}})(\boldsymbol{u})$ is the convolution between the estimated hypothetical density and the kernel k_h with $k_h(\cdot) = k(\cdot/h)/h^2$, w is a weight function. This test procedure is an extension of tests investigated in Rosenblatt (1975), Ghosh, Huang (1991), Liero, Läuter, Konakov (1998). Fermanian (2005) stated conditions ensuring that $\mathcal{J}_n$ properly standardized converges weakly to a normal distribution. Based on this limit theorem the author proposed as test statistic

$$\mathcal{T}_n^{denCM} = \frac{(nh)^2 \, (\mathcal{J}_n - \widehat{A}_n)^2}{\widehat{\sigma}_n^2}$$

with

$$\widehat{A}_n = (nh^2)^{-1} \int k^2(\boldsymbol{z})(c_{\hat{\vartheta}} w)(\boldsymbol{u} - h\boldsymbol{z}) d\boldsymbol{z} d\boldsymbol{u} + (nh)^{-1} \int c_{\hat{\vartheta}}^2 w (k_1^2 + k_2^2)$$

and

$$\widehat{\sigma}_n^2 = 2 \int \widehat{c}_n^2 w \cdot \int (\int k(\boldsymbol{u}) k(\boldsymbol{u} + \boldsymbol{v}) d\boldsymbol{u})^2 d\boldsymbol{v}.$$

Chen and Huang (2007) propose a Cramér-von Mises type test based on a kernel estimator for $\mathbb{C}$, in other words based on a kernel smoothed empirical copula.

2.4 Rosenblatt transformation

Already in 1952 Rosenblatt showed how a d-dimensional vector of continuous r.v.'s can be transformed into a vector of independent uniformly distributed r.v.'s. Applying the Rosenblatt transformation to the copula we obtain the following: The copula $\mathbb{C}$ is the joint distribution function of $\boldsymbol{U} = (S_1, (T_1), S_2(T_2))$. We define the 2-dimensional vector $\boldsymbol{W}$ by

$$W_1 = S_1(T_1) = U_1 \qquad W_2 = \mathbb{C}(U_2 | U_1)$$

where $\mathbb{C}(\cdot|u_1)$ is the law of U_2 given $U_1 = u_1$. The variables W_1 and W_2 are independent and uniformly distributed, i.e. their copula is the independence copula $\mathbb{C}_\perp(\boldsymbol{w}) = w_1 w_2$. Test statistics based of this transformation are considered in Breyman (2003) and Dobrić and Schmidt (2007). However, one has to take into account: The r.v.'s W_1 and W_2 are not observable, one has to replace them by the "pseudo-obersvations" $\widehat{W}_{i1}$ and $\widehat{W}_{i2}$. Moreover, if we test the composite hypothesis, we have in addition an error coming from the estimation of ϑ. Consequently, it is not clear, whether it is useful to compare (measured in some distance) the empirical copula of the transformed data, i.e. $\frac{1}{n}\sum_{i=1}^n 1(\widehat{W}_{i1} > w_1, \widehat{W}_{i2} > w_2)$ with the independence copula.

3 Inference Based on Censored Data

3.1 Censoring for bivariate Data

In survival analysis one or both of the two underlying lifetimes may be subject to random right censoring. In general this can be described as follows: The observations are

$$Y_{ij} = \min(T_{ij}, C_{ij}) \quad \text{and} \quad \Delta_{ij} = 1(T_{ij} \leq C_{ij}) \qquad j = 1, 2 \;\; i = 1, \ldots, n$$

where the C_{ij}'s are random censoring variables.

Nonparametric estimators for the survival function under censoring were discussed by several authors, the following should be mentioned: Dabrowska (1988), Pruitt (1990, 1991), van der Laan (1996) and Prentice et al. (2004). In these approaches the authors do not use a copula model.

In applications there are special forms of censoring and so there are estimation methods which take into account the form of censoring. Here are some examples: Lin and Ying (1993) assumes that there is only one censoring variable C_1 with survival function G, that is, we have the special case $Y_{ij} = \min((T_{ij}, C_i) \quad \text{and} \quad \Delta_{ij} = 1(T_{ij} \leq C_i) \qquad j = 1, 2 \;\; i = 1, \ldots, n$. The C_i's and the r.v.'s (T_{i1}, T_{i2}) are independent; Lin and Ying proposed to estimate the joint survival function S by

$$\widehat{S}_n(t_1, t_2) = \frac{n^{-1}\sum_{i=1}^n 1(Y_{i1} > t_1, Y_{i2} > t_2)}{\widehat{G}_n(\max\{t_1, t_2\})}$$

where $\widehat{G}_n$ is the Kaplan-Meier estimator for G.

Gribkova and Lopez (2015) study nonparametric estimators for the distribution function of the form

$$\widehat{F}_n(t_1, t_2) = \frac{1}{n}\sum_{i=1}^n W_{in} 1(Y_{i1} \leq t_1, Y_{i2} \leq t_2) \tag{8}$$

where W_{in} are random weights designed to compensate asymptotically the bias caused by the particular censoring structure. They consider as a first special case the

situation that only one of the two r.v.'s is censored, say $C_2 = \infty$ and $Y_2 = T_2$ almost surly. For this case they set

$$W_{in} = \frac{\Delta_{i1}}{\widehat{G}_n(Y_{i1})}$$

where $\widehat{G}_n$ is the Kaplan-Meier estimator of the survival function of C_1.

If we have pairs (C_{i1}, C_{i2}) of censoring variables which are independent of (T_{i1}, T_{i2}) and which are distributed according to a copula

$$S_G(x_1, x_2) = \mathsf{P}(C_1 > x_1, C_2 > x_2) = \mathbb{C}_g(G_1(x_1), G_2(x_2))$$

where $\mathbb{C}_G$ is known (or known except of an finite dimensional parameter). Denoting by $\widehat{G}_j$ the Kaplan-Meier estimators of the marginal survival functions of the C_j's the authors propose

$$W_{in} = \frac{\Delta_{i1}\Delta_{2i}}{\mathbb{C}_G(\widehat{G}_1(Y_{i1}), \widehat{G}_2(Y_{i2}))}.$$

Gribkova and Lopez investigated consistency of the copula process based on weights W_{ni} introduced with the definition (8). Furthermore, they derived conditions on $\widehat{F}_n$, on the weights W_{ni} and on the copula function ensuring that the empirical copula process with the weights W_{ni} converges weakly to a Gaussian process.

3.2 GOF test statistics under censoring

In this section we consider how the test procedures introduced in Section 2 are extended to the case that data can be censored.

Tests based on the empirical copula function were considered by Andersen et al. (2005). They proposed tests statistics to check the bivariate survival function of a shared frailty model, which leads to the test an Archimedean copula. The marginal survival functions are estimated by the (one-dimensional) Kaplan-Meier estimators $\widehat{S}_{nj}$, and the joint survival function is estimated by Pruitts's estimator, say $\widehat{S}_{n\mathrm{Pr}}$. The empirical copula is given by $\widehat{C}_n(u_1, u_2) = \widehat{S}_{n\mathrm{Pr}}(\widehat{S}^-_{n1}(u_1), \widehat{S}^-_{n2}(u_2))$ and as test statistic a χ^2-type statistic

$$\mathcal{D}_n^{Chi} = \sum_{j=1}^{m}(A_j - B_j)^2 \tag{9}$$

is proposed. Here the unit square is partitioned into m parts, A_j is the mass assigned by $\widehat{C}_n$ to the jth part, and B_j the mass assigned by the estimated hypothetical copula $\mathbb{C}_{\widehat{\vartheta}}$.

Gribkova and Lopez (2015) considered a Cramér-von Mises type test statistic:

$$\mathcal{D}_n^{CM} = n\int(\widehat{C}_n(\boldsymbol{u}) - \mathbb{C}_{\hat{\vartheta}}(\boldsymbol{u}))^2\widehat{C}_n(d\boldsymbol{u}) \tag{10}$$

where C_n is a weighted copula estimator with weights W_{ni} introduced in (8).

Tests for complete data which are based on the Kendall process were extended to the case of censored data by Wang and Wells (2000), Chen and Fan (2007) and Wang (2010). In the censored case the Kendall function is estimated by

$$\widehat{K}_n(v) = 1 - \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{1}(\widehat{F}_n(Y_{(i)1}, Y_{(j)2}) > v)\, \widehat{F}(\Delta Y_{(i)1}, \Delta Y_{(j)2})$$

where $Y_{(1)1} \leq \cdots \leq Y_{(n)1}$ and $Y_{(1)2} \leq \cdots \leq Y_{(n)2}$ be ordered observations. $\widehat{F}_n$ is a non-parametric estimator of the survival function, and $\widehat{F}(\Delta y_{(i)1}, \Delta y_{(j)2}) = \widehat{F}(y_{(i)1}, y_{(j)2}) - \widehat{F}(y_{(i)1}, y_{(j-1)2}) - \widehat{F}(y_{(i-1)1}, y_{(j)2}) + \widehat{F}(y_{(i-1)1}, y_{(j-1)2})$. Wang and Wells (2000) stated sufficient conditions for the weak convergence of the Kendall process to a mean-zero Gaussian process on the space $\mathcal{D}[\xi, 1]$. Note, since censoring is possible, one cannot recover F on the whole rectangle $[0,1]^2$, also consistency and weak convergence of the Kendall process can be proved only on an interval $[\xi, 1]$.

Based on this result Wang and Wells (2000) gave the limiting distribution under the null hypothesis of the test statistic

$$\mathcal{D}_n^{Ken} = \int_{\xi}^{1} (\widehat{K}_n(v) - K(v, \widehat{\vartheta}))^2 dv$$

where $K(\cdot, \widehat{\vartheta})$ is the estimated hypothetical Kendall function.

Chen and Fan (2007) used this approach for defining a model selection procedure in the class of Archimedean copulas. Wang (2010) considered the approach defined in (7) also in the censored case. Here they apply an imputation technique to handle the distribution of the variables X and V under censoring.

As far as I know, test statistics based on the Rosenblatt transform and tests based on copula densities are not extended to the case of censored data.

3.3 Realization of the test procedures and power considerations

To formulate the test procedures one has to know the distribution of the test statistic under the null hypothesis or at least, to know the limiting distribution. In some cases discussed above such a limiting distribution is known, however it depends on the underlying copula. Moreover, the variance of the limit process is often very complicated. Thus, the limit statement is a theoretic background for a procedure, however for the application of the test procedures other approximations than the limit distribution are required. Such approximations are provided by bootstrap methods. Many procedures consists of the following main steps:

1. Generating of R samples of observations under the hypothetical model.

   1.1 Generation of the lifetimes

   - When $\mathbb{C}_0$ and the marginal distributions under $\mathcal{H}$ are completely known, the generation of the hypothetical lifetimes $(T_{i1}^{*r}, T_{i2}^{*r})$ is obvious.

- When $\mathcal{H}$ is a composite hypothesis, ϑ es estimated from the original data, say $\widehat{\vartheta}$.
- If the marginal distributions are unknown, estimate them nonparametrically.
  Thus, R samples $(T_{i1}^{*r}, T_{i2}^{*r})$ are generated according to $\mathbb{C}_{\hat{\vartheta}}$ with marginals $\widehat{S}_{jn}$, $j = 1, 2$, $r = 1, \ldots, R$.

1.2 Generation of the censoring times

- Estimation of the censoring variables according to the (known or) estimated censoring distribution $\widehat{G}_n$: $(C_{i1}^{*r}, C_{i2}^{*r})$

  The samples

- Finally, set for $r = 1, \ldots, R$, $i = 1, \ldots, n$ and $j = 1, 2$
  $Y_{ij}^{*r} = \min(T_{ij}^{*r}, C_{ij}^{*r})$ and $\Delta_{ij}^{*r} = 1(T_{ij}^{*r} \leq C_{ij}^{*r})$

2. For each r based on the bootstrap sample $(Y_{i1}^{*r}, Y_{i2}^{*r}, \Delta_{i1}^{*r}, \Delta_{i2}^{*r})$ compute the test statistic $\mathcal{D}_n^{*r}$. Note that for the composite test problem this includes the computation of the estimator $\widehat{\vartheta}^{*r}$.

3. Let $\mathcal{D}_{org}$ be the value of the test statistic based on the original data. The bootstrap p-value is given by

$$p_{boot} = \frac{1}{r} \sum_{r=1}^{R} 1(\mathcal{D}_n^{*r} \geq \mathcal{D}_{org}).$$

The procedure described above is only one possibility. In the literature there are several proposals to handle the problem of generating the T_{ij}^{*r}'s according the hypothetical model when the marginals are not specified.

A detailed discussion and simulation study of GOF tests for copulas for shared frailty models is given by Andersen et al.(2005). In this study the test statistic (9) is applied. For the estimation of the hypothetical copula the frailty assumption is used. The authors describe the computation of the p-value by bootstrap and include also a power study.

A bootstrap procedure for the realization of the test procedure based on (10) are also given and demonstrated by an example.

Wang and Wells (2000), Wang (2010) and Chen and Fan (2007) apply also bootstrap procedures for the realization of the tests based on the statistics given in Section 3.2. The simulation results show that the test procedures behaves quite well - under the null hypothesis the prescribed α is met. Furthermore, power consideration show that tests based on the Kendall process distinguish between different Archimedean copulas. However there are still open questions: A comparison between different test procedures. What is the influence of the parameter estimation $\widehat{\vartheta}$ on the power? What is the influence of the censoring?

3.4 Some simulations

In this section some simulations are presented to demonstrate the behavior of the test procedure based on the Kendall function.

In the first simulation the simplest case is considered: The null hypothesis is that the data are distributed according to a Clayton copula, i.e., $\mathcal{H} : \mathbb{C} = \mathbb{C}_{\vartheta=8}^{\text{Clayton}}$. The marginals are completely known and there is no censoring.

We apply the test based on Cramér-von Mises type statistic (6) to a sample of size $n = 250$, a corresponding procedure is given the R package *gofcopula*. The procedure is 200 times repeated with $R = 100$ bootstrap replications. In the simulations the nominal level was not matched – in 12 of the 200 repetitions $\mathcal{H}$ was rejected, i.e. we have 0.06 instead of $\alpha = 0.05$. The empirical distribution function of the p-values is given in Figure 4.

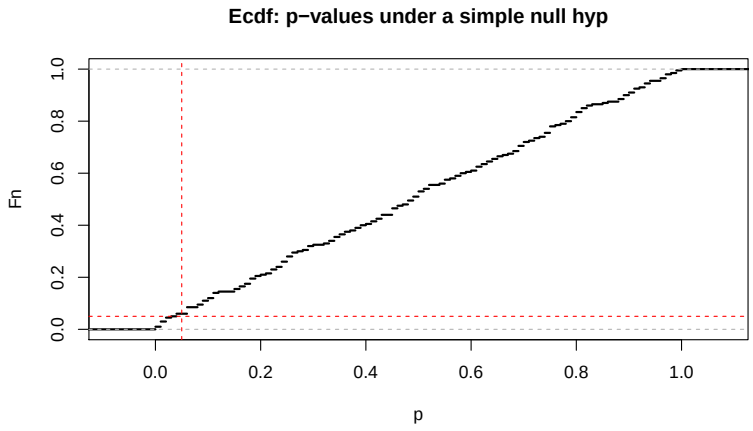


Figure 4: Empirical distribution function of the p-values of 200 simulations. Test statistic is the CvM-distance of Kendall function. The simple null hypothesis is the Clayton copula with $\vartheta = 8$.

In a second simulation the composite hypothesis was checked $\mathcal{H} : \mathbb{C} \in \mathcal{C}$, where $\mathcal{C}$ is the class of the Clayton copulas. Here parameter estimates are used in each bootstrap step. In my simulation there were only 6 rejections, that is the estimated error is 0.03.

In the third case two-dimensional data were simulated according to the copula as above, however the marginal distributions are exponential distributions with parameter $\lambda = 1$. In this case the test uses pseudo-observations. Unfortunately, the empirical distribution of the p-values of the 200 simulations is far from being a uniform distribution, see Figure 5. In my simulation there was no rejection of $\mathcal{H}$.

The last simulation shows the behavior of the test under a fixed alternative. The null hypothesis is that $\mathbb{C}$ belongs to the class of Clayton copulas. The underlying data

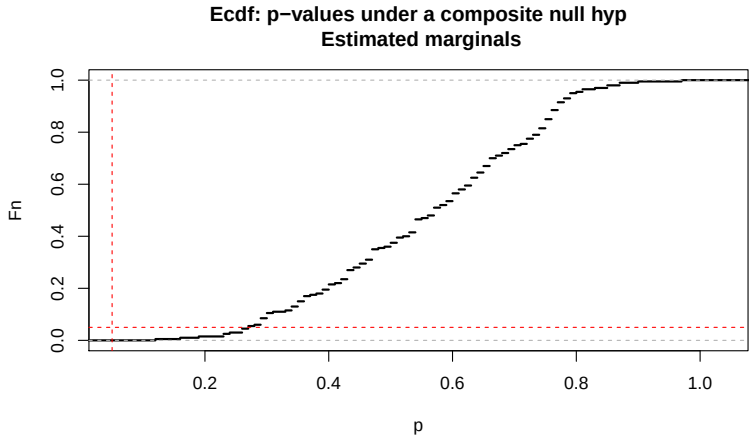


Figure 5: Empirical distribution function of the p-values. Test procedure uses pseudo-observations.

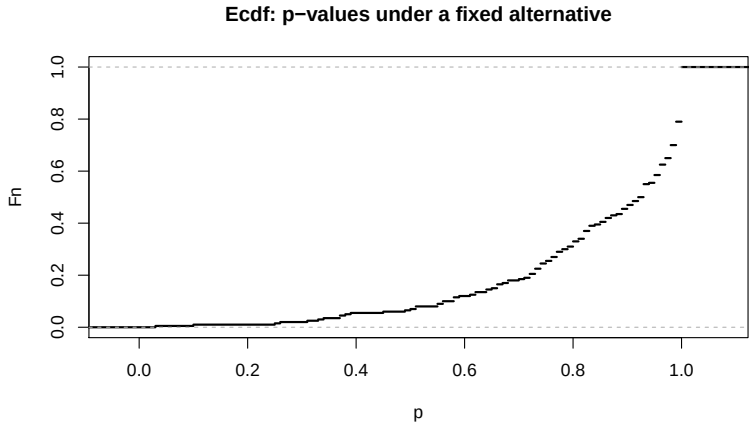


Figure 6: Empirical distribution function of the p-values under the alternative.

were generated according to a Gumbel copula with parameter $\vartheta = 2$. Unfortunately, only one test of the 200 repetitions rejected the null hypothesis. Figure 6 shows the empirical distribution of the p-values.

Test procedures for censored data will be demonstrated and discussed in the talk.

References

[1] Andersens, P. K.; Ekstrom, C.T.; Klein, J. P.; Shu and Zhang, M.-J. (2005). A class of goodness of fit tests for a copula based on bivariate right-censored data. Biometrical Journal **47**, 815–824

[2] Barbe, P. A.; Genest, C.; Ghoudi, K. and Bruno, R. (1996). On Kendall's process. Journal of Multivariate Analysis **58**, 197–229

[3] Breymann, W. ; , Dias, A. and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. Quantitative finance **3**, 1–14.

[4] Chen, X. and Fan, Y. (2007). A model selection test for bivariate failure-time data. Econometric Theory **23**, 414–439

[5] Chen, S. X. and Huang, T. M. (2007). Nonparametric estimation of copula functions for dependence modelling. The Canadian Journal of Statistics **35**, 265 – 282

[6] Chen, Z. (2012). A flexible copula model for bivariate survival data. PhD Thesis, University of Rochester

[7] Dabrowska, D. M. (1988). Kaplan-Meier estimates on the plane. The Annals of statistics **16**, 1475-1489

[8] Deheuvels, P. (1979). La fonction de dépendance empirique et ses proprietès. Un test non paramétrique d'indépendance. Acad. R.Belg., Bull. CI. Sci.5, Série **65**, 274–292

[9] Dobrić, J. and Schmid, F. (2007). A goodness of fit test for copulas based on Rosenblatt's transformation. Computational Statistics and data Analysis **51**, 4633–4642

[10] Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. Journal of Multivariate Analysis **95**, 119–152

[11] Fermanian, J.-D. (2013). An overview of goodness-of-fit tests problems for copulas. In: Jaworski, P.; Durante, F.; Härdle, W. (eds.). Copulae in mathematical and qualitative finance. Lecture notes in Statistics, Springer, Berlin, 61 –69

[12] Genest, C. and Rivest, L.-P. (1993) Statistical inference procedures for bivariate Archimedean copulas. Journal of American Statistics Association **88**, 1034–1043

[13] Genest, C; Queessy, J.-F. and Rémillard, B. (2006). Goodness-of-fit for copula models based on the probability integral transform. Scandinavian Journal of Statistics **33**, 337–366

[14] Genest, C. and Rémillard, B. (2008).Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. Ann. Henri Poincaré **44**, 1096–1127

[15] Ghosh, B. and Huang W. (1991). The power and optimal kernel of the Bickel-Rosenblatt test for goodness of fit. Ann. Statist. **19**, 999–1009

[16] Gribkova, S. and Lopez, O. (2015). Nonparametric copula estimation under bivariate censoring. Scandinavian Journal of Statistics **42**, 925–946

[17] Liero, H.; Läuter, H. and Konakov, V. (1998). Nonparametric versus parametric goodness of fit. Statistics **31**, 115–149

[18] Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. Biometrika **80**, 573–581

[19] Prentice, R. L., Zoe Moodie, F. and Wu, J. (2004). Hazard-based nonparametric survivor function estimation. Journal of the Royal Statistical Society B **66**, 305–319

[20] Pruitt, R. G. (1990). Strong consistency of self consistent estimators: general theory and an application to bivariate survival analysis. Technical Report 543, University of Minnesota, School Statistics.

[21] Pruitt, R. G. (1991). On negative mass assigned by the bivariate Kaplan-Meier estimator. The Annals of Statistics **9**, 879–885

[22] Rosenblatt, M. (1952). Remarks on a multivariate transformation. The Annals of Mathematical Statistics **23**, 470–472

[23] Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Annals Statistics **3**, 1– 14

[24] Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. Bernoulli **18**, 764–782

[25] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. Publications de l' Institut de statistique de l'Université de Paris **8**, 229–231

[26] van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. The Annals of Statistics **24**, 596–627

[27] Wang, W. and Wells, M.T. (2000) Model selection and semiparametric inference for bivariate failure-time data. Journal of the American Statistical Association **95**, 62–72

[28] Wang, A. (2010). Goodness-of-fit tests for archimedean copula models. Statistica Sinica **20**, 441–453

[29] Yilmaz, Y. E. (2009). Estimation and goodness of fit for multivariate survival models based on copulas. Thesis, University Waterloo

On Survival Categorical Methods from Grouped Right Censored Data with Unobserved Status after Censoring

Irina Yu. Malova[1,3] and Sergey V. Malov[2,3]

[1] Higher School of Technology and Energy, St. Petersburg State University of Industrial Technologies and Design, St. Petersburg, Russia

[2] *Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia*

[3] *Department of Mathematics, St. Petersburg Electrotechnical University (LETI), St. Petersburg, Russia*

e-mail: malovs@sm14820.spb.edu

Abstract

We develop methods of categorical data analysis applicable for a number of survival experimental designs. The interval right censored data model with an unobserved status of individual after censoring as an alternative to the life table survival data model considered. We create maximum likelihood estimator for grouped failure time distribution by a sample from the interval right censored data and obtain its consistency and asymptotic normality. Moreover, we create Wald's type tests for categorical homogeneity hypothesis based on the interval right censored data.

Keywords: survival data, right censoring, independent censoring, hypothesis testing, categorical tests, contrasts, Wald's test, actuarial estimator, Kaplan–Meier estimator.

Introduction

Common experimental design in epidemiology is to screen a cohort of individuals for disease endpoints during a time interval. Study participants are disease free at the baseline (time point zero) and they are followed up until a failure time or missing at follow up. Right censored survival data model is widely applicable in practice in spite of in most cases the event times (failure or censoring) are not observed precisely and the investigator observes time interval containing a failure time for each of not missed at follow up individuals having symptoms of disease at the endpoint. Under fixed observation times the life table survival data model describes more accurately the real data, but there is no consistent estimator of the distribution function at the observation times in this case. The life table survival data model assumes that the status of individual failed before censoring is observed even if the individual was missed at follow up until the following observation time, that can be failed under some experimental designs. In this work we investigate the case of unobserved status of any individual after censoring. In other words, the individual failed and missed at follow up between two successive observation times is assumed to be missed even if it was failed before missing at follow up.

Let T be a failure time or time of appearance symptoms of disease. Distribution of T depends on covariate z and can be given by a distribution function $F_z(x) = P(T \leq x|z)$ or by a survival function $S_z(x) = 1 - F_z(x)$. Assume that the covariate z is a categorical variable having d levels. We are interesting to compare distributions of failure time under different values of covariate. Let $\gamma_T = \min_{i \in 1,\ldots,d} \sup\{x : F_i(x) < 1\}$. The null hypothesis is

$$H_0^* : S_1(x) = \ldots = S_d(x) \quad \text{for all} \quad x \in [0, \gamma_T].$$

To formulate the problem in terms of categorical data analysis set $0 < t_1 < \ldots < t_s < \gamma_T$. Consider $p_{1|z} = P(T \in [0, t_j]|z)$ and $p_{j|z} = P(T \in]t_{j-1}, t_j]|z)$, $j = 2, \ldots, s+1$, where $t_{s+1} = \infty$. We formulate weaker null hypothesis

$$H_0 : p_{j|1} = \ldots = p_{j|d} \quad \text{for all} \quad j = 1, \ldots, s$$

or, using the survival function,

$$H_0 : S_1(t_j) = \ldots = S_d(t_j) \quad \text{for all} \quad j = 1, \ldots, s. \tag{1}$$

It is clear that H_0 is closing to H_0^* if $p_{j|z} \to 0$ as $s \to \infty$.

The contingency table experimental design is universal for wide range of applications. There are examples of application of classical categorical tests in right censored data case [8, 16, 17]. Limitations on application of classical categorical tests for right censored survival data are discussed in [12].

The right censored survival data model is commonly used for such kind of experimental design. Categorical tests for grouped right censored survival data are presented widely in literature. Likelihood ratio tests with grouped right censored survival data are investigated in [18]. A chi-square type test for survival data due to Habib & Thomas [7]. Advanced properties of chi-square type tests are obtained in [1, 2]. Hollander & Pena [10] consider chi-square test statistic for simple null hypotheses in censored data case and investigate its limit behavior. Categorical tests on independence for survival data based on contrasts obtained from limit theorems for Nelson–Aalen and Kaplan–Meier estimators are given in [12]. Exact event (failure or censoring) times are required for all these approaches.

Let T and U be independent failure and censoring times respectively. Right censored observation is given by the event time $X = T \wedge U$ and the indicator $\delta = \mathbb{1}_{\{T \leq U\}}$. In practice often the event time is not observed exactly, but at some fixed observation times $0 = x_0 < x_1 < \ldots < x_s < \infty$, which divide the time line into a fixed number of r disjoint finite intervals $I_1, \ldots, I_s$: $I_1 = [0, x_1]$, $I_k =]x_{k-1}, x_k]$, $k = 2, \ldots, s$, and the infinite interval $I_0 =]x_s, \infty[$.

The observed status of participant after censoring leads to life table survival data [3, 5, 6]. The life table (or actuarial) estimator was rather famous in early survival analysis although it is inconsistent. Consistency conditions of the actuarial estimator and its extensions are investigated in [4, 14]. Survival categorical tests for life table survival data are given in [13].

In the case of unobserved status of participant after censoring any observation brings more information on distribution of censoring time. Any single observation can be given as a set of binary (dummy) variables $C_0 = \mathbb{1}_{\{U \in I_k\}}$, $C_k = \mathbb{1}_{\{T_i > x_{j-1}; U_i \in I_k\}}$ and $D_k = \mathbb{1}_{\{T \in I_k; U > x_k\}}$, $k = 1, \ldots, s$. Estimators for the corresponding parameters

can be obtained from a sample as the sample counts and, in contrast of the life table data, there is a consistent estimator of the survival function of failure time in the observation points.

In this work we consider grouped right censored survival data with unobserved status of individuals after censoring. The maximum likelihood estimator of survival function at the observation times and its asymptotic properties are obtained in Section 1 and Wald type tests for homogeneity are created in Section 2.

1 Grouped right censored data

We consider grouped right censored data with the unobserved status of participant after censoring. Let the failure time T and the censoring time U be independent random variables having distribution functions F and G respectively; $\gamma_T = \sup\{x : F(x) < 1\}$ and $\gamma_G = \sup\{x : G(x) < 1\}$. Assume that (T_i, U_i) be a sample from the distribution of (T, U); (X_i, δ_i), where $X_i = T_i \wedge U_i$ and $\delta_i = \mathbb{1}_{\{T_i \le U_i\}}$, $i = 1, \dots, n$, be the right-censored survival data. Let $0 = x_0 < x_1 < \ldots < x_s < \gamma_T$ be fixed time points (observation times); $I_1 = [0, x_1]$ and $I_j =]x_{j-1}, x_j]$. The survival function $S \equiv 1 - F$ in the time points $x_1, \dots, x_s$ is defined as follows:

$$S(x_k) = \prod_{i=1}^{k} (1 - \lambda_i),$$

where $\lambda_i = (S(x_{i-1}) - S(x_i))/S(x_{i-1})$. The observed data is given by $(D_{i1}, \dots, D_{is})$ and $(C_{i1}, \dots, C_{is})$, where $D_{ij} = \mathbb{1}_{\{T_i \in I_j; U_i > x_j\}}$, $C_{ij} = \mathbb{1}_{\{T_i > x_{j-1}; U_i \in I_j\}}$, $j = 1, \dots, s$, $i = 1, \dots, n$. Remark that $\sum_{j=1}^{s} D_{ij} + \sum_{j=1}^{s} C_{ij} \in \{0, 1\}$.

Denote $\eta_j = (F(x_j) - F(x_{j-1}))(1 - G(x_j))$ and $\nu_j = (1 - F(x_j))(G(x_j) - G(x_{j-1}))$, $j = 1, \dots, r$. The observed data can be reduced to the sample counts $(D_1, \dots, D_s, C_1, \dots, C_s, K)$, having the multinomial distribution with parameter $(\eta_1, \dots, \eta_s, \nu_1, \dots, \nu_s, \kappa_0)$, where $\kappa_0 = 1 - \sum_j (\eta_j + \nu_j)$, $D_j = \sum_{i=1}^{n} D_{ij}$, $C_j = \sum_{i=1}^{n} C_{ij}$, $j = 1, \dots, s$ and $K = n - \sum_{j=1}^{s} (C_j + D_j)$. Introduce also $Y_0 = n$, $Y_k = n - \sum_{j=1}^{k} (C_j + D_j)$ is the number of individuals at risk (non censored and non failed) at time x_k, and $Y_k^* = n - \sum_{j=1}^{k-1} (C_j + D_j) - C_k$, $k = 1, \dots, s$

The maximum likelihood estimate $\hat{\boldsymbol{\mu}} = (\hat{\eta}_1, \dots, \hat{\eta}_s, \hat{\nu}_1, \dots, \nu_s)$ of the parameter $\boldsymbol{\mu} = (\eta_1, \dots, \eta_s, \nu_1, \dots, \nu_s)'$ is given directly by the sample method $\hat{\eta}_i = D_i/n$ and $\hat{\nu}_i = C_i/n$, $i = 1, \dots, s$. It is well known that the estimators are asymptotically normal with the covariance matrix

$$\mathbf{R}_{\eta\nu} = \mathrm{diag}(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}',$$

where $\mathrm{diag}(\boldsymbol{\mu})$ is the diagonal matrix with $(\eta_1, \dots, \eta_s, \nu_1, \dots, \nu_s)$ on the diagonal.

Denote $\theta_j = F(x_j)$ and $\psi_j = G(x_j)$, $j = 1, \dots, r$. The parameter of interest $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ can be recovered from the parameter $\boldsymbol{\mu}$ along with the parameter $\boldsymbol{\psi} = (\psi_1, \dots, \psi_r)$ by using $\nu_1 = G(x_1)$ and the following recurrence formulas

$$\nu_j/\eta_{j-1} = (G_j^*(x_j))(1 - F_j^*(x_{j-1}))/F_j^*(x_{j-1})$$

and

$$\eta_j/\nu_j = (F^*_{j-1}(x_j))(1 - G^*_j(x_{j-1}))/F^*_j(x_{j-1}),$$

where $F^*_j(x) = (F(x) - F(x_j))/(1 - F(x_j))$ and $G^*_j(x) = (G(x) - G(x_j))/(1 - G(x_j))$ are truncated distribution functions of T and U respectively, $j = 1, \ldots, r$. Note that $\lambda_k = F^*_{k-1}(x_k)$. Introduce also $\lambda^G_k = G^*_{k-1}(x_k)$. The explicit formulas follow immediately $\lambda_k = \eta_k/(1-\eta^-_k)$ and $\lambda^G_k = \nu_k/(1-\nu^-_k)$, where $\eta^-_k = \sum_{j=1}^{k-1}(\nu_j+\eta_j)+\nu_k$ and $\nu^-_k = \sum_{j=1}^{k-1}(\nu_j + \eta_j)$. Then the parameters of interest can be recovered in the following way

$$S(x_k) = \prod_{j=1}^{k}\Big(1 - \eta_k/(1 - \eta^-_k)\Big) \quad \text{and} \quad \overline{G}(x_k) = \prod_{j=1}^{k}\Big(1 - \nu_k/(1 - \nu^-_k)\Big),$$

$k = 1, \ldots, r$, where $\overline{G} \equiv 1 - G$. Then, $\hat{\boldsymbol{\theta}} = (\hat\theta_1, \ldots, \hat\theta_d)'$ with

$$\hat\theta_k = \hat S(x_k) = \textstyle\prod_{j=1}^{k}(1 - \frac{D_k}{Y^*_k}),$$

$k = 1, \ldots, d$, is the nonparametric maximum likelihood estimator for $\boldsymbol{\theta}$.

Using delta-method we obtain for all $i = 1, \ldots, d$

$$\sqrt{n}(\hat\lambda_i - \lambda_i) = \sqrt{n}(\hat\eta_i - \eta_i)/(1 - \eta^-_i) + \sqrt{n}(\hat\eta^-_i - \eta^-_i)\eta_i/(1 - \eta^-_i)^2 + O_P(1).$$

Note that $(\eta_1, \ldots, \eta_d, \eta^-_1, \ldots, \eta^-_d)' = \mathbf{Q}\boldsymbol{\mu}$, where

$$\mathbf{Q} = \begin{pmatrix} \mathbf{I}_d & \mathbf{0} \\ \mathbf{T}_d - \mathbf{I}_d & \mathbf{T}_d \end{pmatrix}$$

where $\mathbf{I}_d$ is the identity matrix; $\mathbf{0}$ is the matrix of zeroes;

$$\mathbf{T}_d = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 1 & 1 & \ldots & 0 & 0 \\ 1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \ldots & 1 & 1 \end{pmatrix}$$

is the lower triangular matrix. Therefore, $\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) = \sqrt{n}\,\mathbf{J}\mathbf{Q}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) + O_P(1) = \sqrt{n}\,\widehat{\mathbf{J}}\mathbf{Q}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) + O_P(1)$, where $\mathbf{J} = (\mathbf{J}_1, \mathbf{J}_2)$, $\mathbf{J}_1 = \operatorname{diag}\big(1/(1 - \boldsymbol{\eta}^-)\big)$, $\mathbf{J}_2 = \big(\boldsymbol{\eta}/(1-\boldsymbol{\eta}^-)^2\big)$ are the diagonal matrices with the elements $1/(1-\eta^-_1), \ldots, 1/(1-\eta^-_d)$ and $\eta_1/(1 - \eta^-_1)^2, \ldots, \eta_d/(1 - \eta^-_d)^2$ respectively. Hence,

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \Rightarrow N(0, \boldsymbol{\Upsilon}), \tag{2}$$

where $\boldsymbol{\Upsilon} = \mathbf{J}\mathbf{Q}\mathbf{R}_{\nu\eta}\mathbf{Q}\mathbf{J}'$, that implies convergence for cumulative intensities

$$\sqrt{n}(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}) \Rightarrow N(0, \mathbf{T}_d\boldsymbol{\Upsilon}\mathbf{T}'_d),$$

where $\mathbf{\Lambda} = (\Lambda(x_1), \ldots, \Lambda(x_d))$ and $\Lambda(x_i) = \sum_{j=1}^{i} \lambda_j$. On the other hand, using deltha-method we obtain that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{n}\,\mathbf{B}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) + O_P(1)$ with $\mathbf{B} = \operatorname{diag}\big(1/(1-\boldsymbol{\lambda})\big)\,\mathbf{T}_d\,\operatorname{diag}\big(\boldsymbol{\theta}\big)$. Therefore,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \Rightarrow N(0, \mathbf{B}\boldsymbol{\Upsilon}\mathbf{B}'). \tag{3}$$

The asymptotic variance estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{B}}\widehat{\boldsymbol{\Upsilon}}\widehat{\mathbf{B}}'$ can be obtained by substitution the corresponding estimators insead of true values of parameters to $\mathbf{R}_{\eta\nu}$, $\mathbf{J}$ and $\mathbf{B}$ matrices. Particularly, the estimator for $1/(1-\lambda_k)$ can be given as Y_k^*/Y_{k+1}, where $Y_{k+1} = n - \sum_{i=1}^{k+1}(D_i + C_i)$ be the number of individuals at risk (non censored and not failed) at the observation time x_{k+1}, $k = 1, \ldots, s-1$

2 Categorical survival tests

Let $0 < t_1 < \ldots < t_s < \gamma_T$ be the observation times. We pose the problem of testing categorical homogeneity hypothesis (1) by d samples from grouped right censored data with unobserved status after censoring; F_z and G_z be distribution functions of failure time and censoring time respectively are such that $S_i(t_{k_1}) - S_i(t_{k_2}) > 0$ for all $1 \le k_2 < k_1 < \gamma_T$, $i = \{1, \ldots, d\}$. Assume for simplicity that the set of breakpoints $\{t_1, \ldots, t_s\}$ coincides with the set of observation times $\{x_1, \ldots, x_s\}$. Generalization of tests below to the case $\{t_1, \ldots, t_s\} \subseteq \{x_1, \ldots, x_{s'}\}$, $s \le s'$, is trivial.

Introduce the parameters $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{is})'$, where $\theta_{zj} = S_z(t_j)$. The hypothesis (1) can be rewritten as follows:

$$\widetilde{H}_0 : \theta_{j1} = \ldots = \theta_{jd} \quad \text{for all} \quad j = 1, \ldots, s.$$

Note that (3) implies weak convergence

$$\sqrt{n_i}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \Rightarrow N(0, \boldsymbol{\Sigma}_i) \quad \text{as} \quad n_i \to \infty$$

with $\Sigma_i = \mathbf{B}_i\boldsymbol{\Upsilon}_i\mathbf{B}_i'$, $\boldsymbol{\Upsilon}_i = \mathbf{J}_i\mathbf{Q}\mathbf{R}_i\mathbf{Q}'\mathbf{J}_i$, $\mathbf{J}_i = (\mathbf{J}_{1i}, \mathbf{J}_{2i})$, $\mathbf{J}_{1i} = \operatorname{diag}(1 - \boldsymbol{\eta}_i^-)$, $\mathbf{J}_{2i} = \operatorname{diag}\big(\boldsymbol{\eta}_i/(1-\boldsymbol{\eta}_i^-)^2\big)$, $\boldsymbol{\eta}_i$, $\boldsymbol{\nu}_i$ and $\boldsymbol{\eta}_i^-$ are the expected frequencies from i-th sample, and $\mathbf{R}_i$ is the covariance matrix of $(\boldsymbol{\eta}_i, \boldsymbol{\nu}_i)$. Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \Rightarrow N(0, \boldsymbol{\Sigma}), \tag{4}$$

and $\boldsymbol{\Sigma} = \operatorname{diag}(l_1\boldsymbol{\Sigma}_1, \ldots, l_n\boldsymbol{\Sigma}_d)$ is the block-diagonal matrix, $l_i = n/n_i$ for $i = 1, \ldots, d$.

The consistent estimator $\widehat{\boldsymbol{\Sigma}} = \operatorname{diag}(l_1\widehat{\boldsymbol{\Sigma}}_1, \ldots, l_n\widehat{\boldsymbol{\Sigma}}_d)$, where $\widehat{\boldsymbol{\Sigma}}_i = \widehat{\mathbf{B}}_i\widehat{\boldsymbol{\Upsilon}}_i\widehat{\mathbf{B}}_i'$ with $\widehat{\boldsymbol{\Upsilon}}_i = \widehat{\mathbf{J}}_i\mathbf{Q}\widehat{\mathbf{R}}_i\mathbf{Q}'\widehat{\mathbf{J}}_i$ can be obtained by substitution of the observed frequencies instead of the expected frequencies from the corresponding sample.

Let $\boldsymbol{\psi}_i = \mathbb{A}\boldsymbol{\theta}_i$, where $\mathbb{A} = \|a_{ij}\|$ is $(d-1) \times d$ -matrix of linearly independent contrasts, i. e. $\sum_{j=1}^{d} a_{ij} = 0$ for all i and $\mathbf{rk}(\mathbf{A}) = d - 1$. Then, $\widetilde{H}_0$ can be written in terms of contrasts

$$\widetilde{H}_0 : \boldsymbol{\psi}_1 = \ldots = \boldsymbol{\psi}_{d-1} = 0.$$

Associate with any a_{ij} the diagonal matrix $\mathbf{A}_{ij} = a_{ij}\mathbf{I}_s$, where $\mathbf{I}_s$ is the identity matrix of size s and construct the matrix $\mathbf{A}$ of size $(d-1)s \times ds$ from blocks $\mathbf{A}_{ij}$ in appropriate order. It is obviously that $\mathbf{A}$ is a matrix of linearly independent contrasts and the null hypothesis can be rewritten in vector form

$$\widetilde{H}_0 : \mathbf{A}\boldsymbol{\theta} = 0.$$

Taking into account (4) we obtain that under null hypothesis

$$n\,\hat{\boldsymbol{\theta}}'\widehat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{\theta}} \Rightarrow \chi^2_{(d-1)s},$$

where $\widehat{\boldsymbol{\Omega}} = \mathbf{A}'(\mathbf{A}\widehat{\boldsymbol{\Sigma}}\mathbf{A}')^{-1}\mathbf{A}$, that implies Wald type test immediately.

Analogous tests can be obtained for the null hypothesis

$$\widetilde{H}_0^* : \lambda_{j|1} = \ldots = \lambda_{j|d} \quad \text{for all} \quad j = 1, \ldots, s,$$

using convergence in (2).

References

[1] Bagdonavičius, V., Levuliene, R, Nikulin, M. S. & Tran, Q. X. (2012). On Chi-square Type Tests and Their Applications in Survival Analysis and Reliability. *Zapiski nauchnih seminarov POMI*. Vol. **408**, pp. 43–61.

[2] Bagdonavičius, V. & Nikulin, M. S. (2011). Chi-squared Goodness-of-fit Test for Right Censored Data. *International Journal of Applied Mathematics and Statistics*. Vol. **24**, pp. 30–50.

[3] Berkson and Gage (1950). Calculation of survival rates for cancer. *Proceedings of Staff Meetings of the Mayo Clinic*, Vol. **25**, pp. 270–286.

[4] Breslow, N. and Crowley, J. (1974) A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *The Annals of Statistics*, Vol. **2**(3), pp. 437–453.

[5] Cutler and Ederer (1958). Maximum utilization of the life table method in analyzing survival. *J. Chron.Dis.*, . Vol. **8**, pp. 699–712.

[6] Gehan (1969). Estimating survival function from the life table. *J.Chron.Dis.*, Vol. **21**, pp. 629–644.

[7] Habib & Thomas (1986). Chi-Square Goodness-of-Fit Tests for Randomly Censored Data. *The Annals of Statistics*, Vol. **14**(2), pp. 759–765.

[8] Hendrickson, S.L., Lautenberger, J.A., Chinn, L.W., Malasky, M., Sezgin, E., Kingsley, L.A., Goedert, J.J., Kirk, G.D., Gomperts, E.D., Buchbinder, S.P., Troyer, J.L. and O'Brien, S.J. (2010). Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression *PLoS ONE*, Vol. **5**(9), art. no. e12862, pp. 1-8

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

[10] Hollander & Pena (1992). A Chi-Squared Goodness-of-Fit Test for Randomly Censored Data. *Journal of the American Statistical Association*, Vol. **87**(418), pp. 458-463.

[11] Malov S.V. (2017). Life Table Estimator Revisited. *Submitted to Lifetime Data Analysis.*

[12] Malov S.V. & O'Brien S.J. (2013). On Survival Categorical Methods with Applications in Epidemiology and AIDS Research. In *Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control.* Proceedings of the International Workshop AMSA'13 (Novosibirsk, September 25-27, 2013), pp. 173–180.

[13] Malov S.V. & O'Brien S.J. (2015). On Survival Categorical Methods Based on an Extended Actuarial Estimator. In *Applied Methods of Statistical Analysis. Nonparametric Approach.* Proceedings of the International Workshop AMSA'15 (Novosibirsk & Belokurikha, September 14-19, 2015), pp. 147–156.

[14] Malov S.V. & O'Brien S.J. (2015). Life Table Estimator Revisited. *Submitted to Communication in Statistics – Theory and Methods.*

[15] O'Brien, S.J., Hendrickson, S.L. (2013). Host genomic influences on HIV/AIDS. *Genome Biology*, Vol. **14**, p. 201.

[16] O'Brien, S.J., Nelson, G.W., Winkler, C.A., Smith, M.W. (2000). Polygenic and multifactorial disease gene association in man: Lessons from AIDS. *Annual Review of Genetics*, Vol. **34**, pp. 563–591.

[17] Shin, H.D., Winkler, C., Stephens, J.C., Bream, J., Young, H., Goedert, J.J., O'Brien, T.R., Buchbinder, S.f, Giorgi, J.h, Rinaldo, C.i, Donfield, S.g, Willoughby, A.j, O'Brien, S.J. , and Smith, M.W. (2000) Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. **97**(26), pp. 14467-14472.

[18] Turnbull & Weiss (1978). A Likelihood Ratio Statistic for Testing Goodness of Fit with Randomly Censored Data. *Biometrics*, Vol. **34**(3), pp. 367–375.

Asymptotic Efficiency of Bayesian Type Estimates for Unknown Parameter in Competing Risks Model under Random Censoring by Nonobserving Intervals

Abdushukurov A.A. and Nurmukhamedova N.S.
National University of Uzbekistan, Tashkent, Uzbekistan
e-mail: a_abdushukurov@rambler.ru, rasulova_nargiza@mail.ru

Abstract

In this paper we consider locally asymptotical normality of likelihood ratio statistics in competing risks model under random censoring by nonobserving intervals. Using the property of locally asymptotical normality we investigate asymptotical properties of Bayesian type estimates for unknown parameter and prove its asymptotic efficiency.

Keywords: competing risks; random censoring; likelihood ratio statistics; locally asymptotically normality; asymptotic efficiency.

The likelihood ratio statistics (LRS) plays an important role in decision theory. For example, while testing the simple hypothesis H_0 against complicated alternative H_1 with unknown law of distributions the criterions constructed based on LRS, according to Newmann-Pirson's lemma, are uniformly more powerful for any size n of observations (see Leman,1964; Rusas, 1975). Here appears some useful for estimation theory and hypothesis testing asymptotical properties of LRS, when alternative H_1 depends on n and close to H_0, i.e. $H_1 = H_{1n} \to H_0$ for $n \to \infty$. This property is local asymptotically normality (LAN) of LRS. There are set of papers on investigations of the LAN for LRS and it's applications in statistics. The most remarkable works are Le Cam (1960), Hajek (1972), Rusas (1975), Ibragimov, Khas'minskii (1979), which shown that LAN allows the development of asymptotical theory for most maximum likelihood and Bayesian type estimators and proved the contiguality properties of the family of probability distributions. In papers Abdushukurov (2011), Abdushukurov, Nurmuhamedova (2011, 2012a, 2012b, 2012c, 2013, 2014) established properties of LAN for LRS in competing risks model (CRM) under random censoring of observations on the right and both sides. This paper includes investigations of LAN for LRS in CRM under random censoring by nonobserving intervals. Using the property of LAN, we investigate asymptotical properties of Bayesian type estimates for unknown parameter and prove its asymptotic efficiency.

In CRM it's interesting to investigate the random variable (r.v.) X with values from measurable space $(\mathcal{X}, \mathcal{B})$ and events $(A^{(1)}, ..., A^{(k)})$ forming the complete group, where k is fixed. In practice, r.v. X obvious means survival or a reliability time of some object (individual, physical system), exposing to k competing risks and getting out of work in case one of the events $\{A^{(i)},\ i = 1, ..., k\}$ occurs. In such case pairs $\{(X, A^{(i)}),\ i = 1, ..., k\}$ define the time and reason the object getting out of work (see more about CRM on Langberg, Proshan, Quinzi(1978), Burke, Csörgő, Horváth (1981), Abdushukurov(2011)). While repeat the experiment where the aggregate

$(X, A^{(1)}, ..., A^{(k)})$ is observed under homogenous conditions we took the sequence of copies $\{(X_j, A_j^{(1)}, ..., A_j^{(k)}),\ \ j \geq 1\}$. Let $\delta_j^{(i)} = I(A_j^{(i)})$ is indicator of events $A_j^{(i)}$. Every vector $\zeta_j = (X_j, \delta_j^{(1)}, ..., \delta_j^{(k)})$ induces statistical model with sample space $\mathcal{Y} = \mathcal{X} \times \{0,1\}^{(k)} = \mathcal{X} \times \{0,1\} \times ... \times \{0,1\}$ and σ-algebra $\mathcal{C}$ of the sets such as $B \times D_1 \times ... \times D_k$, where $B \in \mathcal{B}$ and $D_i \subset \{0,1\}$, $i = 1, ..., k$. We suppose that distribution of vector ζ_j on $(\mathcal{Y}, \mathcal{C})$ depends on unknown parameter $\theta = (\theta_1, ..., \theta_s) \in \Theta$:

$$Q_\theta^*(B \times D_1 \times ... \times D_k) = P_\theta(X_1 \in B, \delta_1^{(1)} \in D_1, ..., \delta_1^{(k)} \in D_k), \tag{1}$$

where Θ is an open set in R^s. Let distribution (1) is absolutely continuous with respect to σ -finite measure $\nu(x) = \mu(x) \times \varepsilon_1 \times ... \times \varepsilon_k$, where μ is Lebesque measure on R and ε_i are counting measures, concentrated at points $y^{(i)} \in \{0,1\}$, $i = \overline{1,k}$. Later we consider such statistical scheme, in according which aggregate $(X_j, A_j^{(1)}, ..., A_j^{(k)})$ is nonobservable if r.v. X_j fall in interval $[Y_{1j}, Y_{2j}]$, where $\{(Y_{1j}, Y_{2j}),\ \ j \geq 1\}$ is the sequence of independent and identically distributed (i.i.d) random vectors with unknown distribution $G(u,v)$, $(u,v) \in R^2$ (possibly implicitly depending from θ). Here the aggregates $(X_j, A_j^{(1)}, ..., A_j^{(k)})$ and pairs (Y_{1j}, Y_{2j}) assumed to be independent and $P_\theta(Y_{1j} \leq Y_{2j}) = 1$ for every $j \geq 1$. Really they corresponds, for example, the experiments, in those observation for object j with life time X_j might be stopped in random moment Y_{1j} and renew in random moment Y_{2j}. Such statistical model we call as CRM under random censoring by nonobserving intervals. In this case, instead of events $(A_j^{(1)}, ..., A_j^{(k)})$ we observe the events $(D_j^{(0)}, D_j^{(1)}, ..., D_j^{(k)})$, where $D_j^{(0)} = \{\omega : Y_{1j}(\omega) \leq X_j(\omega) \leq Y_{2j}(\omega)\}$ and $D_j^{(i)} = A_j^{(i)} \cap (\{\omega : X_j(\omega) < Y_{1j}(\omega)\} \cup \{\omega : X_j(\omega) > Y_{2j}(\omega)\})$, $i = 1, ..., k$. Let $\Delta_j^{(i)} = I(D_j^{(i)})$, $i = 0, 1, ..., k$ and $w_j = \varepsilon_{1j} + \varepsilon_{2j}$, where $\varepsilon_{1j} = I(X_j < Y_{1j})$ and $\varepsilon_{2j} = I(X_j > Y_{2j})$. It's obviously, that $\Delta_j^{(0)} = 1 - w_j$ and $\Delta_j^{(i)} = w_j \delta_j^{(i)}$. In CRM we interested in properties of pairs $\{(X_j, A_j^{(i)}),\ i = \overline{1,k}\}$ and consequently consider the subdistributions

$$Q_{i\theta}(B) = Q_\theta^*(B \times \{0\} \times ... \times \{0\} \times \{1\} \times \{0\} \times ... \times \{0\}),\ i = 1, ..., k, \tag{2}$$

produced from (1) when $D_i = \{1\}$ and $D_l = \{0\}$, $i \neq l$, $l = 1, ..., k$. Let $Q_\theta(B) = \sum\limits_{i=1}^{k} Q_{i\theta}(B)$. By $h^{(i)}$ and h we define the densities of subdistributions $Q_{i\theta}$ and Q_θ:

$$Q_{i\theta}(B) = \int\limits_B h^{(i)}(x;\theta)\mu(dx),\ \ i = 1, ..., k, \quad Q_\theta(B) = \int\limits_B h(x;\theta)\mu(dx), \tag{3}$$

where $h = h^{(1)} + ... + h^{(k)}$. For $B = (-\infty; x]$ we put $Q_{i\theta}((-\infty; x]) = H^{(i)}(x;\theta)$, $i = \overline{1,k}$ and $Q_\theta((-\infty; x]) = H(x;\theta)$. Later, we define the cumulative hazard functions (c.h.f.) of the pairs $(X, A^{(i)})$:

$$\Lambda^{(i)}(x;\theta) = \int\limits_{(-\infty;x]} \lim_{\Delta \downarrow 0} P_\theta(t < X \leq t + \Delta,\ A^{(i)}/X > t)\mu(dt) =$$

$$= \int\limits_{(-\infty;x]} \frac{dH^{(i)}(t;\theta)}{1-H(t;\theta)}, \;\; i=1,...,k, \;\; x \in R^1. \tag{4}$$

Then c.h.f., corresponding the r.v. X is $\Lambda(x;\theta) = \sum\limits_{i=1}^{k} \Lambda^{(i)}(x;\theta)$. In CRM exponential hazard functionals $F^{(i)}(x;\theta) = 1-\exp\{-\Lambda^{(i)}(x;\theta)\}$, $i = \overline{1,k}$, describe the distribution of the pairs $(X, A^{(i)})$ through the i-th risk. In light of equality $\Lambda(x;\theta) = -\log(1 - H(x;\theta))$, we have

$$1 - H(x;\theta) = P_\theta(X > x) = \prod_{i=1}^{k} (1 - F^{(i)}(x;\theta)). \tag{5}$$

Let's define density $f^{(i)}(x;\theta) = \frac{\partial}{\partial x} F^{(i)}(x;\theta)$, $i = \overline{1,k}$. Then the hazard rate density for i-th risk is $f^{(i)}/(1 - F^{(i)})$. In other hand, by formulas (3-5) for every $(x;\theta) \in R^1 \times \Theta$ and $i = 1, ..., k$:

$$\frac{f^{(i)}(x;\theta)}{1-F^{(i)}(x;\theta)} = \frac{h^{(i)}(x;\theta)}{1-H(x;\theta)},$$

i.e.

$$h^{(i)}(x;\theta) = f^{(i)}(x;\theta) \prod_{j=1;j\neq i}^{k} (1 - F^{(j)}(x;\theta)). \tag{6}$$

Let for n-th stage of experiments to observation available the sample $\mathbb{Z}^{(n)} = (Z_1, ..., Z_n)$, where $Z_j = w_j X_j + (1 - w_j)[Y_{1j}, Y_{2j}]$, it means that every observation Z_j is r.v. X_j(when $w_j = 1$) or an interval $[Y_{1j}, Y_{2j}]$(when $w_j = 0$). As $p(z;\theta)$ we define the density of one observation without the multipliers depending of unknown nuisance distribution G. Then according the representation (6), we have the following "truncated" likelihood function of sample $\mathbb{Z}^{(n)}$:

$$p_n(\mathbb{Z}^{(n)};\theta) = \prod_{m=1}^{n} p(Z_m;\theta) = \prod_{m=1}^{n} \left\{ \left[\prod_{i=1}^{k} \left[f^{(i)}(X_m;\theta) \prod_{j=1;j\neq i}^{k} (1 - F^{(j)}(X_m;\theta)) \right]^{\delta_m^{(i)}} \right]^{w_m} \cdot \right.$$

$$\left. \cdot [H(Y_{2m};\theta) - H(Y_{1m};\theta)]^{1-w_m} \right\} =$$

$$= \prod_{m=1}^{n} \left\{ \left[\prod_{i=1}^{k} \left[h^{(i)}(X_m;\theta) \right]^{\delta_m^{(i)}} \right]^{w_m} [H(Y_{2m};\theta) - H(Y_{1m};\theta)]^{1-w_m} \right\}. \tag{7}$$

Let for every $u \in R^s$, $\theta + n^{-1/2}u = \Psi_n(u;\theta) \in \Theta$ and $\widetilde{Q}_\theta^{(n)}$ is distribution induced by the sample $\mathbb{Z}^{(n)}$. Then we have LRS of the model as

$$L_{n,\theta}(u) = d\widetilde{Q}^{(n)}_{\Psi_n(u;\theta)}(\mathbb{Z}^{(n)}) / d\widetilde{Q}^{(n)}_\theta(\mathbb{Z}^{(n)}) = \frac{p_n(\mathbb{Z}^{(n)}; \Psi_n(u;\theta))}{p_n(\mathbb{Z}^{(n)};\theta)} =$$

$$= \prod_{m=1}^{n} \left\{ \left[\prod_{i=1}^{k} \left[\frac{h^{(i)}(X_m; \Psi_n(u;\theta))}{h^{(i)}(X_m;\theta)} \right]^{\delta_m^{(i)}} \right]^{w_m} \left[\frac{H(Y_{2m}; \Psi_n(u;\theta)) - H(Y_{1m}; \Psi_n(u;\theta)}{H(Y_{2m};\theta) - H(Y_{1m};\theta)} \right]^{1-w_m} \right\}. \tag{8}$$

Let $N^{(i)} = \{x : h^{(i)}(x;\theta) > 0\}$ and $N = \bigcap\limits_{i=1}^{k} N^{(i)}$. Later we need for some regularity conditions:

(C1) The supports $\{N^{(i)},\ i = \overline{1,k}\}$ are independent from θ and $N \neq \varnothing$;

(C2) There exist the derivatives $\frac{\partial^m h^{(i)}(x;\theta)}{\partial \theta_j^m}$, $m = 1,2;\ i = 1,...,k;\ j = 1,...,s$, for all $\theta \in \Theta$;

(C3) $\int\limits_{-\infty}^{\infty} \left| \frac{\partial^m h^{(i)}(x;\theta)}{\partial \theta_j^m} \right| \mu(dx) < \infty,\ m = 1,2;\ i = 1,...,k;\ j = 1,...,s$ for all $\theta \in \Theta$;

(C4) There are finite integrals $I_{lj}^{(i)}(\theta) = M_\theta \left[\frac{\partial}{\partial \theta_l} \log h^{(i)}(X;\theta) \frac{\partial}{\partial \theta_j} \log h^{(i)}(X;\theta) \right]$ for all $l,j = 1,...,s$ and $\theta \in \Theta$;

(C5) The matrix $I_X(\theta) = \left\| I_{lj}^X(\theta) \right\|_{l,j=\overline{1,s}} = \left\| \sum\limits_{i=1}^{k} I_{lj}^{(i)}(\theta) \right\|_{l,j=\overline{1,s}} = \sum\limits_{i=1}^{k} I^{(i)}(\theta)$ positive defined for all $\theta \in \Theta$.

Obviously, that $I^{(i)}(\theta)$ is Fisher's information matrix according to pair $(X, \delta^{(i)})$, and $I_X(\theta)$ is same for r.v. X. Let

$$S_n(\mathbb{Z}^{(n)};\theta) = \frac{\partial \log p_n(\mathbb{Z}^{(n)};\theta)}{\partial \theta} = \sum_{j=1}^{n} l_\theta(X_j, Y_{1j}, Y_{2j}, w_j),$$

where

$$l_\theta(x, y_1, y_2, w) = w \sum_{i=1}^{k} \delta^{(i)} \frac{\partial \log h^{(i)}(x;\theta)}{\partial \theta} + (1-w) \frac{\partial \log(H(y_2;\theta) - H(y_1;\theta))}{\partial \theta}.$$

We note that $\mathbb{J}(\theta) = \mathbb{J}_1(\theta) + \mathbb{J}_2(\theta)$, where

$$\mathbb{J}_1(\theta) = \sum_{i=1}^{k} \int\limits_{-\infty}^{\infty} \int\limits_{y_1}^{\infty} \left[\int\limits_{-\infty}^{y_1} \frac{\partial \log h^{(i)}(x;\theta)}{\partial \theta} \left(\frac{\partial \log h^{(i)}(x;\theta)}{\partial \theta} \right)^T dH^{(i)}(x;\theta) + \right.$$

$$\left. + \int\limits_{y_2}^{\infty} \frac{\partial \log h^{(i)}(x;\theta)}{\partial \theta} \left(\frac{\partial \log h^{(i)}(x;\theta)}{\partial \theta} \right)^T dH^{(i)}(x;\theta) \right] dG(y_1, y_2),$$

$$\mathbb{J}_2(\theta) = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{y_1} \frac{\partial \log(H(y_2;\theta) - H(y_1;\theta))}{\partial \theta} \left(\frac{\partial \log(H(y_2;\theta) - H(y_1;\theta))}{\partial \theta} \right)^T \cdot$$

$$\cdot (H(y_2;\theta) - H(y_1;\theta)) dG(y_1, y_2).$$

Let $(u;v)$ is the scalar multiplication of vectors $u, v \in R^s$. The following theorem asserts LAN for LRS

Theorem 1. *Let us the regularity conditions (C1)-(C5) are hold and* $\det\{\mathbb{J}(\theta)\} \neq 0$. *Then for LRS* $L_{n,\theta}(u)$ *we have representation*

$$L_{n,\theta}(u) = \exp\left\{ n^{-1/2} \sum_{j=1}^{n} \left(l_\theta(X_j, Y_{1j}, Y_{2j}, w_j); u\right) - \frac{1}{2}\left(\mathbb{J}(\theta)u^T; u\right) + R_n(u;\theta) \right\}, \quad (9)$$

where as $n \to \infty$ *for all* $u \in R^s$,

$$R_n(u;\theta) \overset{\widetilde{Q}_\theta^{(n)}}{\to} 0 \quad (10)$$

and

$$\mathcal{L}\left(n^{-1/2} \sum_{j=1}^{n} l_\theta(X_j, Y_{1j}, Y_{2j}, w_j) / \widetilde{Q}_\theta^{(n)} \right) \to \mathbb{N}_s(0; \mathbb{J}(\theta)). \quad (11)$$

From (9) follows that LRS $L_{n,\theta}(u)$ is approximated with exponential density, and $\chi_{n,\theta}(u)$ has asymptotically s -dimensional normal distribution.

Using the properties of LAN for LRS we prove asymptotic normality of Bayesian type estimates. Let $\{\pi(u), u \in \Theta\}$ - non-negative measurable function $l(d;\theta) = (d-\theta)^2$ - loss function on the set $D \times \Theta$, where D is set of possible estimates for θ. We consider estimates $\widehat{\theta}_n \in D$, defined by the relation

$$\widehat{\theta}_n = \arg\min_{d \in D} \frac{\int\limits_{\Theta} l(d;\theta) p_n(\tilde{Z}^{(n)};\theta)\pi(\theta)d\theta}{\int\limits_{\Theta} p_n(\tilde{Z}^{(n)};\theta)\pi(\theta)d\theta}. \quad (12)$$

Note that if θ is r.v. with a priori density π, then $\widehat{\theta}_n$ is Bayesian estimate for θ. We prove the asymptotic normality of estimates $\widehat{\theta}_n$, limit distributions which are not depend on the functions. π.

Theorem 2. Let us the regularity conditions (C1)-(C5) are hold and function $\pi(\theta)$ is is continuous in the neighborhood of θ_0 and $\pi(\theta_0) \neq 0$. Then at $n \to \infty$, $\mathcal{L}\left(\sqrt{n}(\theta_n - \theta_0)/Q_{\theta_0}^{(n)}\right) \to N(0, \mathbb{J}^{-1}(\theta_0))$.

Remark. From the theorem 1, according to Fisher's definition (Ibragimov, Khas'minskii,1979, p.127), estimate $\widehat{\theta}_n$ is asymptotically efficient.

References

[1] Abdushukurov A.A. (2011). Estimators of unknown distributions from incomplete observations and its properties, *LAMBERT Academic Publishing*, 299p. (In Russian).

[2] Abdushukurov A.A., Nurmuhamedova N.S. (2011). Approximation of the likelihood ratio statistics in competing risks model under random censorship from both sides, *ACTA NUUz*, **N.4**, pp.162–172. (In Russian).

[3] Abdushukurov A.A., Nurmuhamedova N.S. (2012a). Asymptotics of the likelihood ratio statistics in competing risks model under multiple right censorship on the right, *In: Statistical Methods of estimation and Hypothesis Testing. Perm. Russia. Perm State University Press*, **Issue 21**, pp.4–15. (In Russian).

[4] Abdushukurov A.A., Nurmuhamedova N.S. (2012b). Locally asymptotically normality in competing risks model, *Uzbek Mathematical Journal*, **N.2**, pp.5–12. (In Russian).

[5] Abdushukurov A.A., Nurmuhamedova N.S. (2012c). Locally asymptotically normality of statical experiments , *LAMBERT Academic Publishing*, 136p. (In Russian).

[6] Abdushukurov A.A., Nurmuhamedova N.S. (2013). Local approximate normality of likelihood ratio statistics in competing ricks model under random censorship fromboth sides, *Far East Journal of Theoretical Statistics,* **v.42**, **N.2**, pp.107–122.

[7] Abdushukurov A.A., Nurmuhamedova N.S. (2014). Locally asymptotically normality of the family of distributions by incomplete observations, *Journal of Siberian Federal University. Mathematics & Physics.* **v.7**, **N.2**, pp.141–154.

[8] Burke M.D., Csörgő S.,Horváth L. (1981). Strong approximations of some biometric estimates under random censorship, *Z. Wahrscheinlich. verw. Gebiete*, **v.56**, pp.87–112.

[9] Hajek J. (1972). Local asymptotic minimax and admissibility in estimation, *Proc. Sixth. Berkeley Symp. on Math. Statist. and Prob*, **v.1**, pp.175–194.

[10] Ibragimov I.A, Khas'minskii R.Z. (1979). Asymptotic theory of estimation, *M.: Nauka*, (In Russian).

[11] Le Cam L. (1960). Locally asymptotically normal families of distributions, *Unif. Calif. Publ. Statist.* **v.3**, pp.37-98.

[12] Leman E. (1964). Testing of statistical hypothesis, *M.: Nauka*, (In Russian).

[13] Langberg N., Proshan F., Quinzi A.J. (1978). Converting dependent models into independents ones, preserving essential features, *Ann. Probab*, **v.6**, pp.174–181.

[14] Rusas J. (1975). Contiguity of probability measures, *M.: Mir,* 254p. (In Russian).

Nonparametric Interval Estimation of the Multivariate Probability Density Function and its Derivatives

Rakhimova G.G.[1], Tursunov G.T.[2]
[1] *Tashkent Avto - Road Institute, Tashkent, Uzbekistan*
[2] *National Universitete of Uzbekistan, Tashkent, Uzbekistan*
e-mail: tursunovgafur52@gmail.com

Abstract

Conditions of asymptotic normality of the estimate based on a random number of observations for derivative of probability density function of the random vector are proved. Obtained asymptotic consistency fixed-width confidence intervals for a derivative of probability density function.

Keywords: random vector, probability density function, derivative, fixed-width, confidence interval, asymptotic normality, asymptotic consistency.

1. Let $\xi_1, \xi_2, ..., \xi_n$ be a independent observations of a m - dimensional random vector ξ having the unknown probability density function $f(x), x = (x_1, x_2, ...x_m) \in R_m$, where R_m m - dimensional Euclidean space with respect to Lebesgue measure , $m \geq 1$ and $\xi_k = (\xi_{k1}, \xi_{k2}, ...\xi_{km}), 1 \leq k \leq n$.

Denote by $f^{(r)}(x) = \frac{\partial^r (f(x))}{\partial^{r_1} x_1 \partial^{r_2} x_2 ... \partial^{r_m} x_m}, r_1 + r_2 + ... + r_m = r, r_i \geq 0, 1 \leq i \leq m$ derivative of r- order of probability density function $f(x)$, $r \geq 0, \left(f^{(0)}(x) \equiv f(x)\right)$. In this paper we consider the estimate of $f^{(r)}(x)$ at point $x \in R_m$ given by

$$f_n^{(r)}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i^{m(r+1)}} K^{(r)} \left(\frac{x - \xi_i}{h_i} \right), \tag{1}$$

where $h_n = n^{-\delta}, \delta > 0$, $K^{(r)} \left(\frac{x-\xi_i}{h_i} \right) = K^{(r)} \left(\frac{x_1 - \xi_{i1}}{h_i}, \frac{x_2 - \xi_{i2}}{h_i}, ..., \frac{x_m - \xi_{im}}{h_i} \right), K^{(r)}(x)$ is derivative of r- order of positive function $K(x), x = (x_1, x_2, ...x_m) \in R_m$ $\left(K^{(0)}(x) \equiv K(x)\right)$. The asymptotic properties estimate (1) for the case $m = 1$ and $r = 0$ has been investigated by Yamato [1] and Davies [2].

In many practical situations the number of observation N_t which we observe in time $(0, t]$ is a random variable. For example, we consider the problem of estimating the probability density function of the waiting times of customers at a service station and we may assume that the number of customers N_t, arriving in time $(0, t]$ is a Poisson random variable with parameter λt, where λ is positive number.

We assume that for any $t > 0$ N_t is an integer valued random variable, which perhaps arbitrary dependent of the random vectors $\xi_1, \xi_2, ..., \xi_n$. In this paper we proved conditions of asymptotic normality of the estimate$f_{N_t}^{(r)}(x)$ for derivative of r-order of probability density function$f^{(r)}(x)$, based on sample $\xi_1, \xi_2, ..., \xi_{N_t}$ and given by:

$$f_{N_t}^{(r)}(x) = \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{1}{h_i^{m(r+1)}}K^{(r)}\left(\frac{x-\xi_i}{h_i}\right)$$

and obtained results are used to nonparametric interval estimation of the multivariate probability density function and its derivatives.

Introduce following conditions:

(N): $\frac{N_t}{t} \xrightarrow{p} N$ as $t \to \infty$, where $P(N{>}0) = 1$.

(K) : $\int\limits_{R_m} K(x)dx = 1$, $\int\limits_{R_m} x_i K(x)dx = 0, 1 \le i \le m$, $\lim\limits_{\|x\|\to\infty} \|x\|^m \left|K^{(r)}(x)\right| < \infty$, $\int\limits_{R_m} \left|K^{(r)}(x)\right| dx < \infty, r \ge 0$.

(δ) : $\frac{1}{(3+2r)m} < \delta < \frac{1}{(1+2r)m}$.

(F): derivative $f^{(r+1)}(x)$ exist and is bounded.

Theorem 1. If conditions $(N), (K), (\delta)$ and (F) are satisfied, then as $t \to \infty$ at every point x of continuity $f(x)$ random variable

$$\sqrt{N_t^{1-(1+2r)\delta m}}\left(f_{N_t}^{(r)}(x) - f^{(r)}(x)\right)$$

converges in distribution to a normal random variable with mean zero and variance $\sigma^2(x) = \alpha(r)k(r)f(x)$, where $\alpha(r) = \frac{1}{1+(1+2r)\delta m}, k(r) = \int\limits_{R_m}\left|K^{(r)}(x)\right|^2 dx$.

Remark. Asymptotic normality of the estimate $f_{N_t}^{(r)}(x)$ for $r = 0$ and $r = 1$ proved in paper Samanta and Mugisha [3] by more strongly condition $P\left(\lim\limits_{t\to\infty}\frac{N_t}{t} = \lambda\right) = 1$, where λ is positive number.

2. Use results of theorem 1 for obtain fixed-width confidence intervals for a derivative of r- order of probability density function $f(x)$.

Let $0 < \gamma < 1$, $a = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$, $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x} e^{-\frac{y^2}{2}}dy$. If $\varepsilon > 0$, then for $n \ge \left(\frac{a\sigma(x)}{\varepsilon}\right)^{\frac{2}{1-(1+2r)\delta m}}$ it follows that

$$P\left\{f^{(r)}(x) \in \left[f_n^{(r)}(x) - \varepsilon, f_n^{(r)}(x) + \varepsilon\right]\right\} \ge P\left\{\sigma^{-1}n^{\frac{1-(1+2r)\delta m}{2}}\left|f_n^{(r)}(x) - f^{(r)}(x)\right| \le a\right\}.$$

Consequently, if $n(\varepsilon, f) = \min(n \ge 1 : n \ge n_0(\varepsilon, f))$, where $n_0(\varepsilon, f) = \left(\frac{a^2\alpha(r)k(r)f(x)}{\varepsilon^2}\right)^{\frac{1}{1-(1+2r)\delta m}}$, then $\lim\limits_{\varepsilon\to 0} n(\varepsilon, f) = \infty$ and if assume that conditions theorem 1 are satisfied for $P\left(N_t = n(\varepsilon, f)\right) = 1$ it follows from theorem 1 that

$$\lim_{\varepsilon\to 0} P\left\{f^{(r)}(x) \in \left[f_{n(\varepsilon,f)}^{(r)}(x) - \varepsilon, f_{n(\varepsilon,f)}^{(r)}(x) + \varepsilon\right]\right\} \ge 2\Phi(a) - 1 = \gamma.$$

The stopping time $n(\varepsilon, f)$ depend on unknown probability density function $f(x)$. Therefore instead $n(\varepsilon, f)$ consider random stopping time $N_\varepsilon = n(\varepsilon, f_n^{(0)})$ or

$$N_\varepsilon = \min\left(n \ge 1 : n \ge \left(\frac{a^2\alpha(r)k(r)}{\varepsilon^2}f_n^{(0)}(x)\right)^{\frac{1}{1-(1+2r)\delta m}}\right).$$

Theorem 2. If conditions $(K), (\delta)$ and (F) are satisfied, then $\frac{N_\varepsilon}{n_0(\varepsilon,f)} \xrightarrow{p} 1$ as $\varepsilon \to 0$.

Theorem 3. If conditions $(K), (\delta)$ and (F) are satisfied, then $\lim\limits_{\varepsilon\to 0} P\left\{f^{(r)}(x) \in \left[f^{(r)}_{N_\varepsilon}(x) - \varepsilon, f^{(r)}_{N_\varepsilon}(x) + \varepsilon\right]\right\} \geq \gamma$. Fixed-widht confidence interval $\left[f^{(r)}_{N_\varepsilon}(x) - \varepsilon, f^{(r)}_{N_\varepsilon}(x) + \varepsilon\right]$ is asymptotic cocsistency.

References

[1] Yamato H. (1971). Sequential estimation of a continuous probability density function and mode. *Bull. Math. Statist.* **14**, pp. 1-12.

[2] Davies H. I. (1973). Strong consistency of a sequential estimation of a probability density function. *Bull. Math. Statist.* **15**, pp. 49-54.

[3] Samanta M., Mugisha R. X. (1981). On a class of estimates of the probability density function and mode based on a random number of observations. *Calcutta statistical association bulletin*, **30**, N 117-118, pp. 23-40.

New Robust Statistical Method for Two-Sample Problem Testing under Right-Censored Data

Petr Philonenko and Sergey Postovalov
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: petr-filonenko@mail.ru, postovalov@ngs.ru

Abstract

In hypothesis testing, there are always such two alternative hypotheses A and B where one two-sample test is more preferable in the alternative A and less preferable in the alternative B than another two-sample test. Hence, the most powerful two-sample test does not exist in the general case. Therefore, we have constructed the types of alternative hypotheses and determined three two-sample tests complementing each other in terms of the test power. They are Bagdonavičus-Nikulin and weighted Kaplan-Meier two-sample tests. Using these test statistics, we have proposed a new two-sample test $MIN3$ that has the test power close to the maximum of these tests power. The power research is studied.

Keywords: survival analysis, hypothesis testing, two-sample problem, test power, right-censored data.

Introduction

Two-sample problem testing is a very popular statistical problem, especially for lifetime data, for example, in medicine and biostatistics research. There are a lot of tests for solving. But any test is more preferable in a small set of alternative hypotheses. We would like to expand a set of alternative hypotheses where one test is more preferable. The preference is the test power.

There are various ways for test statistic construction, for example, distance tests between survival curves [1–6], minimum p-value tests [7-9], test statistics based on a select function among other tests [10], rank-sum tests [11–14], special models of the behavior of survival functions [15–17]. In the case of Bagdonavičus-Nikulin tests, there are tests based on models for one intersection point of survival functions (the single crossing test based on SCE-model [16]), for two intersection points of survival functions (the multiple crossing test based on MCE-model [16]) and for constant ratio of hazard functions.

Our simulations [18] have shown the Bagdonavičus-Nikulin tests can be stronger than other two-sample tests when survival functions have a point of the intersection. In the case without intersection, a distance two-sample test statistic is more preferable. Moreover, using the Wald maximin model (the test for decision-making under risk and uncertainty) on groups of alternative hypotheses in our simulations [18], we have concluded there are three two-sample tests which are complementary in terms of the test power. These tests are the weighted Kaplan-Meier (the distance type test) and Bagdonavičus-Nikulin tests.

Therefore, we propose a new two-sample test that the test statistic is computed through the test statistics and combines their advantages. In our simulation, the application of the new test is a robust way for two-sample problem solving.

1 Two-Sample Tests

Suppose that we have two samples of continues variables ξ_1 and ξ_2 respectively, $X_1 = \{t_{11}, t_{12}, ..., t_{1n_1}\}$ and $X_2 = \{t_{21}, t_{22}, ..., t_{2n_2}\}$ of two survival distributions $S_1(t)$ and $S_2(t)$. The observation $t_{ij} = \min(T_{ij}, C_{ij})$, where T_{ij} and C_{ij} are the failure and censoring times for the j-th object of the i-th group. T_{ij} and C_{ij} are i.i.d. with CDF $F_i(t)$ and $F_i^C(t)$ respectively. Survival curve means the probability of survival in the time interval $(0, t)$

$$S_i(t) = P\{\xi_i > t\} = 1 - F_i(t).$$

Then the null hypothesis is

$$H_0 : S_1(t) = S_2(t).$$

Further, we will suppose that the elements of samples are ordered: $t_{11} < ... < t_{1n_1}$ and $t_{21} < ... < t_{2n_2}$. Also, we pool these samples and sort their elements by ascending order $T = X_1 \cup X_2 = \{t_1, t_2, ..., t_n\}$, where $t_1 < ... < t_n$ and $n = n_1 + n_2$.

Denote the sample indicator v_i and the censoring indicators c_{ij} and c_i as

$$v_i = \begin{cases} 0, & t_i \in X_1, \\ 1, & t_i \in X_2 \end{cases} \quad c_{ij} = \begin{cases} 0, & t_{ij} \text{ is failure}, \\ 1, & t_{ij} \text{is censored} \end{cases} \quad c_i = \begin{cases} 0, & t_i \text{ is failure}, \\ 1, & t_i \text{is censored} \end{cases}$$

Let us consider the two-sample tests for lifetime data.

1.1 The Weighted Kaplan-Meier Test

The weighted Kaplan-Meier [6] is a distance two-sample test. The main idea of the test is to measure the distance between estimates of survival functions. In the case of the weighted Kaplan-Meier, the estimates are Kaplan-Meier estimates. The test statistic can be computed by following way:

$$S_{WKM} = \frac{U}{\sqrt{\sigma}},$$

where

$$U = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \int_0^{t_n} w(t) \left(\hat{S}_1(t) - \hat{S}_2(t) \right) dt, \; w(t) = \frac{(n_1 + n_2)\, \hat{F}_1^C(t-) \hat{F}_2^C(t-)}{n_1 \hat{F}_1^C(t-) + n_2 \hat{F}_2^C(t-)},$$

$$\sigma = - \int_0^{t_n} \left\{ \int_t^{t_n} w(\tau) \hat{S}(\tau) d\tau \right\}^2 \frac{n_1 \hat{F}_1^C(t-) + n_2 \hat{F}_2^C(t-)}{(n_1 + n_2)\, \hat{F}_1^C(t-) \hat{F}_2^C(t-)} \frac{d\hat{S}(t)}{\hat{S}(t) \hat{S}(t-)},$$

and $\hat{F}_1^C(t)$, $\hat{F}_2^C(t)$, $\hat{S}_1(t)$ and $\hat{S}_2(t)$ are Kaplan-Meier estimates of corresponding functions but $\hat{S}(t)$ is a Kaplan-Meier estimate of the pooled sample.

The null hypothesis is rejected with the significance level α if $|S_{WKM}| > z_{1-\frac{\alpha}{2}}$, where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$ quantile of the standard normal distribution.

1.2 The Bagdonavičus-Nikulin Multiple Crossing Test

The test statistic S_{BN2} [16] is

$$S_{BN2} = (U_1, U_2, U_3)\,\Sigma^{-1}(U_1, U_2, U_3)^T,$$

where

$$U_1 = \sum_{j:c_{1j}=0} \frac{Y_2(t_{1j})}{Y(t_{1j})} - \sum_{j:c_{2j}=0} \frac{Y_1(t_{2j})}{Y(t_{2j})},\ U_2 = -\sum_{j:c_{1j}=0} \frac{Y_2(t_{1j})}{Y(t_{1j})}\Lambda(t_{1j}) + \sum_{j:c_{2j}=0} \frac{Y_1(t_{2j})}{Y(t_{2j})}\Lambda(t_{2j}),$$

$$U_3 = -\sum_{j:c_{1j}=0} \frac{Y_2(t_{1j})}{Y(t_{1j})}\Lambda^2(t_{1j}) + \sum_{j:c_{2j}=0} \frac{Y_1(t_{2j})}{Y(t_{2j})}\Lambda^2(t_{2j}).$$

The elements of the matrix Σ can be computed as

$$\sigma_{11} = \sum_{i=1}^{2}\sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})},\ \sigma_{12} = \sigma_{21} = \sum_{i=1}^{2}\sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})}\Lambda(t_{ij}),$$

$$\sigma_{13} = \sigma_{31} = \sigma_{22} = \sum_{i=1}^{2}\sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})}\Lambda^2(t_{ij}),$$

$$\sigma_{23} = \sigma_{32} = \sum_{i=1}^{2}\sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})}\Lambda^3(t_{ij}),\ \sigma_{33} = \sum_{i=1}^{2}\sum_{j:c_{ij}=0} \frac{Y_1(t_{ij})Y_2(t_{ij})}{Y^2(t_{ij})}\Lambda^4(t_{ij}),$$

$$Y(t) = Y_1(t) + Y_2(t),\ Y_i(t) = \sum_{j=1}^{n_i} Y_{ij}(t),\ Y_{ij}(t) = 1_{\{t_{ij}\geq t\}},\ i = 1, 2$$

$$\Lambda(t) = \sum_{i=1}^{2}\sum_{j:c_{ij}-0,t_{ij}\leq t} \frac{1}{Y\left(t_{ij}\right)}.$$

The null hypothesis is rejected with the significance level α if $S_{BN2} > \chi^2_{1-\alpha}(3)$, where $\chi^2_{1-\alpha}(3)$ is the $(1-\alpha)$ quantile of the chi-square distribution with 3 degrees of freedom.

1.3 The Bagdonavičus-Nikulin Constant Hazard Ratio Test

The test statistic S_{BN3} [17] is

$$S_{BN3} = (U_1, U_2)^T \Sigma^{-1} (U_1, U_2),$$

where

$$U_1 = \sum_{j=1}^{n_1} (1 - c_{1j}) \frac{K_1(t_{1j})}{Y(t_{1j})} Y_2(t_{1j}) - \sum_{j=1}^{n_2} (1 - c_{2j}) \frac{K_1(t_{2j})}{Y(t_{2j})} Y_1(t_{2j}),$$

$$U_2 = \sum_{j=1}^{n_1} (1 - c_{1j}) \frac{K_2(t_{1j})}{Y(t_{1j})} Y_2(t_{1j}) - \sum_{j=1}^{n_2} (1 - c_{2j}) \frac{K_2(t_{2j})}{Y(t_{2j})} Y_1(t_{2j}),$$

The elements of the matrix Σ can be computed as

$$\sigma_{ij} = \sum_{r=1}^{2} \sum_{s=1}^{n_r} (1 - c_{rs}) K_i(t_{rs}) \frac{Y_1(t_{rs}) Y_2(t_{rs})}{Y^2(t_{rs})}.$$

The following functions are $K_1(t)$ and $K_2(t)$ may be used

$$K_1(t) = \exp(-\Lambda(t))/\sqrt{n},\ K_2(t) = \frac{\exp(-\Lambda(t)) - 1}{\sqrt{n}},$$

$$Y(t) = Y_1(t) + Y_2(t),\ Y_i(t) = \sum_{j=1}^{n_i} Y_{ij}(t),\ Y_{ij}(t) = 1_{\{t_{ij} \geq t\}},\ i = 1, 2$$

$$\Lambda(t) = \sum_{i=1}^{2} \sum_{j: c_{ij} - 0, t_{ij} \leq t} \frac{1}{Y(t_{ij})}.$$

The null hypothesis is rejected with the significance level α if $S_{BN3} > \chi^2_{1-\alpha}(2)$, where $\chi^2_{1-\alpha}(2)$ is the $(1 - \alpha)$ quantile of the chi-square distribution with 2 degrees of freedom.

1.4 New Two-Sample Test

The test statistic of the proposed $MIN3$ test is

$$S_{MIN3} = \min\{p_{WKM}, p_{BN2}, p_{BN3}\},$$

where

$$p_{WKM} = 2 \cdot \min\left\{F_{N(0,1)}(S_{WKM}), 1 - F_{N(0,1)}(S_{WKM})\right\},$$

$$p_{BN2} = 1 - F_{\chi^2(3)}(S_{BN2}),\ p_{BN3} = 1 - F_{\chi^2(2)}(S_{BN3}),$$

$F_{N(0,1)}(t)$ is a cumulative distribution function (CDF) of the standard normal distribution at time t, $F_{\chi^2(3)}(t)$ is a CDF of the chi-square distribution with 3 degrees of

freedom at time t and $F_{\chi^2(2)}(t)$ is a CDF of the chi-square distribution with 2 degrees of freedom at time t.

The null hypothesis is rejected with the significance level α if $S_{MIN3} > G(S_{MIN3}|H_0)_{1-\alpha}$, where $G(S_{MIN3}|H_0)_{1-\alpha}$ is the $(1-\alpha)$ quantile of the distribution of the test statistic S_{MIN_3} under the null hypothesis H_0. In other words, the test statistic S_{MIN3} has a right-critical area. The simulated (sample size $n_1 = n_2$ is 200 observations, 2 700 000 replications of the Monte-Carlo method) upper percent points of the distribution of the new test statistic S_{MIN3} under H_0 are represented in Table 1.

Table 1: Simulated upper percent points of the distribution of the new test statistic S_{MIN3} under H_0

α	$\hat{G}(S_{MIN_3}\|H_0)_{1-\alpha}$	α	$\hat{G}(S_{MIN_3}\|H_0)_{1-\alpha}$
0.0010	0.9765	0.1500	0.6419
0.0025	0.9618	0.2000	0.5812
0.0050	0.9440	0.3000	0.4729
0.0075	0.9296	0.4000	0.3832
0.0100	0.9186	0.5000	0.3020
0.0250	0.8666	0.6000	0.2280
0.0500	0.8037	0.7000	0.1605
0.0750	0.7553	0.8000	0.0999
0.1000	0.7136	0.9000	0.0449
0.1250	0.6761	0.9990	0.0003

2 Test power

In addition to the considered tests, we consider some two-sample tests for a test power research. They are the generalized Wilcoxon tests (Peto S_P and Gehan S_G tests), log-rank test S_{LG}, Cox-Mantel test S_{CM}, Q-test S_Q, Bagdonavičus-Nikulin single crossing test S_{BN1}, weighted log-rank tests (Tarone-Ware $S_{WLG(TW)}$, Peto-Prentice $S_{WLG(PP)}$ and Prentice $S_{WLG(P)}$) and maximum value test S_{MAX}.

We have constructed some types of alternative hypotheses with 0, 1 and 2 points of intersection represented in Figure 1. For every type, some alternative hypotheses correspond with various families of failure time distribution $F(t)$ (Weibull, Gamma, Exponential, Log-Normal distributions). These alternative hypotheses are represented in [18]. The test power research is studied for various families of censored time distribution $F^C(t)$ (Weibull, Gamma and Exponential), the sample size is $n_1 = n_2 = 200$ observations, the test size $\alpha = 0.05$, the size of the Monte-Carlo replications is 150 000 and a censored rate 0%-50%.

The result of the simulation is represented in Table 2. In the top line, there are maximal values of a corresponding column. Every value of a test power is a

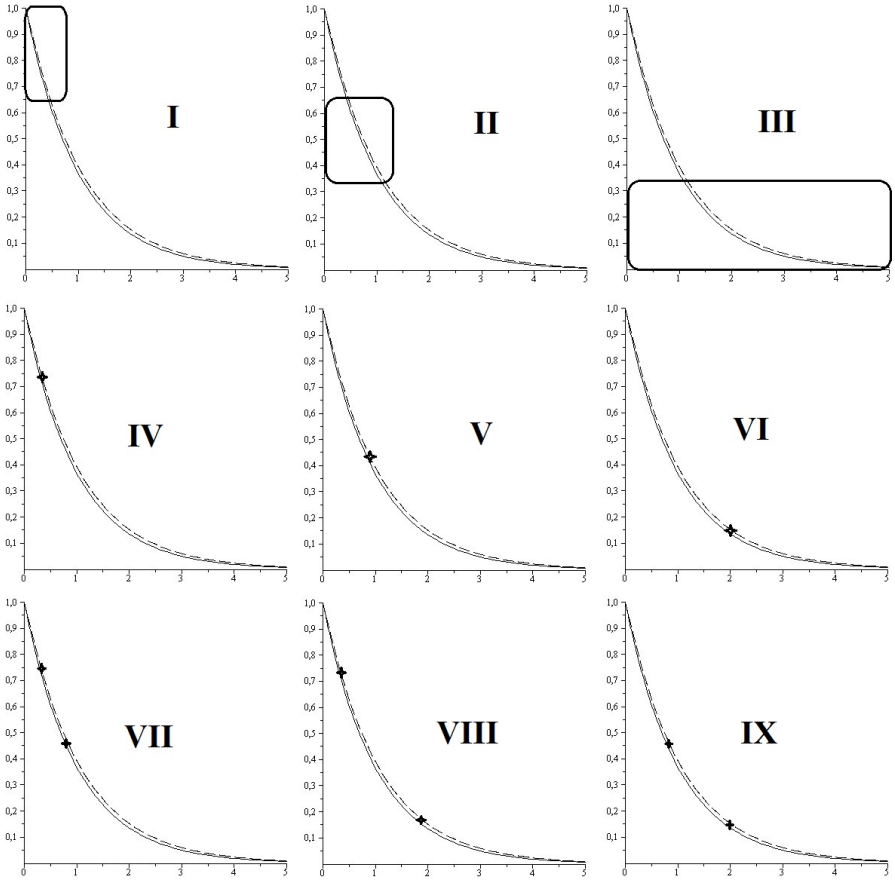


Figure 1: The types of considered alternative hypotheses

minimum among all alternative hypotheses, censored rates and distributions $F^C(t)$ for a corresponding type.

One can see two-sample tests have high test power on alternative hypotheses without points of intersections in early and middle time, with one point of intersection in late time and with two intersections in middle and late time.

Early time is a set of values t between $S(t) = 1.00$ and $S(t) = 0.67$. Middle time is a set of values t between $S(t) = 0.67$ and $S(t) = 0.33$. Late time is a set of values t between $S(t) = 0.33$ and $S(t) = 0.00$.

Table 2 shows that the proposed two-sample $MIN3$ test has a test power close to the maximal value on a certain type of alternative hypothesis. The difference does not exceed 0.03. It makes possible to conclude that the $MIN3$ test statistic is a stable test statistic when the type of an alternative hypothesis is unknown.

Conclusions

Two-sample problem testing is one of fundamental hypotheses in statistics. Such testing is applied in industry, medicine, biostatistics, sociology and so on. The proposed

Table 2: Simulated test power under considered types of alternative hypotheses

Type	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**
MAX	**0.597**	**0.207**	**0.093**	**0.079**	**0.078**	**0.243**	**0.070**	**0.097**	**0.275**
S_P	0.427	0.180	0.069	0.049	0.052	0.189	0.051	0.050	0.083
S_G	0.531	0.207	0.066	0.052	0.056	0.243	0.051	0.054	0.131
S_{LG}	0.209	0.132	0.077	0.051	0.050	0.115	0.050	0.049	0.051
S_{CM}	0.212	0.133	0.077	0.052	0.049	0.114	0.051	0.050	0.052
S_Q	0.303	0.156	0.073	0.051	0.051	0.169	0.052	0.052	0.070
S_{BN1}	0.434	0.148	0.069	0.072	0.076	0.196	0.064	0.053	0.147
S_{BN2}	0.490	0.134	0.093	0.079	0.073	0.173	0.070	0.097	0.275
S_{BN3}	0.489	0.156	0.066	0.078	0.078	0.209	0.065	0.063	0.180
S_{WKM}	0.535	0.206	0.066	0.051	0.057	0.242	0.051	0.053	0.131
$S_{WLG(TW)}$	0.411	0.182	0.071	0.050	0.052	0.191	0.051	0.050	0.078
$S_{WLG(PP)}$	0.047	0.059	0.065	0.077	0.058	0.050	0.055	0.048	0.069
$S_{WLG(P)}$	0.048	0.059	0.065	0.078	0.056	0.049	0.054	0.050	0.070
S_{MAX}	0.478	0.180	0.075	0.057	0.056	0.205	0.060	0.059	0.117
S_{MIN3}	0.597	0.180	0.083	0.074	0.071	0.227	0.063	0.077	0.256

$MIN3$ test statistic has a test power close to the maximal value on a certain type of an alternative hypothesis. The maximal difference does not exceed 0.03. This test statistic is a stable two-sample test statistic and can be applied when the type of an alternative hypothesis is unknown.

This research has been supported by Russian Ministry of Education and Science as a part of the state task (project 1.1009.2017/4.6) and by Novosibirsk State Technical University as a young scientist project (project C-15, 2017).

References

[1] Smirnov N.V., "Table for estimating the goodness of fit of empirical distributions", Ann. Math. Stat., vol. 19, pp. 279–281, 1948.

[2] Lehmann E.L., "Consistency and unbiasedness of certain nonparametric tests", Ann. Math. Statist., vol. 22, no. 1, pp. 165–179, 1951.

[3] Rosenblatt M., "Limit theorems associated with variants of the von Mises statistic", Ann. Math. Statist., vol. 23, no. 1, pp. 617–623, 1952.

[4] Scholz F.W., Stephens M.A., "K-sample Anderson-Darling Tests", Journal of the American Statistical Association, vol. 82, no. 399, pp. 918–924, 1987.

[5] Pettitt A.N., "A two-sample Anderson-Darling rank statistic", Biometrika, vol. 63, no. 1, pp. 161–168, 1976.

[6] Pepe M.S. and Fleming T.R., Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics. 1989; 45 : 497–507.

[7] Philonenko P, Postovalov S., A new two-sample test for choosing between log-rank and Wilcoxon tests with right-censored data. J Stat Comput Simul. 2015; 85(14) : 2761–2770. doi:10.1080/00949655.2014.941533.

[8] Petr Philonenko, Sergey Postovalov and Artem Kovalevskii (2016), The limit test statistic distribution of the maximum value test for right-censored data, Journal of Statistical Computation and Simulation, 86:17, 3482-3494, DOI: 10.1080/00949655.2016.1164703

[9] Philonenko P., The limit test statistic distribution of the maximum value test for right-censored data / P. Philonenko, S. Postovalov, A. Kovalevskii // Journal of Statistical Computation and Simulation. - 2016. - Vol. 86, iss. 17. - P. 3482-3494

[10] Martinez Ruvie L.M.C., Naranjo Joahua D., "A pretest for choosing between Logrank and Wilcoxon tests in the two-sample problem", International Journal of Statistics, vol. LXVIII, no. 2, pp. 111–125, 2010.

[11] Gehan E.A., "A generalized Wilcoxon test for comparing arbitrarily singlycensored samples", Biometrika, vol. 52, no. 1/2, pp. 203–223, 1965.

[12] Peto R., Peto J., "Asymptotically efficient rank invariant test procedures", Journal of the Royal Statistical Society, Series A (General), vol. 135, no. 2, pp. 185–207, 1972.

[13] Savage, I. R. (1956). Contributions to the Theory of Rank Order Statistics: The Two Sample Case. Annals of Mathematical Statistics, 27, 590–615.

[14] Mantel N., "Evaluation of survival data and two new rank order statistics arising in its consideration", Cancer Chemotherapy Rep., vol. 50, pp. 163–170, 1967.

[15] Bagdonavičius V.B., Levuliene R.J., Nikulin M.S., et al., "Tests for equality of survival distributions against non-location alternatives", Lifetime Data Analysis, vol. 10, no. 4, pp. 445–460, 2004.

[16] Bagdonavičius V.B., Nikulin M., "On goodness-of-fit tests for homogeneity and proportional hazards", Applied Stochastic Models in Business and Industry, vol. 22, no. 1, pp. 607–619, 2006.

[17] Bagdonavičius, V., Kruopis, J. and Nikulin, M. S. (2013) Censored and Truncated Data, in Non-parametric Tests for Censored Data, John Wiley & Sons, Inc, Hoboken, NJ, USA.

[18] P. Philonenko, S. N. Postovalov (2016) Test Power in Two-Sample Problem Testing as the Utility Function in the Theory of Decision Making Under Risk and Uncertainty // XIII International Scientifical-Technical Conefernce "Actual problems of electronic instrument engineering (APEIE-2016)", October 03-06, 2016 // V.1, P.2, pp.369-373, ISBN 978-5-7782-2991-4 - supported by Russian Ministry of Education and Science (project 2.541.2014K).

A Comparative Analysis of the Wiener, Gamma and Inverse Gaussian Degradation Models

Ekaterina V. Chimitova, Evgeniya S. Chetvertakova,
Sofiya A. Sergeeva and Evgeniya A. Osinceva
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: chimitova@corp.nstu.ru, chetvertakova@corp.nstu.ru,
sergeeva.2013@stud.nstu.ru, j.osinceva@gmail.ru

Abstract

In this paper, the construction of degradation models are considered. We make a description of gamma and Wiener degradation models as frequently used and inverse Gaussian model as an alternative model. To avoid the misspecification of degradation model, the identification approach, which is based on the application of goodness-of-fit tests, is proposed. Also, we give an example of the application of the proposed approach for the GaAs lasers data.

Keywords: degradation process, Gamma degradation model, Wiener degradation model, Inverse Gaussian degradation model, testing goodness-of-fit, GaAs lasers data, reliability.

Introduction

The reliability is one of the most important characteristics of high-quality devices. Therefore, the quality assurance and analysis of the reliability of technical items have been recently paid more attention in manufacturing. Generally, the reliability is defined using the information about failures of tested items. However, usually only a few number of high-reliable devices is assigned to the experiment and in most cases the failure information is not enough for carrying out the adequate analysis. Thus, it is necessary to use an additional lifetime data – so-called degradation data. A time moment, when the degradation index gets a critical level, is taken as a failure time. For the further statistical analysis, the degradation model should be constructed.

The most popular statistical degradation models are gamma and Wiener degradation models [2]- [4] which are well described in literature and widely applicable because of the possibility to obtain the reliability estimates analytically. However, there are some examples, when the gamma and Wiener degradation models are not appropriate [5], [6]. In this case, it is necessary to consider and compare different degradation models and define the most appropriate one for each set of data.

In this paper, we consider the example of GaAs lasers data analysis, which is often described in publications devoted to the investigation of degradation models. In [4], this data have been analyzed using gamma degradation model and Wiener degradation model. Inverse Gaussian degradation model is described in [9] as another variant of the model for lasers data. In this research, we have considered Gamma, Wiener and Inverse Gaussian degradation models, then we have proposed the approach to identification of the degradation model using goodness-of-fit tests. Finally, we have

analyzed the GaAs lasers data using considered degradation models and defined the most appropriate model for these data.

1 Gamma Degradation model

Stochastic process $Z(t)$, $t \geq 0$ characterizing degradation process is referred to as the gamma degradation process, if

- $Z(0) = 0$;
- $Z(t)$ is a stochastic process with independent increments;
- increments $\Delta Z(t) = Z(t + \Delta t) - Z(t)$ have the gamma distribution with probability density function

$$f(t; \gamma, \sigma\Delta m(t)) = \left(\frac{t}{\gamma}\right)^{\sigma\Delta m(t)-1} \frac{e^{-t/\gamma}}{\gamma\Gamma(\sigma\Delta m(t))},$$

  where $\sigma\Delta m(t)$ is a shape parameter and γ is a scale parameter, $m(t)$ is a positive increasing trend function.

If random variables ξ_1 and ξ_2 follow the gamma distribution with the scale parameter γ and shape parameters σ_1 and σ_2, correspondingly, then $\xi_1 + \xi_2$ follows the gamma distribution with the scale parameter γ and the shape parameter $\sigma_1 + \sigma_2$. This property explains the fact of using the gamma distribution as a distribution of increments.

Let the mathematical expectation of degradation process $Z(t)$ is

$$M(Z(t)) = m(\gamma; t),$$

where $m(t)$ is a trend function of the degradation index.

In this paper, we have considered a linear trend functions: $m(t) = \gamma t$.

The time-to-failure is a random variable

$$T = sup\{t : Z(t) < z_0\},$$

where z_0 is a critical value of the degradation path. Then, the reliability function for the gamma degradation model is given by:

$$S(t) = P\{T > t\} = P\{Z(t) < z_0\} = F(z_0; \gamma, \sigma m(t)).$$

Let the realization of degradation path $Z^i(t)$ for the i-th item is denoted as

$$Z^i = \{(0, Z_0^i = 0), (t_1^i, Z_1^i), \ldots, (t_{k_i}^i, Z_{k_i}^i)\}, i = \overline{1, n},$$

where k_i is a number of time moments measured degradation. Then, a sample of independent degradation increments can be written as:

$$\mathbf{X_n} = \{X_j^i = Z_j^i - Z_{j-1}^i, i = \overline{1, n}, j = \overline{1, k_i}\}. \quad (1)$$

Maximum likelihood estimates of parameters σ and γ are calculated by maximization of the log-likelihood function:

$$\ln L(\mathbf{X}_n) = \sum_{i=1}^{n} \sum_{j=1}^{k} \ln f(X_j^i; \gamma, \sigma\Delta m(t)) \to \max_{\gamma,\sigma}.$$

2 Wiener Degradation model

The Wiener degradation process $Z(t)$, $t \geq 0$ is defined as the stochastic process with the following suppositions:

- $Z(0) = 0$;
- $Z(t)$ is a stochastic process with independent increments;
- increments $\Delta Z(t) = Z(t + \Delta t) - Z(t)$ have the normal distribution with probability density function

$$f(t; \theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi}\theta_2}\right) exp\left(-\frac{(x-\theta_1)^2}{2{\theta_2}^2}\right),$$

  where $\theta_1 = \gamma \Delta m(t)$ is a shift parameter , $\theta_2 = \sigma\sqrt{(\Delta m(t))}$ is a scale parameter, $m(t)$ is a positive increasing trend function:

$$M(Z(t)) = m(t).$$

Obviously, if random variables ξ_1 and ξ_2 follow normal distribution with shift parameters μ_1 and μ_2 and scale parameters σ_1 and σ_2, correspondingly, then $\xi_1 + \xi_2$ follows the normal distribution with the shift parameter $\mu_1 + \mu_2$ and the scale parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

Then, taking into account the given assumptions, the stochastic process $Z(t)$ at time moment $t = t_k$ has normal distribution with the shift parameter equal to $\gamma m(t_k)$ and the scale parameter equal to $\sigma\sqrt{m(t_k)}$. In this case, the reliability function for the Wiener degradation model is given by:

$$S(t) = F\left(z_0; \gamma m(t), \sigma\sqrt{m(t)}\right) = \Phi\left(\frac{z_0 - \gamma m(t)}{\sigma\sqrt{m(t)}}\right),$$

where $\Phi(\cdot)$ is the standard normal distribution function.

The log-likelihood function for this model can be written as:

$$\ln L(\mathbf{X}_n) = \sum_{i=1}^{n}\sum_{j=1}^{k} \ln f(X_j^i; \gamma \Delta m(t), \sigma\sqrt{\Delta m(t)}),$$

where $\mathbf{X}_n$ is the sample of independent degradation index increments (1).

3 Inverse Gaussian Degradation model

The Inverse Gaussian (IG) degradation process $Z(t)$, $t \geq 0$ is defined as the stochastic process satisfying:

- $Z(0) = 0$;
- $Z(t)$ is a stochastic process with independent increments;

- increments $\Delta Z(t) = Z(t + \Delta t) - Z(t)$ have an inverse Gaussian distribution with probability density function

$$f(t; \gamma\Delta m(t), \sigma(\Delta m(t))^2) = \sqrt{\frac{\sigma(\Delta m(t))^2}{2\pi t^3}} exp\Big[-\frac{\sigma(\Delta m(t))^2(t-\gamma\Delta m(t))^2}{2\gamma^2(\Delta m(t))^2 t}\Big],$$

  where $\gamma\Delta m(t)$ is a shape parameter and $\sigma(\Delta m(t))^2$ is a scale parameter, $m(t)$ is a positive increasing trend function:

$$M(Z(t)) = m(t).$$

The inverse Gaussian distribution also has the following property: if random variables ξ_1 and ξ_2 follow the inverse Gaussian distribution, then $\xi_1 + \xi_2$ follows the inverse Gaussian distribution.

The reliability function for the IG degradation model is defined as

$$S(t) = F(z_0; \gamma m(t), \sigma m(t)^2) =$$
$$\Phi\Bigg(\sqrt{\frac{\sigma m(t)^2}{z_0}}\Big(\frac{z_0}{\gamma m(t)} - 1\Big)\Bigg) + exp\Big(\frac{2\sigma m(t)^2}{\gamma m(t)}\Big)\Phi\Bigg(-\sqrt{\frac{\sigma m(t)^2}{z_0}}\Big(\frac{z_0}{\gamma m(t)} + 1\Big)\Bigg),$$

where $\Phi(\cdot)$ is the standard normal distribution function.

The log-likelihood function for the IG degradation model can be written as:

$$\ln L(\mathbf{X}_n) = \sum_{i=1}^{n}\sum_{j=1}^{k} \ln f(X_j^i; \gamma\Delta m(t), \sigma(\Delta m(t))^2) =$$
$$\sum_{i=1}^{n}\sum_{j=1}^{k} \ln\Big[\sqrt{\frac{\sigma(\Delta m(t))^2}{2\pi X_j^i}} exp\Big[-\frac{\sigma(\Delta m(t))^2(X_j^i-\gamma\Delta m(t))^2}{2\gamma^2(\Delta m(t))^2 X_j^i}\Big]\Big].$$

4 The identification of the degradation model

In many publications, authors use the gamma or Wiener model as a degradation model, which describes the variation of the degradation index. However, there are some examples, where these models are not appropriate. One of the most popular approach for comparing probabilistic models is to use the information criteria, such as AIC and BIC. The AIC value is defined as follows:

$$AIC = 2k - 2ln(\hat{L}),$$

where k is the number of estimated model parameters and $\hat{L}$ is the maximum value of the likelihood function.

The BIC is formally defined as

$$BIC = ln(n)k - 2ln(\hat{L}),$$

where k is the number of estimated model parameters, n is the number of observations in $\mathbf{X_n}$ and $\hat{L}$ is the maximum value of the likelihood function.

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC or BIC value. However, the information criteria do not provide a test of a model in the sense of testing a null hypothesis of goodness-of-fit, so they can tell nothing about the quality of the model. If all the candidate models fit poorly, AIC and BIC will not give any warning of that. By this reason, we suggest to compare models by p-value of goodness-of-fit tests, such as Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests.

However, the degradation index increments have the distribution with different values of parameters, and the goodness-of-fit tests can not be applied directly. Thus, we introduce the following transformation:

$$U_j^i = F\left(X_j^i; \theta_j^i\right), i = \overline{1,n}, j = \overline{1,k_i}, \tag{2}$$

where θ_j^i is the vector of distribution parameters, which depends on the corresponding increments $\Delta m(t)$.

If the hypothesis H_0 on goodness-of-fit of the degradation model with distribution of increments $F(\cdot)$ is correct, then:

$$U_j^i \sim Uniform(0,1), \, i = \overline{1,n}, j = \overline{1,k_i}.$$

Thus, we need to test the uniform distribution for the random variates U_j^i, $i = \overline{1,n}, j = \overline{1,k_i}$.

We denote the elements of the variational series by $U_{(1)}^* \leq U_{(2)}^* \leq ... \leq U_{(M)}^*$, $M = \sum\limits_{i=1}^{n} k_i$, which constructed according to the sample

$$\mathbf{U}_M = \left\{U_j^i, i = \overline{1,n}, j = \overline{1,k_i}\right\}.$$

The value of the Kolmogorov statistic with Bolshev's correction [1] by sample data $\mathbf{U}_M$ is given by

$$S_k = \frac{6MD_M + 1}{6\sqrt{M}}, \tag{3}$$

where $D_M = \max(D_M^+, D_M^-)$, $D_M^+ = \max\limits_{1 \leq i \leq M} \left\{\frac{i}{M} - U_{(i)}^*\right\}$, $D_M^- = \max\limits_{1 \leq i \leq M} \left\{U_{(i)}^* - \frac{i-1}{M}\right\}$. The Cramer-von Mises Smirnov test statistic can be written by

$$S_\omega = M\omega_M^2 = \frac{1}{12M} + \sum_{i=1}^{M} \left\{U_{(i)}^* - \frac{2i-1}{2M}\right\}^2, \tag{4}$$

and the Anderson-Darling statistic can be presented as

$$S_\Omega = -M - 2\sum_{i=1}^{M} \left\{\frac{2i-1}{2M} \ln U_{(i)}^* + \left(1 - \frac{2i-1}{2M}\right) \ln(1 - U_{(i)}^*)\right\}. \tag{5}$$

Let us denote the distribution of a test statistic under hypothesis H_0 as $G(s|H_0)$. In the case of testing simple hypothesis the distributions $G(s|H_0)$ of the considered statistics do not depend on the tested distribution. Statistic S_K belongs to the Kolmogorov distribution, S_ω and S_Ω belong to the $a1$ and the $a2$ distributions, respectively. For composite hypothesis the test statistic distributions $G(s|H_0)$ are affected by a number of factors: the form of the tested distribution $F(x;\theta)$, the type and the number of estimated parameters, the method of parameter estimation, as well as the form of the trend function and the time moments of measuring degradation index.

The preferred model among all candidate models for the data is the one with the maximum p-value. Moreover, the p-value provide us the information about the quality of the chosen model: if p-value is rather high (larger than the given significance level), then we have no reason to reject the hypothesis of goodness-of-fit, and we can be sure that the chosen model fits the data.

The distribution of test statistics under true null hypothesis can be obtained according to the following algorithm:

1. Generate the sample of increments $\mathbf{X_n}$ basing on the degradation model under test, the appropriate plan of the experiment and time moments, in which the degradation index is measured.
2. Estimate model parameters by the sample $\mathbf{X_n}$ using maximum likelihood method.
3. Transform the sample of increments into the sample $\mathbf{U}_M$ according to (2).
4. Calculate the values of goodness-of-fit test statistics (1), (4) or (5) by the sample $\mathbf{U}_M$.
5. Repeat points $1-4$ N times, and obtain the empirical distribution $G_N(s|H_0)$.

Thus, we can calculate the p-value $\alpha_n = 1 - G_N(S_n|H_0)$, where S_n is a value of test statistics, calculated for the original sample of degradation increments.

5 Analysis of the reliability for the GaAs lasers data

In this section, we have considered the analysis of the degradation data of gallium arsenide (GaAs) lasers which have been taken from Meeker and Escobar [8]. Gallium arsenide (GaAs) lasers are used in telecommunication systems, processing of materials, various fields of medicine. The aging process of some lasers leads to deterioration of light output throughout the whole life cycle. The lasers fail when the consumption current exceed nominal value on 10%. Developing the lasers, engineers had some requirements: lasers need to work no less than 200000 hours under temperature of 20°C without failure. During the accelerated experiment 15 lasers were tested under the stress of 80°C for 40000 hours.

This data set has been extensively studied in modeling degradation processes [5], [9], [10], [11]. It has been demonstrated, in literatures, that the Gamma and Wiener degradation models do not fit to this data set.

In this paper, we have considered Gamma, Wiener and Inverse Gaussian degradation models. In accordance with the identification approach of degradation model, we have estimated parameters of these degradation models, as well as we computed the values AIC and BIC. The results are given in Table 1.

Table 1: The estimates of the distribution parameters of increments for the GaAs laser data

	Gamma model	Wiener model	IG model
$\hat{\gamma}$	0.002	0.013	$2.0372e^{-3}$
$\hat{\sigma}$	0.073	0.002	$5.4492e^{-5}$
AIC	-135.20	-87.14	-146.07
BIC	-134.44	-86.37	-145.31

Obtained values of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests statistics and corresponding p-values for all considered models are given in Table 2.

Table 2: The value of the test statistics and p-value for the considered models

	Kolmogorov		Cramer-von Mises-Smirnov		Anderson-Darling	
	Value S^*	p-value	Value S^*	p-value	Value S^*	p-value
Gamma model	1.076	0.011	14.503	0.029	75.083	0.107
Wiener model	0.069	0.006	35.385	0.017	179.867	0.093
IG model	0.764	0.208	0.046	0.593	0.330	0.539

It can be seen from tables above that p-values of all considered goodness-of-fit tests in the case of Wiener and Gamma degradation models are larger than p-values for the Inverse Gaussian degradation model. Also, values of AIC and BIC metrics are larger for the Inverse Gaussian degradation model. In this case, it can be concluded that the Inverse Gaussian degradation model is more appropriate degradation model for the GaAs laser data.

Conclusion

In this paper, the Gamma, Wiener and Inverse Gaussian degradation models have been considered. The identification approach of the degradation model using goodness-of-fit tests and information criteria have been developed. The example with the analysis of the GaAs lasers has been considered. It has been shown that the most appropriate model for the given data is the Inverse Gaussian degradation model.

Acknowledgment

The research is supported by the Russian Ministry of Education and Science (project 1.1009.2017/4.6)

References

[1] Bolshev L.N. and Smirnov N.V. (1983). *Tables of Mathematical Statistics*. Moscow: Science, (in Russian).

[2] Bordes L., Paroissin C., Salami A. (2010). Parametric inference in a perturbed gamma degradation process. *Statistics & Probability Letters*. P. 13.

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

[4] Tang L.C., Yang L.C., Xie M. (2004). Planning of step-stress accelerated degradation test. *Reliability and Maintainability Annual Symposium*.

[5] Wang X., Xu D. (2010). An inverse Gaussian process model for degradation data. *Technometrics*. Vol. **52**, pp. 188-197.

[6] Tsai C.C., Tseng S.T., Balakrishnan N. (2011). Mis-specification analyses of gamma and Wiener degradation processes. *Journal of Statistical Planning and Inference*. Vol. **12**, pp. 25-35.

[7] Pan D., Liu J.B., Cao J. (2016). Remaining useful life estimation using an inverse Gaussian degradation model.*Neurocomputing*. Vol. **185**, pp. 64-72.

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

[9] Ye Z., Chen N. (2014). The Inverse Gaussian Process as a Degradation Model. *Technometrics*. Vol. **56**, pp. 302-311.

[10] Peng C. (2014). Inverse Gaussian processes with random effects and explanatory variables for degradation data. *Technometrics*. Vol. **57 (1)**, pp. 100-111.

[11] Peng W.W., Li Y.F., Yang Y.J., Huang H.Z., Zuo M.J. (2014). Inverse Gaussian process models for degradation analysis: a Bayesian perspective. *Reliability Engineering & System Safety*. Vol. **130**, pp. 175-189.

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

The Prototype of the Cognitive Approach to Cancer Diagnosis According to Morphological Studies of the Stomach

Filimonov V.A.[1] and Mozgovoy S.I. [2]
[1] *Sobolev Institute of Mathematics, Omsk, Russia*
[2] *Omsk State Medical University, Omsk, Russia*
e-mail: filimonov-v-a@yandex.ru, simozgovoy@yandex.ru

Abstract

The problem of diagnosis of the gastric cancer is under consideration. Two problems of biopsy analysis are considered. The formulation and the solution of problems are oriented toward understanding by physician of the meaning of the used mathematical methods. By analogy with "4P medicine", "4K diagnostics" is suggested.

Keywords: *analysis of biopsy samples, structural and logical analysis, "4P medicine", "4K diagnostics", cross-technology.*

Introduction

A well-known phenomenon is the complexity of setting and solving problems associated with the application of mathematical approaches in medical diagnostics [6]. We describe here the episodes of diagnosis, reflecting the experience of many years interaction between the authors: a doctor and a mathematician. In the process of this interaction, many factors have to be taken into account, in particular, the one that is called the "systematic error of survivors" [3], as well as the subjectivity of the classification in the process of recognizing of images.

Previously, the authors proposed a model for the diagnosis of cancer based on the reaction of biological material to the chemical effect of a set of drugs (panel markers) [8, 9]. The marker was a specific reaction of a certain type. In this article, two stages of another type of diagnostics are considered. At the first stage, the problem of classifying a sample (biopsy) as being in a state of norm or pathology is considered. The starting material is the image of the biological material being analyzed.

At the second stage the task of testing of the status of the body is considered from which the given set of samples is taken. Separately, the important problem of explaining of the essence of the applied mathematical methods to the physician is considered. We do not describe here the use of the Bayes formula from the considerations of simplicity of explanation. The complexity of explaining it to doctors is represented well in [10]. We offer here a fairly intuitive combinatorial system, which was received well by the readers. The general task is very close to the task of creating of components for the explaining of the rationale for decisions made by expert systems.

The proposed diagnostic methods are preliminary, and require further theoretical and experimental studies. The Conclusion provides the considerations on the use of the arsenal of cognitive methods in medical diagnostics.

1 Analysis of biopsies by the series method

For the predictive assessment of gastric cancer, an approach based on the use of the mathematical model CriSS (Criterion of Serial Sorting) is proposed. The data structure in the model is represented by a square matrix, which is the image of the tissue cut. The task was to use a criterion to test hypotheses about the random nature of the grouping of pathological elements. The analysis of series in the chains of elements was used with this purpose. The general algorithm for estimating of one slice by the series analysis method can be represented as follows.

1. The original image is transformed into a square matrix (two-dimensional array), in which the pathological elements are encoded with the symbol "1", and all the rest with the symbol "0". The dimension of the matrix is determined from the conditions of the problem, the quality of the sample, and so on.

2. A two-dimensional array is transformed into a linear sequence (scanning through rows of the matrix).

3. The number of series is counted for this sequence, i.e. the number of fragments of the sequence consisting of identical symbols.

4. Items 1-3 are executed N times, every time the original image is rotated by $180°/N$ clockwise (or in the opposite direction). You can do twice as few turns, and the number of passes in the matrix analysis is increased due to the implementation in paragraph 2 of an additional pass through the columns.

5. A sequence with a minimum number of series w is chosen of all the resulting sequences.

6. This minimum number w is compared with the threshold W, in case it is less than or equal to W, the zero hypothesis about the absence of pathology is rejected.

To calculate thresholds and probabilities, we introduce the notation: n is the number of elements in the matrix without pathology (coded as "0"), m is the number of elements with a pathology sign (coded as "1"), k is the ratio of the elements; $n \sim k * m$ (this is an auxiliary parameter), w is the number of series in the sequence. W is a variable in the distribution function of the number of runs.

The total number of elements in the matrix is $n + m$, which gives the dimension of the matrix as an integer part of $R = \sqrt{n + m}$. In the cases we are considering, this dimension is roughly from 20 to 70. The values of the criterion are determined using

tables [2], including the asymptotic variants. The complexity of calculations requires further development of simple approximate methods of evaluation.

The series analysis method is able to detect the extended structures, however compact structures can be skipped. For example, a structure in the form of a circle or a square can be characterized by a large number of series. The primary sign of pathology is the presence in the matrix of one or more connected structures (such a structure is further called pre-focus) of pathological elements (symbols "1").

Connectivity is understood as the presence for each element of a structure of a certain type at least one neighbour of the same type. This makes it possible to construct paths connecting two any elements of the structure by passing only along the elements of this structure without going beyond it. The examples of simple connected structures are rectangles and chains. The set of elements in which the number of elements is greater than a given threshold M will be called the focus. The presence of at least one focus is considered to be the sufficient reason to reject the hypothesis that there is no pathology.

2 Threshold amount of pathological samples when assessing the stage of the disease

Sampling sets are taken of the three sections A, B, C of the organ (the patient's stomach), in the number of AN, BN, CN, respectively. In each of these sets, respectively, AG, BG, CG samples with signs of pathology were detected. The established stage S of the general development of disease is considered for each patient. The stages are coded with the numbers 2, 3, 4. An example of the representation of the data sets is given in Table 1.

Table 1: Fragment of the experimental data table

S	AG	AN	CG	CN	BG	BN
2	0	4	2	6	0	3
2	0	4	0	4	1	18
3	1	5	0	8	3	28
3	0	2	4	8	0	14
3	3	13	0	31	0	9
3	2	22	5	29	2	16
3	0	10	1	24	0	12
4	1	3	1	3	1	2
4	2	2	7	16	6	15
4	0	14	0	9	1	13

It is required to create a rule for deciding making on the stage S of the development of the disease of a particular patient based on the results of the morphological analysis

presented by the data of the structure indicated in Table 1. The situation of particular interest was that the total number of samples N was fixed: $N = 4$. In this case, each sample could be taken from any of the 3 parts of the organ. The results for this case are presented in Table 2.

Table 2: Frequency distribution of pathological samples depending on the stage of the disease

Number of pathological samples	0	1	2	3	4
S=4	0.15	0.36	0.36	0.11	0.02
S=2 or S=3	0.45	0.41	0.13	0.01	0.00

The preliminary rule was formulated as follows: in the presence of more than one pathological sample, a decision is made about the high risk of the disease.

3 Conclusions

The appearance of P4 medicine (predictive, personalized, preventive, participatory) requires the improvement of diagnostic methods. The development of the proposed approaches will make it possible to create a procedure of obtaining a predictable estimate of stomach cancer using a standard computer equipment quite simple and understandable for medical personnel.

On the other hand, the development of diagnostics requires the system integration of different approaches. At first we can note the cognitive format of the used methods. Equally important configuration is, i.e. the synthesis of various system representations of the object under consideration. The collective nature of multidisciplinary diagnostics is quite evident. Finally, diagnostics must be converging to specific conclusions. By analogy with "4P medicine" we propose to use the term "4C diagnostics": cognitive, configurational, collective and convergent.

Let us explain the term "configuration". The concept "configurator" was proposed by V.A. Lefebvre in 1967. The object of research is projected onto several screens reflecting various system representations of this object. The screens are connected to each other. The researcher can relate these representations. The simplest example is three planar projections of a 3D image of a part in the drawing. Another example is cognitive graphics using associative links. Cross-technologies [4] are one of the ways to implement configuration.

The results of experimental use of approaches will enable to formulate correctly the problem of human-machine diagnostics of stomach cancer using advanced computer technologies. We mention here the prospect of the use of cognitive analogs of sensory substitution [1, 7] in diagnostic procedures. One of the directions, which we called "music of the biopsy", is based on the addition of a visual image of the biopsy in an auditory way. Of itself, the convertion of an image into sound is a fairly well known

procedure. The problem is to select pre-focuses and focuses in the image, and to form a melody that helps the diagnosis.

References

[1] Abboud S, Hanassy S, Levy-Tzedek S, Maidenbaum S, Amedi A. (2013). EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution *Restorative Neurology and Neuroscience*. DOI: 10.3233/RNN-130338

[2] Bolshev L. N., Smirnov N.V. (1983). *Tables of mathematical statistics. (in Russian)*. Nauka Publ., Moscow.

[3] Ellenberg J. (2014). *How Not to be Wrong: The Hidden Maths of Everyday Life*. Penguin UK.

[4] Filimonov V. A. (2014). Cross-technologies of situational center as a testing ground of Cybernetics (In Russian) *Mathematical structures and modeling*. No. 3 (31), pp. 99-108.

[5] Hood L., Friend S.H. (2011). Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nature Reviews Clinical Oncology* V. 8, pp. 184-187 | doi:10.1038/nrclinonc.2010.227

[6] Kotov Yu. B. (2011). *New mathematical approaches to problems of medical diagnostics. (in Russian)*.URSS, Moscow.

[7] Lenay C., Gapenne O., Hanneton S., Marque C., Genouelle Ch. *Sensory substitution: limits and perspectives*.
https://web.archive.org/web/20061122061515/http://www.utc.fr/gsp/publi/Lenay03-SensorySubstitution.pdf

[8] Mozgovoy S.I., Kononov A.V., Schimanskaya A.G., Markelova M.V., Filimonov V.A. (2014). *Diagnosis of atrophy of the gastric mucosa on the basis of the panel of biomarkers. (in Russian)*. The certificate on the state registration of the computer program № 2014618512 from 22.08.2014.

[9] Mozgovoy S.I., Schimanskaya A.G., Nazarov A.N., Griscshenko R.K., Kononov A.V., Filimonov V.A. (2015). Predictive estimate of intestinal type gastric cancer by molecular cell markers using mathemathical models: a systematic analysis of the problem (In Russian) *Natural and technical Sciences* . N. 12 (90), pp. 154-160.

[10] Yudkowsky E.S. An Intuitive Explanation of Bayes' Theorem. Bayes' Theorem for the curious and bewildered; an excruciatingly gentle introduction. http://yudkowsky.net/rational/bayes

Powers of Some Tests for Exponentiality

Pavel Yu. Blinov and Boris Yu. Lemeshko
Novosibirsk State Technical University, Novosibirsk, Russian Federation
e-mail: blindizer@yandex.ru,lemeshko@ami.nstu.ru

Abstract

In the paper, some statistical tests for testing exponentiality have been considered. The distributions of tests statistics have been studied depending on sample sizes. Comparative analysis of the power of tests under different pairs of competing hypotheses has been conducted. Advantages and disadvantages of individual tests have been shown. Considered tests have been ranked by the test power.

Keywords: exponential distribution, hypothesis testing, test statistic, test power.

Introduction

The exponential distribution law is main distribution law used in reliability theory. Its analytical simplicity makes it attractive to engineers and researchers. However, you need to be sure, that behavior of observable random variable (for example, the moment's of product failure (breakdown)) is consistent by « desirable » exponential distribution before using this model. Otherwise, the benefit from computation simplicity will be repeatedly reduced by losses from conclusion incorrectness caused by deviation of empirical distribution from exponential distribution law.

There are a lot of papers devoted to exponential law, authors of these papers propose different statistical tests for testing hypothesis of exponentiality. The abundance of tests is caused by frequent use of exponential distribution model in applications. However, the frequency of using is defined that usage of simple model leads to the solution of problem grounded only on analytical methods in most cases.

In this paper, a lot of considered tests are studied by the method of statistical simulations. The number of experiments carried out for statistical modeling is assumed equal to 1 660 000 in the study of the distributions of test statistics.

1 The statement of exponentiality testing

Let $x_1, x_2, ..., x_n$ be sample of independent observations of nonnegative random variate X. Belonging of sample to exponential distribution law with density function $f(x) = exp(-x)$ was considered as tested hypothesis H_0.

The set of tests constructed special for exponentiality testing can be used for testing hypothesis H_0 besides classical goodness-of-fit tests. It is quite difficult to divide the special test statistics into the groups due to multiplicity of its. It should be noted that elements of ordered samples $x_1 < x_2 < ... < x_n$ are used in calculation for some test statistics. In another cases sequence order of elements doesn't matter.

Also, in some test statistics, we will use transformed values $z_1, z_2, ..., z_n$, which use estimate of shift and scale.

2 Alternative hypotheses

The exponential distribution has constant failure rate. The distribution laws corresponding to alternative hypotheses can be with: increasing, decreasing and non-monotonic failure rate [13]. The research was carried out for three alternative hypotheses:

$H_1 : LN(1)$ is lognormal distribution with denstity function $f(x) = \left(\theta x\sqrt{2\pi}\right)^{-1} exp\left(-(\ln x)^2/2\theta^2\right)$ and scale parameter $\theta = 1$ as alternative hypothesis with non-monotonic failure rate;

$H_2 : W(0.7)$ is Weibull distribution with density function $f(x) = \theta x^{\theta-1} exp(-x^{\theta})$ and form parameter $\theta = 0.7$ as alternative hypothesis with decreasing failure rate;

$H_3 : W(1.2)$ – is Weibull distribution with form parameter $\theta = 1.2$ as alternative hypothesis with increasing failure rate.

The distribution functions and density functions, which correspond to tested and alternative hypotheses, are presented on Figures 1 and 2, respectively.

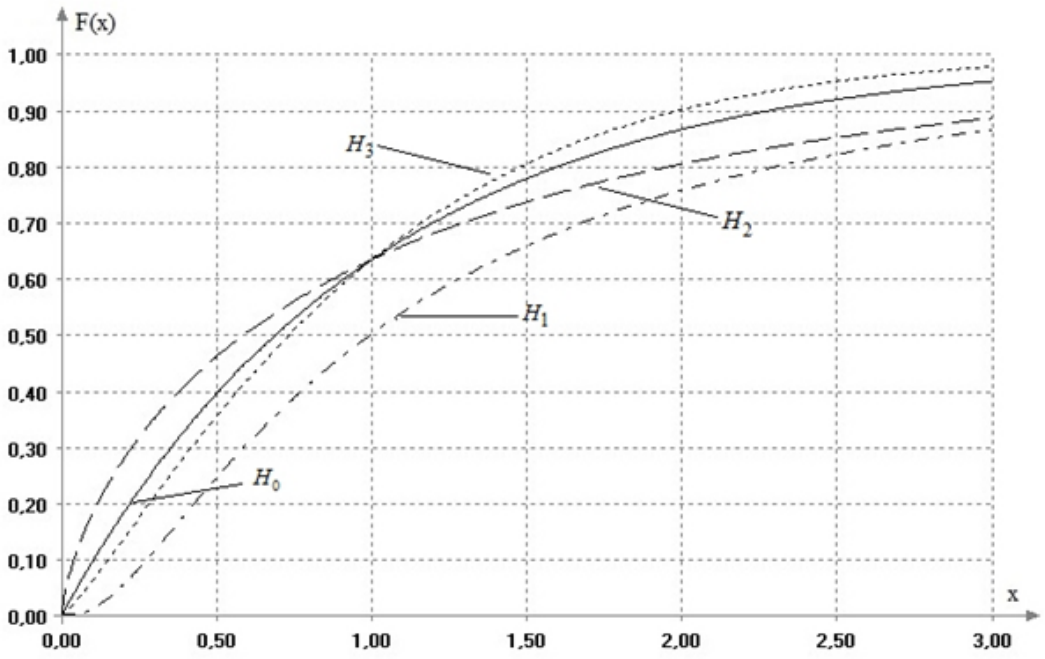


Figure 1: The distribution functions of hypotheses

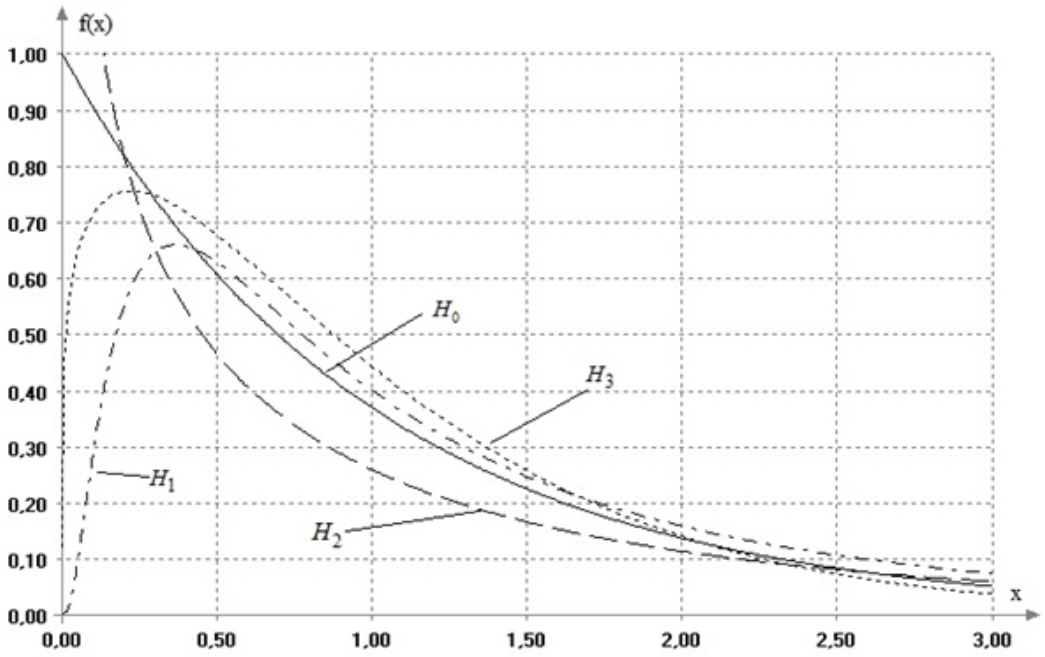


Figure 2: The density functions of hypotheses

3 Considered tests

In this section the description of considered tests are presented. The expressions for rest of test statistics of special uniformity tests are shown in Table 1, where $x_1 \leq x_2 \leq ... \leq x_n$ – ordered sample.

It is necessary to make the following remark. Although most of tests characterized as right-sided tests [8] by authors, but obtained results of research are contradicted to given assertions. The powers under one (or even two) considered alternative hypotheses with increasing of sample sizes leads to zero (values of statistic under validity of hypothesis H_i are less than under validity of hypothesis H_0) under supposition about right-sidedness of these tests. That fact can be described as follows: those tests considered like two-sided in power analysis. Only Frocini test, correlation tests and Kimber-Michael test are right-sided among analyzed tests, other considered tests are two-sided.

The correlation tests. Suppose that we have distribution law of probability $F(x) = 1 - exp\left(\frac{x-\mu}{\nu}\right)$ where μ and ν — non-known parameters, estimates of which can be calculate from formulas:

$$\hat{\nu} = \frac{n(\bar{x}-x_1)}{n-1};\ \hat{\mu} = x_1 - \frac{\hat{\nu}}{n}.$$

The test statistic based on correlation coefficient r for normalized variable $z_i = \frac{x_i-\hat{\mu}}{\hat{\nu}}$ with mathematical expectation i-th order statistic [14] of exponential distribution has the next form:

$$r(z,m)=\frac{\sum\limits_{i=1}^{n}(z_i-\bar{z})(m_i-\bar{m})}{\left[\sum\limits_{i=1}^{n}(z_i-\bar{z})^2(m_i-\bar{m})^2\right]^{1/2}},$$

where $m_i=\sum\limits_{j=1}^{i}\frac{1}{n-j+1}$. Under $n\geq 20$ approximation $\tilde{m}_i=-\ln\left(1-\frac{i}{n+1}\right)$ is possible and corresponding correlation coefficient is denoted by $r(z,\tilde{m})$.

The test statistics are used in form [8, 17]

$$K(z,m)=n\left[1-r^2(z,m)\right] \text{ and } K(z,\tilde{m})=n\left[1-r^2(z,\tilde{m})\right].$$

Kimber-Michael test. Kimber has developed the test [6] based on linear dependence between theoretical $F(x)$ and empirical distribution function $F_n(x)=\frac{i}{n}$ of random variables. For standardized exponential law $F(z_i)=1-exp(-z_i)$, where $z_i=\frac{x_i}{\bar{\nu}}$, $\bar{\nu}=\frac{1}{n}\sum x_i$ is standardized random exponential variable, in order to stabilize dependence and reduce influence of non-equal dispersions, Michel proposed as follows [11]:

$$s_i=\arcsin\sqrt{F(z_i)},\ r_i=\arcsin\sqrt{\tfrac{i-0.5}{n}}.$$

The test statistic has form

$$D=\max_i|s_i-r_i|.$$

Hollander-Proshan test. The test statistic has form [5]

$$T=\sum_{i>j>k}\varphi(x_i,x_j+x_k),$$

$$\varphi(a,b) \quad = \quad \begin{cases}1 & \text{if } a>b\\ 0 & \text{if } a<b\end{cases}$$

The normalized test statistic has form [5]:

$$T^*=\frac{T-M(T)}{\sqrt{D(T)}};\ \text{where } M(T)=\frac{n(n-1)(n-2)}{8},$$
$$D(T)=\frac{3n(n-1)(n-2)}{2}\left[\frac{5(n-3)(n-4)}{2592}+\frac{7(n-3)}{432}+\frac{1}{48}\right].$$

Klimko-Antle-Rademaker-Rockette test. The special exponentiality test based on equality testing of shape coefficient in Weibull distribution to one was considered in [7]. The test statistic can be written as follows

$$\tilde{c}=\sqrt{n}\left(\nu^{-1{,}075}-1\right);\ \nu=\frac{\sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}}{\sum\limits_{i=1}^{n}\left\{x_i-\left[x_1-\frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-x_1)\right]\right\}}$$

where x_1 is first element of ordered samples or minimum.

Table 1: Statistics of considered tests

Number	Test	Test statistic
1	Shapiro-Wilk 1 [15]	$W_E = \frac{n(\bar{x}-x)^2}{(n-1)\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$
2	Shapiro-Wilk 2 [15]	$\widetilde{W}_{E_0} = \frac{\left(\sum\limits_{i=1}^{n} x_i\right)^2}{n\left[(n+1)\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2\right]}$
3	Frosini [3]	$B_n = \frac{1}{\sqrt{n}} \sum\limits_{i=1}^{n} \lvert 1 - exp\left(-\frac{x_i}{\bar{x}}\right) - \frac{i-0.5}{n} \rvert$
4	Bartlett-Moran	$B = \frac{12n^2}{7n+1}\left[\ln\left(\frac{1}{n}\sum\limits_{i=1}^{n} x_i\right) + \frac{1}{n}\sum\limits_{i=1}^{n}\ln x_i\right]$
5	Sherman	$\omega_n = \frac{1}{2n}\frac{\sum\limits_{i=1}^{n}\lvert x_i-\bar{x}\rvert}{\bar{x}}$
6	Epps-Pulley	$c = \sqrt{48n}\left(\frac{1}{n}\sum\limits_{i=1}^{n} exp\left(-\frac{x_i}{\bar{x}}\right) - \frac{1}{2}\right)$
7	Max interval [1]	$\nu_n = \frac{\max(x_i - x_{i-1})}{\sum x_i}$, $i =$, $x_0 = 0$
8	Hartley [4]	$h(n) = \frac{\max x_i}{\min x_i}$
9	Kochar [9]	$T_n = \sqrt{\frac{108}{n}}\frac{\sum\limits_{i=1}^{n} J\left(\frac{i}{n+1}\right)}{\sum\limits_{i=1}^{n} x_i}$; $J\left(\frac{i}{n+1}\right) = 2\left(\frac{n+1-i}{n+1}\right)\left[1 - \ln\left(\frac{n+1-i}{n+1}\right)\right] - 1$

4 Simulation result

The Table 2 contains considered tests ordered by decreasing of power (quantity $1-\beta$) under alternatives H_1, H_2 and H_3 ($n = 50$ and $\alpha = 0.05$).

The best result are shown by Kimber-Michael test under hypothesis H_1, the power of this test are larger than powers of other tests considered. The correlation tests, Shapiro-Wilk tests, Frocini test and Klimko-Antle-Rademaker-Rockette demonstrate good powers as well.

It should be noted that worst results under hypothesis H_1 are shown by tests, limit distribution of which is normal distribution law (Epps-Pulley test [2], Sherman test [16], Hollander-Proshan test [5]).

The Bartlett-Moran test [12] shows the highest powers under alternative hypothesis H_2 and H_3. However, this test demonstrates average power under hypothesis H_1. The Epps-Pulley test, Frocini test and Hollander-Proshan test demonstrate consistently good ability to distinguish those alternative hypotheses from exponential distribution. The low power are presented by correlation tests (especially under hypothesis H_3), Hartley test and max interval test.

Table 2: The tests ranked by power ($n = 50$,$\alpha = 0.05$)

	hypothesis H_1	$1-\beta$	hypothesis H_2	$1-\beta$	hypothesis H_3	$1-\beta$
1	Kimber-Michael	0.516	Bartlett-Moran	0.890	Bartlett-Moran	0.315
2	Correlation test 2	0.377	Epps-Pulley	0.831	Epps-Pulley	0.304
3	Klimko-Antle-Rademaker-Rokette	0.359	Hollander-Proshan	0.818	Frosini	0.291
4	Shapiro-Wilk 1	0.359	Sherman	0.804	Hollander-Proshan	0.280
5	Correlation test 1	0.344	Frosini	0.804	Kimber-Michael	0.279
6	Frosini	0.310	Kochar	0.772	Sherman	0.277
7	Shapiro-Wilk 2	0.290	Kimber-Michael	0.706	Kochar	0.268
8	Max interval	0.254	Shapiro-Wilk 2	0.657	Shapiro-Wilk 2	0.266
9	Hartley	0.225	Klimko-Antle-Rademaker-Rokette	0.621	Klimko-Antle-Rademaker-Rokette	0.223
10	Kochar	0.218	Shapiro-Wilk 1	0.621	Shapiro-Wilk 1	0.223
11	Epps-Pulley	0.171	Max interval	0.347	Hartley	0.177
12	Bartlett-Moran	0.143	Hartley	0.319	Max interval	0.125
13	Sherman	0.140	Correlation test 2	0.311	Correlation test 1	0.053
14	Hollander-Proshan	0.109	Correlation test 1	0.276	Correlation test 2	0.016

Finally, it is significant that most powerful tests among considered tests are exponentiality Frocini and Kimber-Michael test.

Conclusions

Unfortunately, the distributions of most special uniformity tests depend on the sample size, therefore the possibility to estimate the significance level is absent and the researchers should use the tables of percent points.

In program system ISW, the possibility for simulation of the distributions of exponential test statistics in interactive mode is implemented as well as for simulation distribution of statistics of uniformity tests [10]. The estimate of p_{value} is calculated by empirical statistic distribution obtained as a result of modeling. This makes the statistical conclusions about the results of hypothesis testing more informative.

It is recommended to use some series of tests, which have certain advantages for more objective inferences.

Acknowledgements

This work is supported by the Russian Ministry of Education and Science (project 1.4574.2017/6.7).

References

[1] Gnedenko B.V., Belyaev Yu.K, Solov'ev A. D. (1965). *Mathemathical methods in reliability theory (in Russian)*. M:Nauka, Moscow.

[2] Epps T.W., Pulley L.B. (1986). A test of exponentiality vs. monotone-hazard alternatives from the empirical characteristics function. *Journal of the Royal Statistical Society: Series B*. Vol. **48**, pp. 206-216.

[3] Frosini B.V. (1987). On the distribution and power of a goodness-of-fit statistic with parametric and nonparametric applications *In: «Goodness-of-fit» Ed. by Revesz P., Sarkadi K., Sen P. K., Amsterdam-Oxford-New York: North-Holland. Publ. Comp..*, pp. 133-154

[4] Hartley H.O. (1950). The maximum F-ratio as a short-cut test of heterogeneity of variance. *Biometrika*. Vol. **37**, pp. 308-312

[5] Hollander M, Proshan F. (1972). Testing whether new is better than used. *The Annals of Mathematical Statistics*. Vol. **43**, pp. 1136-1146

[6] Kimber A.C. (1985). Tests for the exponential, Weibull and Gumbel distributions based on the stabilized probability plot. *Biometrika*. Vol. **72**, pp. 661-663

[7] Klimko L.A., Antle C.E., Rademaker A.W., Rockette H.E. (1975). Upper bounds for the power of invariant tests for the exponential distribution with Weibull alternative. *Technometrics*. Vol. **17**, pp. 357-360.

[8] kobzar A.I. (2006). *Applied Mathematical Statistics. For engineers and scientists. (in Russian)*. M:FIZMATLIT, Moscow.

[9] Kochar S.C. (1985). Testing exponentiality against monotone failure rate average. *Communications in Statistics – Theory and Methods*. Vol. **14**, pp. 381-392.

[10] Lemeshko B.Yu., Blinov P.Yu. (2015). *Tests for checking the deviation from uniform distribution law. Guide on the application. (in Russian)*. INFRA-M, Moscow.

[11] Michael J.R. (1983). The stabilized probability plot. *Biometrika*. Vol. **70**, pp. 11-17.

[12] Moran P.A.P. (1951). The random division of an intervals, II. *Journal of the Royal Statistical Society: Series B*. Vol. **13**, pp. 147-150.

[13] Rogozhnikov A.P., Lemeshko B.Yu. (2012). A Review of Tests for Exponentiality. *11th International Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*. Vol. **1**, pp. 159-166.

[14] Sarhah A.E. (1954). Estimation of the mean and standard deviation by order statistics. *The Annals of Mathematical Statistics*. Vol. **25**, pp. 317-328.

[15] Shapiro S.S., WIlk M.B.(1972). An analysis of variance test for the exponential distribution (complete samples). *Technometrics*. Vol. **14**, pp. 355-370

[16] Sherman B.A. (1950). A random variable related to the spacing of sample values. *The Annals of Mathematical Statistics*. Vol. **21**, pp. 339-361.

[17] Spinelli J.J., Stephens M.A.(1987). for exponentiality when origin and scale parameters are unknown. *Technometrics*. Vol. **29**, pp. 471-476.

On the Application of Homogeneity Tests

Boris Yu. Lemeshko, Irina V. Veretelnikova, Stanislav B. Lemeshko, Alena Yu. Novikova
Novosibirsk State Technical University, Novosibirsk, Russian Federation
e-mail: Lemeshko@ami.nstu.ru

Abstract

The properties of the homogeneity tests of Smirnov, Lehmann–Rosenblatt, Anderson-Darling, k-sampling tests of Anderson-Darling and Zhang are studied. For k-sampling Anderson-Darling test, models of limit distributions for a different number compared samples are built. A comparative analysis of the power of the homogeneity tests has been performe. The tests are ordered in terms of power relative to various alternatives. Recommendations on the application of tests are given.

Keywords: hypothesis testing, homogeneity test, Smirnov's test, Lehmann – Rosenblatt test, Anderson–Darling test, Zhang's tests, test power.

Introduction

Statistician constantly encounter with the need to solve problems of testing hypotheses about the belonging of two (or more) random variables samples to the same general population (homogeneity check) in various applications. In this case, there are problems of correct application and selection of the most preferable test. The problem of checking the homogeneity of samples is formulated as follows. Let x_{ij} be the j observation of the i sampling $j = \overline{1, n_i}, i = \overline{1, k}$. Let's pretend that $F_i(x)$ corresponds to i sample. It is necessary to test the hypothesis $H_0 : F_1(x) = F_2(x) = \cdots = F_k(x)$ for any x without specifying the common for them distribution law. The empirical distribution function corresponding to i sample is designated as $F_{in_i}(x)$.

In practice, two-sampling test of Smirnov [1] and Lehmann–Rosenblatt are most often used [1, 2, 3]. Significantly less mention is made of the use of the Anderson-Darling test [4] (Anderson-Darling-Petit) or its k-sampling [5], and even more rarely of the k-sampling variants of the Smirnov or Lehmann–Rosenblatt test [6, 7, 8] application. It is practically not said about the use of Zhang's homogeneity test [9, 10].

The goal of this paper, which is the development of [11], was to study the distributions of statistics and the homogeneity test power for limited sample sizes, to refine the sample sizes, from which one can use the limiting distributions, to clarify the nature of the alternatives concerning which tests have a poweradvantage. In carrying out the research, a computer simulation and analysis of statistical regularities methodology was used, which has proved itself in analogous works [12, 13, 14, 15, 16, 17, 18], based mainly on the statistical modeling method.

1 The tests under consideration

1.1 The Smirnov test

The Smirnov homogeneity test is proposed in [19]. It is assumed that the distribution functions $F_1(x)$ and $F_2(x)$ are continuous. The Smirnov test statistics measure the distance between the empirical distribution functions constructed from the samples

$$D_{n_1,n_2} = \sup_x \mid F_{1,n_1}(x) - F_{2,n_2}(x) \mid.$$

In practical use of the test of statistics D_{n_1,n_2} is calculated in accordance with the relations [1]:

$$D^+_{n_1,n_2} = \max_{1\le r\le n_1}[\tfrac{r}{n_1} - F_{2,n_2}(x_1 r)] = \max_{1\le s\le n_2}[F_{1,n_1}(x_2 s) - \tfrac{s-1}{n_2}],$$

$$D^-_{n_1,n_2} = \max_{1\le r\le n_1}[F_{2,n_2}(x_1 r) - \tfrac{r-1}{n_1}] = \max_{1\le s\le n_2}[\tfrac{s}{n_2} - F_{1,n_1}(x_2 s)],$$

$$D_{n_1,n_2} = \max(D^+_{n_1,n_2}, D^-_{n_1,n_2}).$$

If the hypothesis is valid statistics of the Smirnov test

$$S_C = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1,n_2} \tag{1}$$

the limit is a subject to the Kolmogorov distribution $K(S)$ [1].

However, for limited values n_1 and n_2 random variable D_{n_1,n_2} is discrete, and the number of its possible values is the smallest common multiple of n_1 and n_2 [1]. The stepwiseness of the conditional distribution $G(S_C \mid H_0)$ of statistics S_C with equal n_1 and n_2 remains even with $n_i = 1000$. Therefore, it is preferable to apply the test when the sample sizes n_1 and n_2 are not equal and are in fact the prime numbers.

Another drawback of the test with statistics (1) is that the distributions $G(S_C \mid H_0)$ with n_1 and n_2 and growth slowly approach the limiting distribution on the left and with bounded n_1 and n_2 substantially differ from $K(s)$. Thereby, a simple modification of the statistics (1) was proposed in [11]:

$$S_C M = \sqrt{\tfrac{n_1 n_2}{n_1+n_2}}(D_{n,m} + \tfrac{n_1+n_2}{4.6 n_1 n_2}),$$

which practically does not have the last drawback.

1.2 The Lehmann-Rosenblatt test

The Lehmann-Rosenblatt homogeneity test is a ω^2 type test. The test was proposed in [2] and was investigated in [3]. Statistics of the test is used in the form [1]

$$T = \frac{1}{n_1 n_2 (n_1 + n_2)} (n_2 \sum_{i=1}^{n_2} (r_i - i)^2 + n_1 \sum_{j=1}^{n_1} (s_j - j)^2) - \frac{4 n_1 n_2 - 1}{6(n_1 + n_2)}, \qquad (2)$$

where r_i- ordinal number (rank) x_{2i}; s_j – ordinal number (rank) x_{1j} in the combined variational series. It was shown in [3] that the statistic (2) in the limit is distributed as $a1(t)$ [1].

In contrast to Smirnov's test, the distribution of statistics converges rapidly to the limiting $a1(T)$. When $n_1 = n_2 = 100$ distribution visually coincides with $a1(T)$, while in practice deviation $G(T \mid H_0)$ from $a1(T)$ when $n_1, n_2 \geq 45$ can be neglected.

1.3 The Anderson–Darling test

The two-sampling the Anderson-Darling test (test for homogeneity) was considered in [4]. The statistics of the applied test is determined by the expression

$$A^2 = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1 + n_2 - 1} \frac{(M_i (n_1 + n_2) - n_1 i)^2}{i (n_1 + n_2 - i)}, \qquad (3)$$

where M_i – the number of elements in the first sample that are less than or equal to i element of the variation series of the combined sample.

The limiting distribution of the statistics (3) with the validity of the hypothesis being tested H_0 is the same distribution $a2(t)$ [4], which is the limit for Anderson-Darling's consent statistics.

Convergence of distribution $G(A^2 \mid H_0)$ statistics (3) $a2(A^2)$ with limited sample volumes was investigated in [20], where it was shown that when $n_1, n_2 \geq 45$ deviation of the distribution function $G(A^2 \mid H_0)$ $a2(A^2)$ does not exceed 0.01.

1.4 The k-sampling Anderson-Darling test

The k-sampling variant of the Anderson-Darling's consent test was proposed in [5]. Assuming continuity $F_i(x)$ the sample is built on the base of analyzed samples and generalized a total volume $n = \sum_{i=1}^{k} n_i$ and ordered $X_1 \leq X_2 \leq \cdots \leq X_n$. The statistics of the test has the form [5]:

$$A^2_{kn} = \frac{1}{n} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n-1} \frac{(n M_{ij} - j n_i)^2}{j (n - j)}, \qquad (4)$$

where M_{ij} – number of elements in i sample, which are not greater than X_j. The hypothesis to be tested H_0 deviates at large values of the statistics (4).

In [5], the table of upper percentage points is not presented for statistics (4), but for statistics of the form:

$$T_{kn} = \frac{A^2_{kn} - (k-1)}{\sqrt{D[A^2_{kn}]}}. \tag{5}$$

The parameter of the scale of statistics A^2_{kn} is given by [5]

$$D[A^2_{kn}] = \frac{an^3+bn^2+cn+d}{(n-1)(n-2)(n-3)}$$

at

$$\begin{gathered} a = (4g-6)(k-1) + (10-6g)H, \\ b = (2g-4)k^2 + 8hk + (2g-14h-4)H - 8h + 4g - 6, \\ c = (6h+2g-2)k^2 + (4h-4g+6)k + (2h-6)H + 4h, \\ d = (2h+6)k^2 - 4hk, \end{gathered}$$

where

$$H = \sum_{i=1}^{k} \frac{1}{n_i}, h = \sum_{i=1}^{n-1} \frac{1}{i}, g = \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \frac{1}{(n-i)j}.$$

Dependence of the limiting distributions of statistics (5) on the number of compared samples is k illustrates in fig.1. With increasing number of compared samples, this distribution slowly converges to the standard normal law.

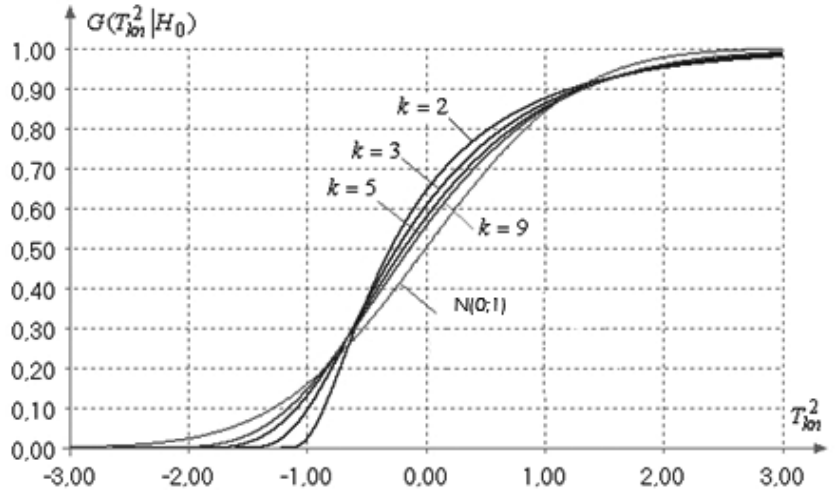


Figure 1: Dependence of limit distributions of statistics (5) of the number of samples being compared

The study of statistical distributions by methods of statistical modeling showed that when using test, the difference between the distributions of statistics from the corresponding limiting ones does not have practical significance for $n_i \geq 30$.

The table of upper percentage points of the statistic (5) limit distributions is presented in [5]. Also interpolation polynomials are constructed there, allowing to find critical values $T^2_{kn}(\alpha)$ for the number of samples being compared k, absent in the table.

As a result of studies of statistical distributions (5), statistical modeling ($n_i = 1000$ and the number of simulation experiments $N = 10^6$) we have somewhat refined and expanded the table 1 of critical values.

Table 1: Refined upper critical values $T^2_{kn}(\alpha)$ of statistics (5)

k	$1-\alpha$				
	0.75	0.90	0.95	0.975	0.99
2	0.325	1.228	1.966	2.731	3.784
3	0.439	1.300	1.944	2.592	3.429
4	0.491	1.321	1.925	2.511	3.277
5	0.523	1.331	1.900	2.453	3.153
6	0.543	1.333	1.885	2.410	3.078
7	0.557	1.337	1.870	2.372	3.017
8	0.567	1.335	1.853	2.344	2.970
9	0.577	1.334	1.847	2.323	2.927
10	0.582	1.3345	1.838	2.306	2.899
11	0.589	1.332	1.827	2.290	2.867
∞	0.674	1.282	1.645	1.960	2.326

Simultaneously, for the limiting distributions of statistics (5), approximate models of laws (for $k = 2 \div 11$) were built. Good models were [21] laws of the family of beta distributions of the third kind with density

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

for specific values of the law $B_{III}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ parameters, found on the basis of the statistics samples obtained as a result of modeling $N = 10^4$.

The models $B_{III}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ presented in table 2 with the given parametersvalues, allow to find p_{value} with an appropriate number k compared samples from the statistics values calculated from the relation (5).

When $k = 2$ the test with statistics (5) is equivalent in power to the two-sample Anderson-Darling test with statistics (3).

1.5 Test for the homogeneity of Zhang

The tests of homogeneity proposed by Zhang [9, 10] are the of the Smirnov, Lehmann-Rosenblatt and Anderson-Darling tests development enabling us to compare $k \geq 2$ samples. Zhang's goodness-of-fit test [9] show some advantage in power compared

Table 2: Models of limit distributions of statistics (5)

k	Model
2	B_{III}(3.1575, 2.8730, 18.1238, 15.0000, –1.1600)
3	B_{III}(3.5907, 4.5984, 7.8040, 14.1310, –1.5000)
4	B_{III}(4.2657, 5.7035, 5.3533, 12.8243, –1.7500)
5	B_{III}(6.2992, 6.5558, 5.6833, 13.010, –2.0640)
6	B_{III}(6.7446, 7.1047, 5.0450, 12.8562, –2.2000)
7	B_{III}(6.7615, 7.4823, 4.0083, 11.800, –2.3150)
8	B_{III}(5.8057, 7.8755, 2.9244, 10.900, –2.3100)
9	B_{III}(9.0736, 7.4112, 4.1072, 10.800, –2.6310)
10	B_{III}(10.2571, 7.9758, 4.1383, 11.186, –2.7988)
11	B_{III}(10.6848, 7.5950, 4.2041, 10.734, –2.8400)
∞	N(0.0,1 .0)

to the Kramer-Mises-Smirnov and Anderson-Darling goodness-of-fittests [22], but the drawback that limits the use of Zhang's test is the dependence of statistical distributions on sample volumes. The same drawback is possessed by variants of Zhang's test for checking the homogeneity of laws. To overcome this disadvantage, the author [9] proposes to use the Monte Carlo method for p_{value} estimation. The problem of modeling distributions of the Zhang homogeneity test statistics,is much simpler in comparison with a similar problem for the goodness-of-fittest, since it is necessary to model the distributions of statistics $G(S \mid H_0)$ in the case of analyzed samples belonging to the uniform law.

Let $x_{i1} \le x_{i2} \le \cdots \le x_{in_i}$ be ordered samples of continuous random variables with distribution functions $F_i(x)$, $(i = \overline{1,k})$ and $X_1 < X_2 < \cdots < X_n$, $n = \sum\limits_{i=1}^{k} n_i$, a combined ordered sample. Rank j of the ordered observation x_{ij} i sample in the combined sample is denoted as R_{ij}. Let $X_0 = -\infty, X_{n+1} = +\infty$, and ranks $R_{i,0} = 1, R_{i,n_i+1} = n+1$.

The modification of the empirical distribution function $\hat{F}(t)$ is used in the tests, which is equal $\hat{F}(X_m) = (m-0.5)/n$ [9] at break points $X_m, m = \overline{1,n}$.

Z_k the Zhang homogeneity test has the form [9]:

$$Z_K = \max_{1\le m\le n} \sum_{i=1}^{k} [F_{i,m} \ln \frac{F_{i,m}}{F_m} + (1 - F_{i,m}) \ln \frac{1 - F_{i,m}}{1 - F_m}], \tag{6}$$

where $F_m = \hat{F}(X_m)$, so that $F_m = (m-0.5)/n$, and the calculation $F_{i,m} = \hat{F}_i(X_m)$ is carried out as follows. At the initial moment the values are $j_i = 0, i = \overline{1,k}$. If $R_{i,j_i+1} = m$, then $j_i := j_i+1$ and $F_{i,m} = (j_i-0.5)/n_i$, otherwise, if $R_{i,j_i} < m < R_{i,j_i+1}$, then $F_{i,m} = j_i/n_i$.

Right-hand test: testable hypothesis H_0 deviates at **large** values of the statistics

(6). The distributions of statistics depend on n_i, k. Decision-making is influenced by the discreteness of statistics, which, with growth of k becomes less pronounced (see fig. 2).

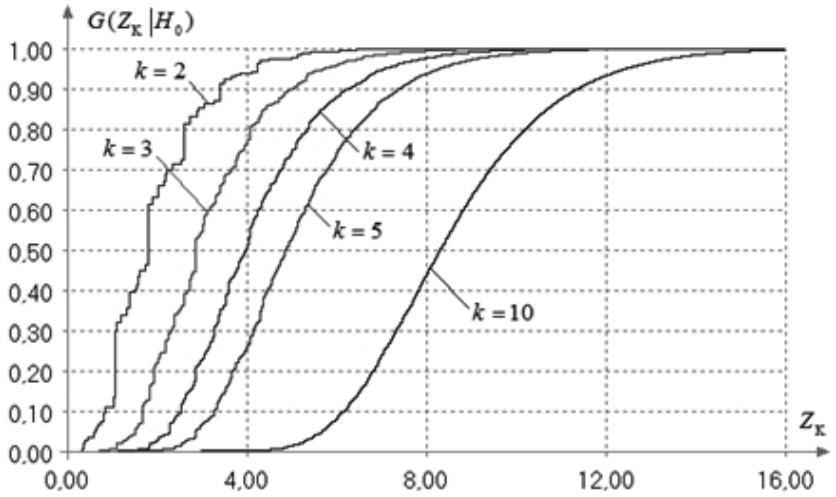


Figure 2: Dependence of the distributions of statistics (6) on k where $n_i = 20$

Statistics Z_A of the Zhang homogeneity testis determined by the expression [9]:

$$Z_A = -\sum_{m=1}^{n}\sum_{i=1}^{k} n_i \frac{F_{i,m}\ln F_{i,m} + (1-F_{i,m})\ln(1-F_{i,m})}{(m-0.5)(n-m+0.5)}, \tag{7}$$

where F_m and $F_{i,m}$ are calculated as defined above.

Left-sided test: verifiable hypothesis H_0 deviates for **small** values of the statistics (7). The distributions of statistics depend on n_i, k.

Statistics Z_C the test for homogeneity of samples is calculated in accordance with expression [9]:

$$Z_C = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} \ln\left(\frac{n_i}{j-0.5} - 1\right)\ln\left(\frac{n}{R_{i,j}-0.5} - 1\right). \tag{8}$$

The test is also left-handed: the hypothesis being tested H_0 deviates at **small** values of the statistics (8). The distributions of statistics depend on n_i, k.

The lack of information on the distribution laws of statistics and tables of critical values in modern conditions is not a serious disadvantage of Zhang's test, since in software supporting the application of test it is not difficult to organize the calculation of the achieved significance levels p_{value}, using methods of statistical modeling.

2 Comparative analysis of the test power

The power of homogeneity testing test has been investigated with respect to a number of alternatives. For definiteness, the hypothesis tested H_0 corresponded to the samples

with same standard normal distribution law with density

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

and the shift parameters $\theta_0 = 0$ and scale $\theta_1 = 1$.

With all alternatives, the first sample always corresponded to the standard normal law, and the second sample to some other one. In particular, with a shift alternative in the case of a competing hypothesis H_1 the second compilation corresponded to the normal law with the shift parameter $\theta_0 = 0.1$ and scale parameter $\theta_1 = 1$, in the case of a competing hypothesis H_2 – normal law with parameters $\theta_0 = 0.5$ and $\theta_1 = 1$.

When the scale is changed in the case of a competing hypothesis H_3 the second assembly corresponds to the normal law with parameters $\theta_0 = 0$ and $\theta_1 = 1.1$, in the case of a competing hypothesis H_4 – normal law with parameters $\theta_0 = 0$ and $\theta_1 = 1.5$.

In the case of a competing hypothesis H_5 the second assembly corresponded to the logistic law with density

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

and parameters $\theta_0 = 0$ and $\theta_1 = 1$. Normal and logistic laws are very close and difficult to distinguish using the goodness-of-fittest.

The obtained power estimates of the considered test for equal n_i when k with respect to competing hypotheses $H_1 - H_5$ – are presented in the table 3, where the test are ordered in descending order with respect to the corresponding H_i. Power ratings k-sampling tests where $k = 4$ with respect to competing hypotheses H_1, H_3, H_5 are given in the table 4.

Naturally, with the increase in the number of compared samples of the same volumes, the power of the test relative to similar competing hypotheses decreases. For example, it is more difficult to single out the situation and give preference to a competing hypothesis, when only one of the samples analyzed belongs to some other law. This can be seen by comparing the corresponding power ratings in Tables 3 and 4.

Table 3: Estimates of the power of test relative to alternatives $H_1 - H_5$ where $k = 2$ with equal n_i and $\alpha = 0.1$

Test	$n_i = 20$	$n_i = 50$	$n_i = 100$	$n_i = 300$	$n_i = 500$	$n_i = 1000$
Concerning the alternative H_1						
AD	0.114	0.137	0.175	0.319	0.447	0.691
LR	0.115	0.136	0.173	0.313	0.438	0.678
Z_C	0.114	0.134	0.164	0.278	0.382	0.600
Sm	0.111	0.132	0.164	0.280	0.381	0.617
Z_A	0.113	0.133	0.162	0.272	0.374	0.583
Z_K	0.111	0.126	0.152	0.238	0.333	0.526

Test	$n_i = 20$	$n_i = 50$	$n_i = 100$	$n_i = 300$	$n_i = 500$	$n_i = 1000$
Concerning the alternative H_2						
AD	0.435	0.768	0.959	1	1	1
LR	0.430	0.757	0.954	1	1	1
Z_C	0.425	0.743	0.946	1	1	1
Z_A	0.419	0.733	0.941	1	1	1
Sm	0.365	0.703	0.910	1	1	1
Z_K	0.344	0.650	0.906	1	1	1
Concerning the alternative H_3						
Z_A	0.108	0.128	0.164	0.318	0.464	0.745
Z_C	0.107	0.127	0.163	0.320	0.468	0.748
Z_K	0.107	0.127	0.154	0.268	0.390	0.624
AD	0.104	0.112	0.128	0.202	0.290	0.528
Sm	0.105	0.108	0.120	0.150	0.186	0.297
LR	0.103	0.107	0.114	0.149	0.191	0.324
Concerning the alternative H_4						
Z_A	0.267	0.651	0.937	1	1	1
Z_C	0.256	0.640	0.936	1	1	1
Z_K	0.248	0.552	0.849	1	1	1
AD	0.185	0.424	0.777	1	1	1
LR	0.154	0.280	0.548	0.989	1	1
Sm	0.152	0.288	0.510	0.964	0.999	1
Concerning the alternative H_5						
Z_K	0.105	0.110	0.122	0.179	0.266	0.429
Z_A	0.104	0.108	0.115	0.177	0.275	0.563
Z_C	0.104	0.108	0.116	0.172	0.265	0.556
AD	0.103	0.108	0.117	0.156	0.203	0.343
Sm	0.104	0.110	0.121	0.159	0.198	0.319
LR	0.103	0.106	0.113	0.142	0.178	0.288

Analysis of the obtained power estimates allows us to draw the following conclusions.

Concerning competing hypotheses corresponding to a change in the shift parameter, Smirnov's (Sm), Lehmann–Rosenblatt (LR), Anderson-Darling-Petite (AD) test and Zhang's test with statisticians Z_K, Z_A, Z_C in descending order are in the following order:

$$AD \succ LR \succ Z_C \succ Z_A \succ Sm \succ Z_K.$$

Concerning competing hypotheses corresponding to a change in the scale parameter, the test are already arranged in a different order:

$$Z_A \succ Z_C \succ Z_K \succ AD \succ LR \succ Sm.$$

Table 4: Estimates of the test power relative to alternatives H_1, H_3, H_5 where $k = 4$ with equal n_i and $\alpha = 0.1$

Test	$n_i = 20$	$n_i = 50$	$n_i = 100$	$n_i = 300$	$n_i = 500$	$n_i = 1000$
Concerning the alternative H_1						
AD	0.112	0.131	0.164	0.301	0.433	0.701
Z_C	0.111	0.126	0.155	0.260	0.368	0.595
Z_A	0.111	0.127	0.153	0.255	0.360	0.579
Z_K	0.109	0.121	0.141	0.219	0.300	0.502
Concerning the alternative H_3						
Z_C	0.106	0.122	0.158	0.306	0.468	0.761
Z_A	0.107	0.124	0.158	0.305	0.463	0.745
Z_K	0.106	0.120	0.145	0.249	0.367	0.606
AD	0.104	0.110	0.123	0.180	0.254	0.474
Concerning the alternative H_5						
Z_A	0.103	0.107	0.116	0.179	0.274	0.566
Z_C	0.103	0.107	0.115	0.173	0.257	0.555
Z_K	0.103	0.107	0.114	0.161	0.222	0.410
AD	0.102	0.106	0.113	0.143	0.179	0.291

However, the difference in the power with statisticians Z_A and Z_C to small. In a situation where, under a competing hypothesis, one sample belongs to the normal law and the second to the logistic one, the test are ordered in terms of power as follows:

$$Z_K \succ Z_A \succ Z_C \succ AD \succ Sm \succ LR.$$

When k sample in similar situations, the same order of preference is maintained for k-sampling variants of the Anderson-Darling and Zhang test. In particular, with respect to changing the shift parameter, the order of preference is:

$$AD \succ Z_C \succ Z_A \succ Z_K.$$

Regarding the change in the scale parameter –

$$Z_C \succ Z_A \succ Z_K \succ AD.$$

In this case, the test with statistics Z_A and Z_C are practically equivalent in power, and the Anderson-Darling test is noticeably inferior to all. Regarding the situation when the three samples belong to the normal law, and the fourth to the logistic one, the test are arranged according to the power in the following order:

$$Z_A \succ Z_C \succ Z_K \succ AD.$$

One can draw attention to the fact that the Zhang test have an advantage in power relative to the alternatives associated with changing scale characteristics, and are inferior in power under shift alternatives.

3 Application examples

The application of the test considered in the section for checking the homogeneity of laws is considered by analyzing the three samples below, each with a volume of 40 observations.

0.321	0.359	-0.341	1.016	0.207	1.115	1.163	0.900	-0.629	-0.524
-0.528	-0.177	1.213	-0.158	-2.002	0.632	-1.211	0.834	-0.591	-1.975
-2.680	-1.042	-0.872	0.118	-1.282	0.766	0.582	0.323	0.291	1.387
-0.481	-1.366	0.351	0.292	0.550	0.207	0.389	1.259	-0.461	-0.283

0.890	-0.700	0.825	1.212	1.046	0.260	0.473	0.481	0.417	1.825
1.841	2.154	-0.101	1.093	-1.099	0.334	1.089	0.876	2.304	1.126
-1.134	2.405	0.755	-1.014	2.459	1.135	0.626	1.283	0.645	1.100
2.212	0.135	0.173	-0.243	-1.203	-0.017	0.259	0.702	1.531	0.289

0.390	0.346	1.108	0.352	0.837	1.748	-1.264	-0.952	0.455	-0.072
-0.054	-0.157	0.517	1.928	-1.158	-1.063	-0.540	-0.076	0.310	-0.237
-1.109	0.732	2.395	0.310	0.936	0.407	-0.327	1.264	-0.025	-0.007
0.164	0.396	-1.130	1.197	-0.221	-1.586	-0.933	-0.676	-0.443	-0.101

The empirical distributions corresponding to these samples are shown in Fig. 3.

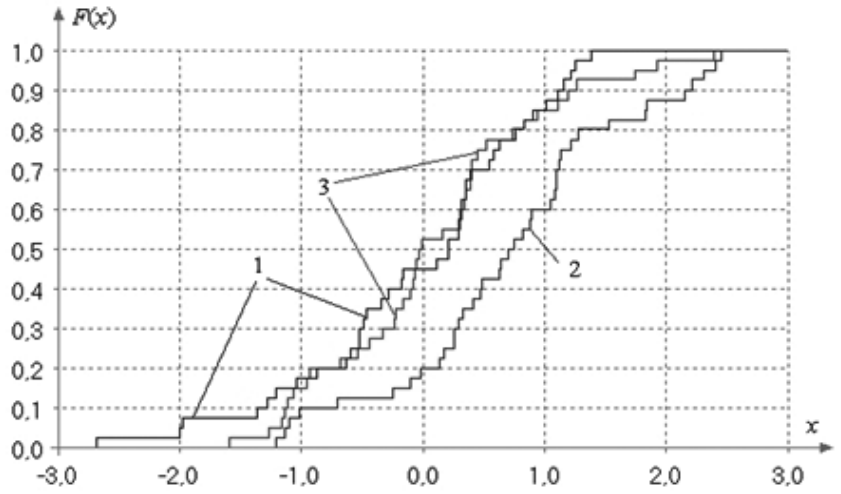


Figure 3: Empirical distributions corresponding to the samples

Let us test the hypothesis of homogeneity of the 1st and 2nd samples. Table 5 shows the results of the check: the values of the test statistics and the achieved significance levels p_{value}. Estimates p_{value} were calculated from the value of statistics in accordance with the distribution $a2(A^2)$ for the Anderson-Darling test, in accordance with the distribution $a1(T)$ for the Lehmann-Rosenblatt test, in accordance with the distribution $K(S)$ for the Smirnov test, in accordance with the beta distribution of the third kind from Table 2 for $k = 2$, k-sampling Anderson-Darling test. The

distributions of statistics (6), (7) and (8) of the Zhang test and estimates p_{value} were the result of modeling. It is obvious that the hypothesis of homogeneity should be rejected by all tests.

Table 6 shows the results of testing the hypothesis of homogeneity of the first and third samples. Here the estimates p_{value} by all test are very high, therefore the hypothesis of homogeneity to be tested should not be rejected.

Table 7 shows the results of testing the hypothesis of homogeneity of the three samples considered k-sampling Anderson-Darling and the Zhang tests. In this case, the estimate p_{value} for the Anderson-Darling test was calculated in accordance with the beta distribution of the third kind from Table 2 for $k = 3$, and for the Zhang test on the basis of statistical modeling carried out in an interactive mode, with the number of simulation experiments $N = 10^6$. The result shows that the hypothesis to be tested must be rejected.

Table 5: The results of checking the homogeneity of the 1st and 2nd samples

Tests	Statistics	p_{value}
Anderson-Darling	5.19801	0.002314
k-sampling Anderson-Darling	5.66112	0.003259
Leman-Rosenblatt	0.9650	0.002973
Smirnov	1.5625	0.015101
Smirnov's modified	1.61111	0.011129
Zhang Z_A	2.99412	0.0007
Zhang Z_C	2.87333	0.0008
Zhang Z_K	5.58723	0.0150

Table 6: The results of checking the homogeneity of the 1st and 3rd samples

Tests	Statistics	p_{value}
Anderson-Darling	0.49354	0.753415
k-sampling Anderson-Darling	-0.68252	0.767730
Leman-Rosenblatt	0.0500	0.876281
Smirnov	0.447214	0.989261
Smirnov's modified	0.495824	0.966553
Zhang Z_A	3.1998	0.332
Zhang Z_C	3.07077	0.384
Zhang Z_K	1.7732	0.531

In this case, the results of the test were fairly predictable, since the first and third samples were modeled in accordance with the standard normal law, and the resulting

Table 7: The results of testing the homogeneity of 3 samples

Tests	Statistics	p_{value}
k-sampling Anderson-Darling	4.73219	0.0028
Zhang Z_A	3.02845	0.0015
Zhang Z_C	2.92222	0.0017
Zhang Z_K	7.00231	0.0217

pseudorandom values were rounded to 3 significant digits after the decimal point. While the second sample was obtained in accordance with the normal law with a shift parameter of 0.5 and a standard deviation of 1.1.

Conclusions

Since the distribution of the statistics (2) converges very rapidly to the distribution, its use as a distribution of the statistics of the Lehmann-Rosenblatt test is correct also for small volumes of compared samples. The same can be said about the convergence of the distribution of statistics (3) of the Anderson-Darling homogeneity test to the distribution $a2(t)$.

The models of limited distributions of statistics (5)constructed in this paper using k-sampling homogeneity Anderson-Darling test for analysis k compared samples ($k = 2 \div 11$) gives an opportunity to find estimates p_{value}, which will undoubtedly make the statistical conclusion results more informative and substantiated.

In the case of the Smirnov test, due to the stepped nature of the statistics distribution (1) (especially, for equal sample sizes), the use of the Kolmogorov limit distribution $K(S)$ for the experimenter will be associated with a very approximate knowledge of the actual level of significance (the probability of error of the first kind) and the corresponding critical value. In case of constructing the procedures for testing homogeneity by the Smirnov test, it is recommended: 1) to choose $n_1 \neq n_2$ so that they are relatively prime numbers, and their least common multiple k was maximal and equal $n_1 n_2$; 2) Use a modification of Smirnov's statistics. Then the application of the Kolmogorov distribution as the distribution of the modified Smirnov test statistic will be correct for relatively small n_1 and n_2.

The test Zhang with statisticians Z_K, Z_A, Z_C with respect to some alternatives have a noticeable advantage in power. The drawback that limits their use is the dependence of the distributions of statistics on sample volumes. This disadvantage is easily overcome by using the Monte Carlo method to construct empirical distributions $G_N(Z \mid H_0)$ for statistics Z_K, Z_A, Z_C at specific sample sizes with subsequent evaluation of the values p_{value}. This procedure is easily realized, since in the construction $G_N(Z \mid H_0)$ comparable samples are modeled according to a uniform law on the interval [0,1]. When processing the measurement results in statistical quality management tasks, they usually deal with samples of a rather limited or very small

volume. It should be clearly understood that the homogeneity test due to low power for small sample sizes is not able to distinguish close competing laws. Therefore, the checked hypothesis about homogeneity of samples, even in the case of its injustice, will not be rejected more often. The shift to 0.1σ or an increase in the scaling parameter 10homogeneity, most likely, "will not be noticed", but large deviations in the laws corresponding to the samples will be noted. For example, in order that, if the Lehmann-Rosenblatt test is applied, the probabilities of errors of the first α and the second kind β did not exceed 0.1 in the presence of a shift 0.1σ (alternative H_1) the sample sizes should be of the order of 2000, and with the shift 0.5σ (alternative H_2) the likelihood of errors will not exceed 0.1 for a sample size of not more than 100.

Acknowledgements

The studies were carried out with the support of the Ministry of Education and Science of the Russian Federation in the framework of the state work "Ensuring the conduct of scientific research" (No. 1.4574.2017 / 6.7) and the design part of the state task (No. 1.1009.2017 / 4.6).

References

[1] Bol'shev L. N., Smirnov N. V. (1983).*Tables for Mathematical Statistics [in Russian]*. Nauka, Moscow.

[2] Lehmann E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Ann. Math. Statist.* Vol. **22**, No 1, pp. 165–179.

[3] Rosenblatt M. (1952). Limit theorems associated with variants of the von Mises statistic. *Ann. Math. Statist.* Vol. **23**, pp. 617–623.

[4] Pettitt A. N. (1976). A two-sample Anderson-Darling rank statistic. *Biometrika*. Vol. **63**, No.1, pp. 161-168.

[5] Scholz F. W., Stephens M. A. (1987). K-Sample Anderson–Darling Tests. *Journal of the American Statistical Association*. Vol. **82**, No. 399, pp. 918-924.

[6] Kiefer J. (1959). K-Sample Analogues of the Kolmogorov-Smirnov and Cramer-v. Mises Tests. *Annals of Mathematical Statistics*. Vol. **30**, No. 2, pp. 420-447.

[7] Conover W. J. (1965). Several k-sample Kolmogorov-Smirnov tests. *The Annals of Mathematical Statistics*. Vol. **36**, No. 3, pp. 1019-1026.

[8] Conover W. J. (1999).*Practical Nonparametric Statistics, 3rd ed.* John Wiley and Sons, New York.

[9] Zhang J. (2001).*Powerful goodness-of-fit and multi-sample tests. PhD Thesis*. York University, Toronto.

[10] Zhang J. (2006). Powerful Two-Sample Tests Based on the Likelihood Ratio. *Technometrics*. Vol. **48**, No. 1, pp. 95-103.

[11] Lemeshko B. Yu., Lemeshko S. B. (2005). Statistical distribution convergence and homogeneity test power for Smirnov and Lehmann–Rosenblatt tests. *MeasurementTechniques*. Vol. **48**, No. 12, pp. 1159–1166.

[12] Lemeshko B. Yu., Postovalov S. N., Chimitova E. V. (2011). *Statistical Data Analysis, Simulation and Study of Probability Regularities. Computer Approach: monograph.* NSTU Publisher. Novosibirsk.

[13] Lemeshko B. Yu., Lemeshko S. B. (2008). Power and robustness of test used to verify the homogeneity of means. *Measurement Techniques*. Vol. **51**, No. 9, pp. 950–959.

[14] Lemeshko B., Mirkin E. (2004). Bartlett and Cochran tests in measurements with probability laws different from normal. *Measurement Techniques*. Vol. **47**, No. 10, pp. 960–968.

[15] Lemeshko B. Yu., Lemeshko S. B., Gorbunova A. A. (2010). Application and power of test for testing the homogeneity of variances. Part I. Parametric test. *Measurement Techniques*. Vol. **53**, No. 3, pp. 237–246.

[16] Lemeshko B. Yu., Lemeshko S. B., Gorbunova A. A. (2010). Application and power of test for testing the homogeneity of variances. Part II. Nonparametric test. *Measurement Techniques*. Vol. **53**, No. 5, pp. 476-486.

[17] Lemeshko B. Yu., Sataeva T. S. (2017). Application and Power of Parametric Test for Testing the Homogeneity of Variances. Part III. *Measurement Techniques*. Vol. **60**, No. 1, pp. 7-14.

[18] Lemeshko B. Yu., Sataeva T. S. (2017). Application and Power of Parametric Test for Testing the Homogeneity of Variances. Part IV. *Measurement Techniques*. No. 5, pp. 12-17.

[19] Smirnov N. V. (1939). Otsenka rashozhdeniya mezhdu empiricheskimi krivyimi raspredeleniya v dvuh nezavisimyi hvyiborkah *Byul. MGU, Seriya A.*. Vol. **2**, No.2, pp. 3–14.

[20] Postovalov S. N. (2013).*Using of computer modeling for expanding application of classical methods of statistics hypothesis checking. DEA Thesis.* NSTU Publisher. Novosibirsk.

[21] Lemeshko B. Yu. (2017).*Tests for homogeneity. Guide on the application.* M: INFRA–M.

[22] Lemeshko B. Yu. (2014).*Nonparametric goodness-of-fit tests. Guide on the application.* M: INFRA–M.

Stochastic Algorithm to Solve the Problem of Linear Programming with Backward Calculations

E.B. Gribanova
Tomsk State University of Control Systems and Radioelectronics, Russia
e-mail: katag@yandex.ru

Abstract

The article focuses on solving linear programming tasks with stochastic algorithms. The cases when the use of stochastic algorithms are justified, as well as the advantages and disadvantages of such methods are indicated. A stochastic algorithm for finding a solution using backward calculations is suggested. Two examples of solving problems with two variables are considered.

Keywords: Linear programming, random search, constrained optimization, stochastic algorithms, economic models.

Introduction

Optimization methods are used in various fields and, in particular, in the economy, due to the need to make optimal decisions and to rationalize the allocation of available resources (people, machines, materials, money, information). The purpose of such methods is to define some optimum, such as the condition of the facility or the mode of its operation, which is, in terms of the chosen criterion, the best possible. The optimization task has the following appearance: find the value of an argument (arguments) that provides the best value (maximum or minimum) of the target function according to the existing constraints. If the number of arguments is greater than one, the task is called a multidimensional optimization task. A target function can have several local minimums at a given interval. In this context, local and global optimization tasks are encountered. Local optimization tasks assume that there is only one local minimum, the global optimization tasks define all of the local minimum of functions, or select the best one. In this work are considered linear programming problems in which all constraints and objective function are linear. In such tasks, the target function needs to be maximized/minimized:

$$f(x) = \sum_{j=1}^{n} c_j \cdot x_j \to \min$$

for some of the specified constraints:

$$h(x) = \sum_{j=1}^{n} a_{ij} \cdot x_j \{\leq, \geq, =\} b_i, i = 1, ..., m$$

$$x_j \geq 0, j = 1, ..., n.$$

where x_j - the required arguments; i - number of restrictions; j - number of variables; c_j, a_{ij}, b_j -some previously known values defined by the purpose of the task.

The fundamental role in the study of this kind of problems belongs to the by Kantorovich L.V. in 1939 published work, "Mathematical Methods of organization

and planning of Production", which described some of the optimization tasks of the economy (rational distribution of fuel, minimization of waste, best use of mechanisms, etc.) and suggested a method of allowing multipliers to solve them. Later Danzig J. has developed a simplex-method which is now the main method for dealing with linear programming.

This area of mathematical programming has been widely distributed, and the literature addresses a large number of tasks in different branches of the economy: industrial and agricultural planning, organization of transport logistics, distribution of investments in the construction of public utilities, development of a diversified policy of the organization, determination of prices for products to maximize revenue and many others.

1 Stochastic methods to solve the optimization problems

The methods used to solve optimization problems can be divided into stochastic, using random numbers, and deterministic. The use of stochastic methods for solving linear programming problems is caused by the following factors.

1. The task has an infinite number of decisions, but the results must be obtained. Also, the task may not have a solution, The use of stochastic methods in this case will yield the result, which is closest to the valid scope.

2. The number of arguments in the model is very large and/or they can only accept integer values (full or mixed integer programming tasks). The work [1] notes that in the order of 100, the exact algorithm is no longer able to find a solution in real time. The integer condition of the solution occurs when it is involved, for example, in finding the number of machines and people needed to complete the work. To the special case of the problem of integer programming, are meant problems in which variables can take only two values: 0 or 1 (classical task of tout, the task of optimizing the public transport movement, etc.). In this case, rounding the resulting solution to a whole does not always produce the best results, and using the branch and boundary method and the Gomori's method may require large computational resources and time.

3. A solution need to be obtained, that is close to the optimal task by using a simple algorithm, or it can searched for an initial solution for use by other methods.

4. It is necessary to verify the correctness of the found by using a different solution method.

5. Introduction to the task of linear programming in the initial stages of a material study. Simpler solutions provide a better understanding of the task.

6. It is necessary to generate several different variants of solving the problem by ordering them according to the value of the objective function in order to conclude how much the optimal solution differs from the nearest one, and to make the final choice, having weighed factors that can not be formalized [2]. The decision may also be related to the Pareto optimality study. For example, in a diet task, a set of products whose cost is slightly above the minimum value can provide more of the product (through discounts) and be preferred.

7. The functions (costs, limitations) considered in the task are not continuous and differentiable (e.g. have a piecewise view or are calculated algorithmically). In the study of transport problems, tasks in such a formulation are often considered when the delivery cost is a piecewise function, depending on the quantity, which is related to the discounts provided by suppliers. In a diet task, you can see different rules for granting discounts, free extra unit of goods, and so on.

The disadvantage of using stochastic algorithms is that the solution obtained is suboptimal, and will differ in different iterations.

The simplest stochastic method for solving linear programming problems is random search, its appearance and development in our country is connected with the research of Rasstrigin L.A. Despite the fact that a simple random search can be effective for solving many problems, there are two shortcomings: A large number of implementations are required to find a solution; In the global optimization task, it is possible to find a local minimum instead of a global level. In this regard, various modifications to the algorithm are being developed to improve the effectiveness of the search and to improve it to determine the global minimum.

The problem of linear programming by generating random numbers can be represented by the following algorithm:

1. To specify the initial values x_i that are accepted as the current solution, and to calculate the value of the target function.

2. To generate random values x_i from specified intervals.

3. To check that the existing constraints are met. If the values lie within the scope of the available solutions, to go to step 4, or to return to step 2.

4. To calculate the value of the target function. If the resulting value is less than the value corresponding to the current solution, to remember it as the current solution. If the necessary number of iterations has been executed, then the work of the algorithm is completed, otherwise the transition to step 2 is carried out.

A random search option is a two-way search: From the value x_i a step is set in two directions: $x_i \pm \Delta x$ and the random value is generated from this interval, i. e. the boundaries of the interval depend on the current solution. It should be taken into account that the values of the variables can not take negative values, therefore, for example, if the value of the variable is zero, only numbers from the interval (0;

$x_i + \Delta x$) will be modeled. With the increase in the number of iterations, the resulting solution will approach the optimal one.

A punitive function is also used to solve conditional tasks [3,4]. To the expression of the objective function, a term equal to zero is added if the conditions are met and some value, if the conditions are not satisfied. Next, the task is solved by a simple random search without restrictions. This method allows you to find the closest solution to a valid scope if there is no solution.

The article [1] introduces greedy algorithms for conditional pseudo-Boolean optimization, where the target function needs to be maximized and the constraint is lesser or equal. In these algorithms, from the possible solutions of the next level that satisfy the constraint, one is chosen for which $\frac{f(x)}{h(x)}$ the value is the maximum.

In a variable probability method (VPM) [5], a new solution is received according to the probability vector that is subsequently changed based on the search result. This method was originally designed to meet the challenges of unconditional pseudo-Boolean optimization, and later modifications were developed to meet the limitation challenges, in particular in the work [3] a parallel implementation is presented to enable the resolution of the conditional tasks of a large size.

The work [6] presents two adaptive algorithms of unconditional global optimization (ARSET and DRASET), where the search scope is first reduced to determine the local minimum, and after a solution with an acceptable precision has been found, it expands.

Tabu search [7] allows to avoid circularity with a return to the local minimum in solving optimization problems by creating a list of prohibitions based on previous solutions.

To increase the efficiency of finding a solution, heuristic algorithms that model the behavior of animate and inanimate nature are also used, which are more complex than random search. Among them, a large class is formed by evolutionary algorithms that model the mechanisms of natural selection, whose research became popular in the early 1960s and is associated with the names of Alex Fraser, Elliot Burnel, Jack Crosby. The most popular algorithm for linear programming is the genetic algorithm that belongs to the evolutionary algorithm class and implies the following mechanisms for finding a solution: selection, crossing, mutation [2]. The downside of these evolutionary algorithms is more complex program logic, and the creation of objects requires additional computational resources.

The presented work is aimed at increasing the efficiency of random search by developing an algorithm based on the generation of random variables according to existing constraints, which allows to simulate numbers with a smaller spread around the range of admissible values. Also in the presented method, operations similar to breeding and crossing in the genetic algorithm are performed.

2 Stochastic algorithm to solve the problem of linear programming

Let's consider the solution of the linear programming problem by generating random variables. Here the points are modeled without taking into account the existing constraints, therefore only a part of them will lie in the range of admissible values, from which the choice is made according to the value of the objective function. Figure 1 provides an example of a task solution ($x_1 \geq 0, x_2 \geq 0$):

$$f(x) = 17x_1 + 58x_2 \to \min \tag{1}$$

$$\begin{cases} 40x_1 + 1,5x_2 \geq 75; \\ 0,2x_1 + 25,2x_2 \geq 14. \end{cases}$$

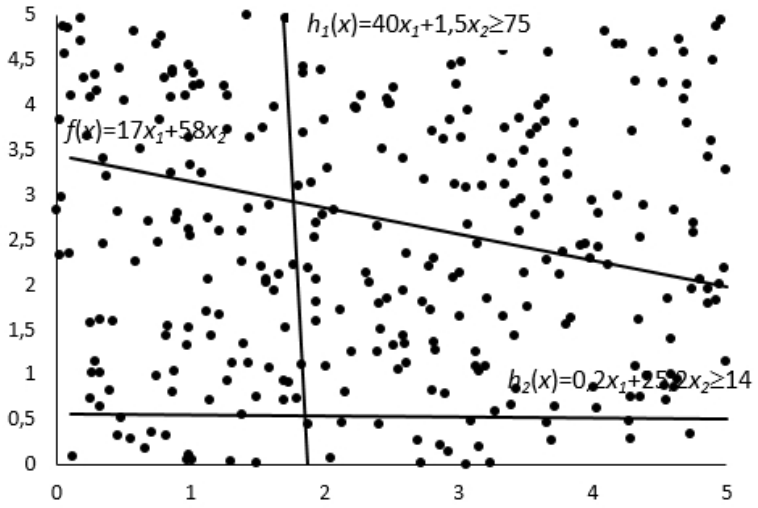


Figure 1: Solving a task by using a random search

Solution received in this implementation: $x_1 = 2,153, x_2 = 0,812, f(x) = 83,7$. The proposed approach is based on the generation of random values according to the existing constraints, so that for each of them the inverse problem is solved by inverse calculations. In this case, the structure of the scorecard is represented as a tree, where the top-level element of h_i sets the desired target, which is achieved by modifying the lower-level elements of the x. The resulting solutions are then combined. Let's consider the algorithm for solving the problem:

1. At a given interval, to generate a point whose coordinates are accepted as initial points.

2. To generate m solutions by solving an inverse problem with the help of inverse calculations for each constraint (assuming that the relation has an equal sign). Inverse calculations were offered by Odinzov B.E.[8] and are designed to meet

the challenges of the following: Determination of the Δx increments of the arguments of the direct function on the basis of their initial value, the desired value of the function and additional information coming from the person, which may be the coefficients of the relative importance of the arguments (α, β). The values of the relative importance factors can also be simulated randomly or calculated on the basis of equation factors. In general, in the case of the two arguments, the task is to solve the system of equations:

$$\begin{cases} y \pm \Delta y = f(x_1 \pm \Delta x_1(\alpha), x_2 \pm \Delta x_2(\beta)); \\ \frac{\Delta x_1}{\Delta x_2} = \frac{\alpha}{\beta}; \\ \alpha + \beta = 1; \end{cases}$$

where $\Delta x_1, \Delta x_2$ is the increment of the arguments; α, β - coefficients of relative importance of the arguments, x_1, x_2 respectively; $y, \Delta y$ - Initial value and increment of the resulting function.

Thus, in order to generate a solution, it is necessary to randomly simulate factors of relative importance and then to define values Δx. For the solutions that you have received, to perform the "crossing" operation. The implementation of this mechanism depends on the type of constraint. If the two restrictions are in the form of "greater or equal", then "crossing" operation can be written as a formula (x_k^f the the biggest value of the two solutions, a random number r between 0 and 1):

$$x_k = x_k^f + r\left(x_k^m + x_k^f\right).$$

3. The solutions are ranked and the best one is chosen. The t of the best solutions is preserved. Return to step 1.

The use of adaptive random searches can improve the efficiency of the algorithm. Applied to the problem under consideration, the conditions for adaptive change of parameters will have the following form (the simulation of the point x is carried out from the interval $(x_{mid} - \Delta; x_{mid} + \Delta)$):

- If the solution does not satisfy the constraints, it is to increase Δ;
- If the modeled solution is in a valid scope and is the best of the previous one, then x_{mid} it becomes equal to the resulting point and the step is reduced Δ.

In this case, the minimum and maximum step values are specified, and the output of values beyond the established limits is checked.

3 Experimental Results

The solution to the task (1) using the (nonadaptive) algorithm presented is shown in Figure 2. Solution received in this implementation ($r = 0,01$): $x_1 = 1,854, x_2 = 0,546, f(x) = 63,20$.

Let's look now a case where the target function is not continuous:

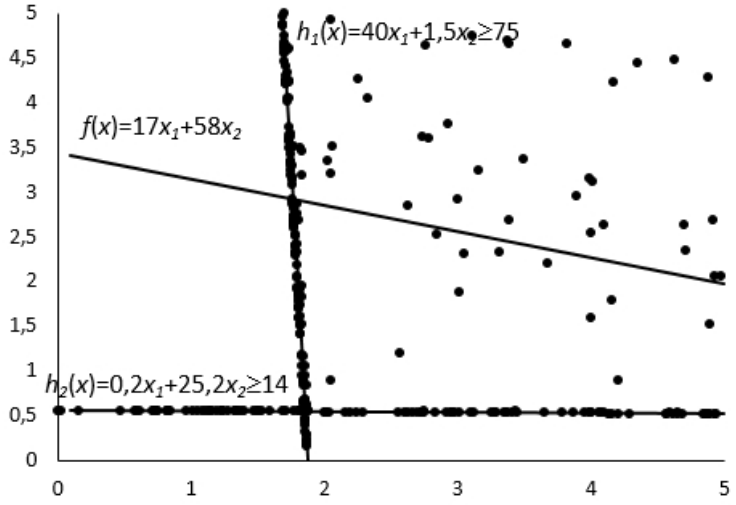


Figure 2: Solving a task using an algorithm that uses backward calculations

$$f(x) = s_1 \cdot x_1 + s_2 \cdot x_2 \to \min,$$

where

$$s_1 = \begin{cases} 30, x_1 < 1; \\ 25, x_1 \geq 1. \end{cases} \quad s_2 = \begin{cases} 40, x_2 < 1; \\ 35, x_2 \geq 1. \end{cases}$$

Restrictions on the composition of products:

$$\begin{cases} 5x_1 + 7x_2 \geq 10; \\ 7x_1 + 3x_2 \geq 5. \end{cases}$$

Figure 3 provides a histogram that shows the deviation of the solver from the optimal solution. The graph shows average values, and the simulation was performed within 1000 implementations.

Conclusions

The possibility of using stochastic algorithms for solving linear programming problems is investigated in the article. An algorithm for solving the problem is proposed by generating random numbers according to existing constraints. For this purpose, the back calculation device was used, followed by a combination of the solutions obtained. As an example, the solution of linear programming problems with two variables is considered. The introduced algorithm made it possible to find a more accurate solution in comparison with a simple random search.

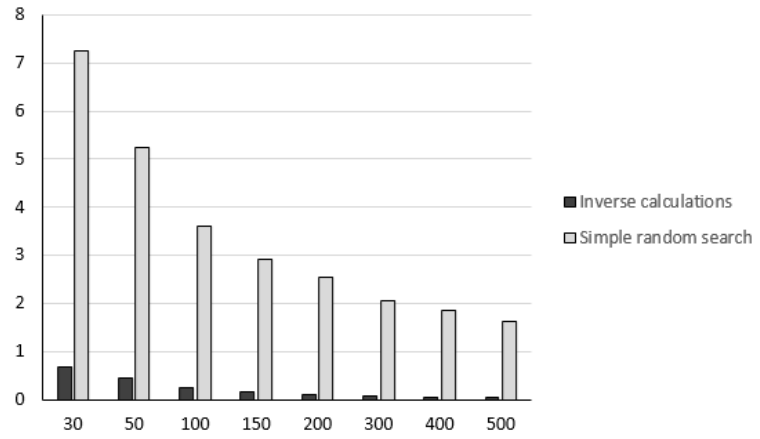


Figure 3: Rejection of the received decisions from the optimal

References

[1] Antamoshkin A.N. (2003). Greedy and Local Search Algorithms for the Conditional Pseudoboolean Optimization. *Investigated in Russia*. Vol. **177**, pp. 2143-2149.

[2] Datta S., Garai C., Das C. (2012). Efficient Genetic Algorithm on Linear Programming Problem for Fittest Chromosomes. *Journal of Global Research in Computer Science*. Vol. **6**, pp. 1-7.

[3] Kazakovtsev L.A., Stupina A.A. (2012). Parallel Realizationof the Probability Changing Method. *Problems of modern science and education*. Vol. **4**, pp. 1-9.

[4] Gribanova E.B. (2016). Stochastic algorithms to solve the economic analysis inverse problems with constraints. *Reports of TUSUR*. Vol. **19**, pp. 112-116.

[5] Antamoshkin A.N., Kazakovtsev L.A. (2014). Application of the probability changing method for optimal location problems on a network. *Vestnik SibGAU*. Vol. **57**, pp. 10-19.

[6] Hamzacebi C., Kutay F. (2006). A heuristic approach for finding the global minimum: adaptive random search technique. *Applied Mathematics and Computation*. Vol. **173**, pp. 1323-1333.

[7] Glover F. (1989). Tabu search—Part I. *INFORMS Journal on Computing*. Vol. **1**, pp. 190-206.

[8] Odintsov B.Ye. (2004). *Inverse calculations in Making Economic Decisions*. Finance and Statistics, Moscow.

How to Detect Topology of a Manifold to Approximate Multidimensional Data

Michael G. Sadovsky[1,2] and Anatoly N. Ostylovsky[2]
[1] *Institute of computational modelling SB RAS, Krasnoyarsk, Russia*
[2] *Siberian federal university, Krasnoyarsk, Russia*
e-mail: msad@icm.krasn.ru, hinayana@g-service.ru

Abstract

New method is proposed to identify topology of a low-dimensional manifold approximating multidimensional datasets. The method is based on the implementation of the compliment for the discrete set of data. Some essential properties and constraints of the method are discussed. New method is proposed to identify clusters in datasets. The method is based on a sequential elimination of the longest distances in dataset, so that the relevant graph looses some edges. The method stops when the graph becomes disconnected.

Keywords: order, unexpectedness, cluster, connectivity.

Introduction

Clustering techniques become number one issue in up-to-date methodology of massive data treatment. It should be stressed that there is no rigid, unambiguous and self-consistent problem formulation, for clustering. Whether such formulation is possible, is an arguable matter, itself. There is no trick here: a variety of admissible configurations of datasets is huge, and exceeds any list of methods to cluster them.

Thus, a modern approach to treat the multidimensional data is to model them and approximate with manifolds of lower dimension. Generally speaking, such approach had taken start in principle component analysis (PCA), with linear vector space as approximating manifold. The point is that PCA is the linear procedure; it hardly could be feasible for analysis of data exhibiting significant non-linearity.

A sounding progress has been achieved in this direction, both in theory [1, 2, 3], and in specific applications [10, 11, 12]. Meanwhile, the topology of an approximating manifold becomes the essential constraint here, since the key idea of the approximation is the maintenance of the topology of the manifold used to model (or approximate) the data. Yet, even a simple configuration (that is a two-dimensional torus) makes a problem: one fails to identify reliably any cluster pattern within such data set, if a genus-one manifold is not used to approximate the dataset. A growth of the data set dimension just makes the problem worse.

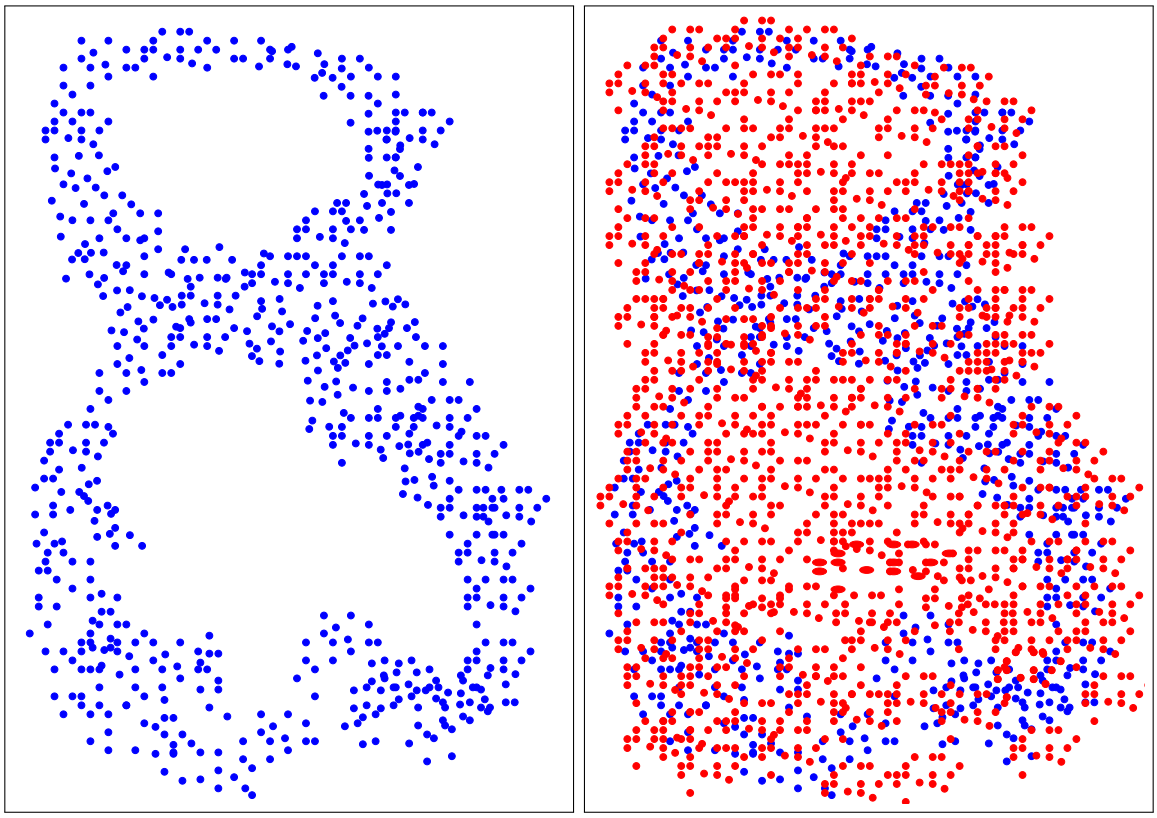
1 The method

Basically, the idea of the method is to change a study of the original set for the complement of that former. Let $\mathfrak{M}$ be the set of multidimensional data points $\mathfrak{m}_j \in$

$\mathfrak{M}$ with index j enlisting the points in it, so that $|\mathfrak{M}| = M$ and $\forall j\ \mathfrak{m}_j \in \mathbb{R}^n$; $\mathfrak{m}_j = (x_1^j, x_2^j, x_3^j, \ldots, x_{n-1}^j, x_n^j)^\intercal$. Here $|\cdot|$ is the capacity of a set.

Suppose then, that the point $\mathfrak{m}_j \in \mathfrak{M}$ are located in space rather densely, and one is always able to distinguish quite a simple body Ω gathering $\mathfrak{M}$ outer space (see Fig. 1(a) for illustration). let the density of $\mathfrak{M}$ in Ω be equal to d. Let color all the points $\mathfrak{m}_j \in \mathfrak{M}$ in blue. Disperse then randomly and independently another set $\mathfrak{L}$ of points $\mathfrak{l}_j \in \mathfrak{L}$ over Ω, colored in red (see Fig. 1(b)), supposing that $|\mathfrak{M}| \sim |\mathfrak{L}|$.

Finally, to reveal the gaps and breaches in $\mathfrak{M}$, one must eliminate all the red points located closely enough to blue ones. There could be few ways to identify those points, see Sec. 1.1. As soon, as the added (red) points are eliminated, one must eliminate the original set $\mathfrak{M}$ (blue points), so that the rest points colored in red to represent the complement $\widehat{\mathfrak{L}}$ to $\mathfrak{M}$; $\widehat{\mathfrak{L}} \subset \mathfrak{L}$. Finally, one should study the complement with a number of convenient methods and techniques; Figs. 1 and 2 illustrate the method.



(a) Original set $\mathfrak{M}$.

(b) Superposition of the sets $\mathfrak{M}$ and $\mathfrak{L}$.

Figure 1: Illustration of the method to develop the complement $\widehat{\mathfrak{L}}$.

1.1 How to remove proximal points

Let now discuss three more issues in the method implementation; these are the method to identify the red (added) points located closely enough to the original

(blue) ones (Sec. 1.1); the problem of a "fuzzy" pattern of an original set $\mathfrak{M}$ and the choice of parameters of the distribution of the added points set $\mathfrak{L}$.

A simple idea is to eliminate the added (red) points located close enough to the points $\mathfrak{m}_j \in \mathfrak{M}$. Since the point $\mathfrak{m}_j \in \mathfrak{M}$ a stipulated to belong to some metric space $\mathbb{R}^n$, then one is always able to determine a distance between $\mathfrak{m}_j$ and $\mathfrak{l}_k$. Nonetheless, there are two options here to define the exclusion rule.

I. A red point to be eliminated is determined locally, with respect to the nearest blue ($\mathfrak{m}_j \in \mathfrak{M}$) points.

II. Another way is to develop some effective "mean field" function to provide a discretion of the points $\mathfrak{l}_k \in \mathfrak{L}$ to be eliminated.

Let now discuss these two options in more detail.

The first option is to cover each point $\mathfrak{m}_j \in \mathfrak{M}$ with a ball of the radius r, and label all the added (red) points located inside the union of those balls. Remove the labeled points $\mathfrak{l}_j \in \mathfrak{L}$, thus making the complement $\widehat{\mathfrak{L}}$ of $\mathfrak{L}$ that is supposed to represent the gaps and breaches in $\mathfrak{M}$. The radius r here is the fitting parameter, and the structure of $\widehat{\mathfrak{L}}$ depends on that latter.

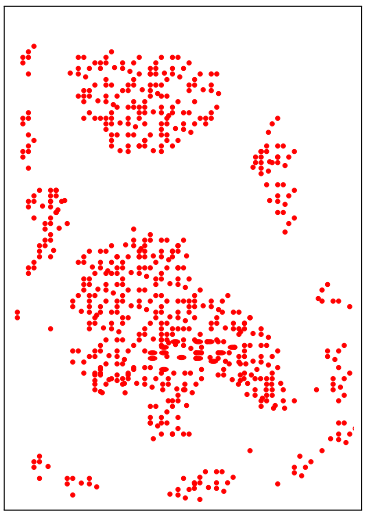


Figure 2: The complement $\widehat{\mathfrak{L}}$.

Alternatively, one can take start from the point $\mathfrak{l}_k \in \mathfrak{L}$: considering each point $\mathfrak{l}_k \in \mathfrak{L}$ as a center, cover them with the balls of the radius r, and remove all the centers of the balls containing points from $\mathfrak{M}$. These two procedure yield the same set $\mathfrak{L}^*$ of the points to be removed, due to the symmetry of the relation *to be inside a ball*. The symmetry results from the equivalence of radii r, both for centers located in $\mathfrak{m}_j \in \mathfrak{M}$, and centers located in $\mathfrak{l}_k \in \mathfrak{L}$.

The second option is to develop a mean-field function to make a decision towards the elimination of a point. To do that, supply each point $\mathfrak{m}_j \in \mathfrak{M}$ with a bell-shaped function $f(r)$

$$f(r) = f_{\mathfrak{m}_j}(r)\,, \qquad r = \sqrt{\sum_{i=1}^{n} \left(x_i^j - x_i\right)^2}\,, \tag{1}$$

where x_i^j is the i-th coordinate of the point $\mathfrak{m}_j$, but x_i is the i-th coordinate of a point $\mathbf{x}$. Then one must make a sum of all the functions (1)

$$\mathcal{F}(x_1, x_2, x_3, \ldots, x_{n-1}, x_n) = \sum_{\mathfrak{m}_j \in \mathfrak{M}} f_{\mathfrak{m}_j}(r) \tag{2}$$

to get an averaged "potential field". Function $f(r)$ may be chosen in a number of ways; they must integrable (probably, with its square) in the entire space. Monotonicity is another important constraint consists for these functions. Practically, one might want to implement the following functions, in (1):

1. Gaussian function $f(r) = \exp\{-\frac{r^2}{\beta^2}\}$; this is the classical function to develop a mean-field like approximation, and the motivation to use it comes from probability theory (the law of large numbers). Here parameter β is adjustable one, changing the typical width of a bell surrounding a center.
2. Exponential function $f(r) = \exp\{-\frac{r}{\beta}\}$ with β having the same meaning.
3. Resonance curve function $f(r) = \dfrac{1}{\beta^2 r^2 + 1}$ ($f(r) = \dfrac{1}{\beta^{n-1} r^{n-1} + 1}$, for n-dimensional case) with β having the same meaning.

Of course, there could be other functions meeting the constraints mentioned above.

As soon, as the function (2) is developed, one should choose the cut-off (or glue-off) level γ, and finally remove from $\mathfrak{L}$ all the point $\mathfrak{l}^* \in \mathfrak{L}$ so that

$$\mathcal{F}(x_1^{\mathfrak{l}^*}, x_2^{\mathfrak{l}^*}, x_3^{\mathfrak{l}^*}, \ldots, x_{n-1}^{\mathfrak{l}^*}, x_n^{\mathfrak{l}^*}) > \gamma\,, \tag{3}$$

thus yielding the complement $\widehat{\mathfrak{L}}$. Surely, there are two adjusting parameters, β and γ both affecting the implementation of the complement $\widehat{\mathfrak{L}}$.

Unlike for the case of the local determination of an eliminated point, the "mean-field" approach is not symmetrical, whether you develop a function (2) for the set $\mathfrak{M}$, for the set $\mathfrak{L}$. Moreover, two competing functions, $\mathcal{F}_{\mathfrak{m}\in\mathfrak{M}}(x_1, x_2, x_3, \ldots, x_{n-1}, x_n)$ and $\mathcal{F}_{\mathfrak{l}\in\mathfrak{L}}(x_1, x_2, x_3, \ldots, x_{n-1}, x_n)$ may explicate the effective mean field to get the complement. The second method to exclude the points $\mathfrak{l} \in \mathfrak{L}$ seems to be less hard, in computation, since an examination of a value of n-dimensional function costs less in comparison to the examination of the embedment of a point into a number of n-dimensional balls.

1.2 Fuzzy border and other parameters of $\mathfrak{L}$

We started from the case where the set $\mathfrak{M}$ could be almost unambiguously identified (as embedded into Ω), so that no problem takes place with the definition of the set $\mathfrak{L}$. A configuration of $\mathfrak{M}$ meeting this supposition might be met in a number of situations, nonetheless, there could be alternative patterns with fuzzy "border"; here we quote this word, since no one has clear, concise, self-consistent and productive definition of the border, for discrete sets.

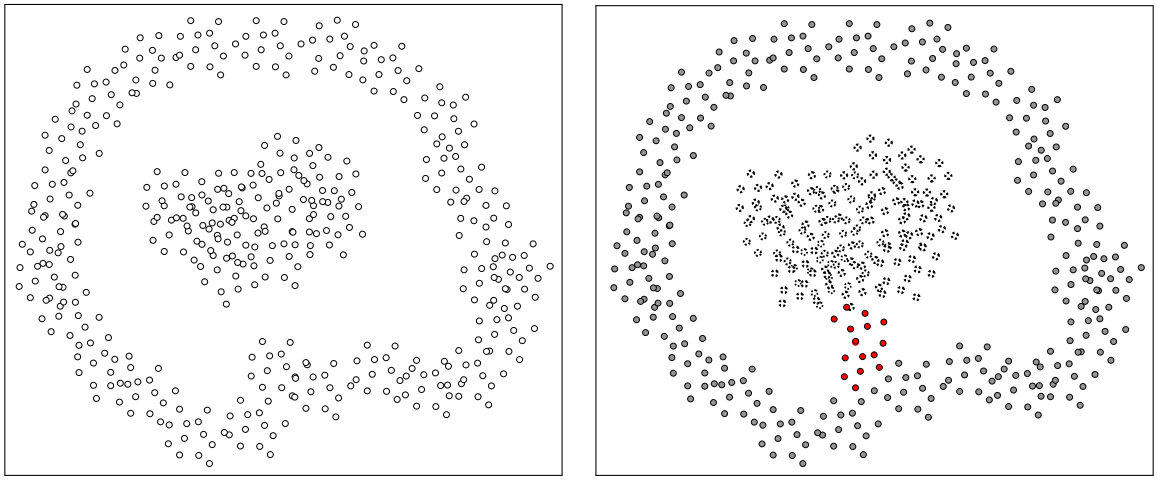
Less evident is the situation, if $\mathfrak{M}$ looks like a gradually dispersal set, as one goes outside from the center of that former. Thus, some difficulties in determination of Ω might take place. Here few options could be implemented, to overcome the problem. Firstly, one can follow a standard technique of image filtration [8]. A glance at Fig. 2 allows to see that the effect of contouring is present even for rather compact sets $\mathfrak{M}$, so the filtration should be applied at any chance.

Another option is to develop the area Ω artificially, say, building up the body Ω due to an ellipse of scattering implementation: counting all the eigenvectors of the covariance matrix for the set $\mathfrak{M}$, one can then build up the corresponding (n-dimensional) ellipsoid. By scaling that latter, one can fit the best subset $\mathfrak{M}^* \subset \mathfrak{M}$ completely falling inside the ellipsoid; thus, that latter might represent Ω.

As soon, as the complement is developed, one may treat it with standard and custom techniques, to find out, say, its cluster structure. As one can see from Figs. 1 and 2, the approximating manifold, at this example, is to be a part of plane with two holes inside. In other words, this is must be a manifold of genius two type. The occurrence of two holes could easily be detected with K-means technique applied to the complement $\widehat{\mathfrak{L}}$.

2 Straps in the datasets

The problems arises from the method itself: there might be two (or more) clusters that are apparently identified by a researcher, while the connectivity determined for giver ball radius r or glue level γ remains intact, so that the method fails to dissociate the dataset into clusters. An example of such data configuration is shown in Fig. 3(b). Indeed, the dataset shown in Fig. 3(a) seems to consist of two cluster; a strap consisting of the points labeled in red in Fig. 3(b) joins two clusters in one.



(a) Disconnected set.

(b) A set with a strap.

Figure 3: An example of dataset configuration consisting of two clusters, while the connectivity remains intact, for given radius r, or γ level.

There are few ways to address the problem. A simple answer is that the method does not identify two clusters, so the dataset has no cluster structure. On the other hand, an example shown in Fig. **??** makes an existence of two clusters (these are a "ring" and a "ball") rather obvious. An elimination of few points (shown in red, in Fig. 3(b)) changes a single-piece pattern for two clusters. So, the point is how to find out such a strap in a dataset.

Suppose, one has treated a dataset $\mathfrak{M}$ in order to reveal the complement $\widehat{\mathfrak{L}}$ and found that the structure of $\mathfrak{M}$ look pretty close to that one shown in Fig. 3(b). In other words, a set $\mathfrak{M}$ is suspected to consist of two (at least) clusters connected with a strap (or few straps, maybe). Hence, the strap must be detected in order to get the clusters clearly. To do that, one should develop special weighted graph, and

an interplay between metric and topological properties of that latter addresses the problem.

First, build up a complete graph $G(V, E)$ over the set $\mathfrak{M}$ with $\mathfrak{M}$ being the set of vertices V, and all pairwise distances

$$\rho\left(\mathfrak{m}_j, \mathfrak{m}_k\right) = \sqrt{\sum_{i=1}^{n} \left(x_i^j - x_i^k\right)^2} \tag{4}$$

being the edges; the distance (4) itself is the weight of an edge. Let then fix some $\varepsilon > 0$, and develop ε-connected component(s) of the graph $G(V, E)$. ε-connected component is defined as follows:

Definition 1. *Starting from a vertex $\mathfrak{m}_0$, let find all other vertices $\{\mathfrak{m}_s\}$ with the edges incident to $\mathfrak{m}_0$ and having the weight $\leqslant \varepsilon$. Going on, one builds up the* ***ε-connected component****.*

Here we implement it to search for straps. Definitely, whether the graph $G(V, E)$ is connected one, depends on ε figure: choosing the figure, one may get entire spectrum of the components, ranging from a set of M null-graphs to a single connected graph. Let then fix the minimal ε^* yielding a connected graph $G^*(V, E)$ (so that there exists a single connected component). This graph gathers all the points of $\mathfrak{M}$ (as vertices), but they are connected with the short segments not exceeding ε^* in length.

As soon as the graph $V^*(V, E)$ is developed, one can start up to search for a strap but before we need to define what is strap. Intuitively, a strap is rather small subset of data points that may not be excluded by a clustering method, while the elimination results in clear and unambiguous cluster implementation. More exactly,

Definition 2. *suppose a set $\mathfrak{M}$ consists of two explicitly identified subsets $\mathfrak{M}_1$ and $\mathfrak{M}_2$ (stipulated to be clusters) and $\mathfrak{S}$, so that $\mathfrak{M} = \mathfrak{M}_1 \sqcup \mathfrak{M}_2 \sqcup \mathfrak{S}$. Then $\mathfrak{S}$ is called* ***strap****, if*

a) $|\mathfrak{M}_1| \sim |\mathfrak{M}_2| \gg |\mathfrak{S}|$;

b) $\mathfrak{S}$ could be approximated by a manifold of lower dimension, in comparison to those modelling $\mathfrak{M}_1$ and $\mathfrak{M}_2$, and

c) elimination of $\mathfrak{S}$ from $\mathfrak{M}$ splits that latter into to unambiguous clusters.

Obviously, $\mathfrak{M}$ might have several straps; the definition could easily be generated for this case. Strap is a generalization of a bridge, in graph theory. Let now get back to a strap search.

To begin with, find the diameter d_{V^*} of $V^*(V, E)$; that latter itself yet does not answer the question on the identification of the strap. Meanwhile, the diameter contains the vertices from a strap, for sure. Key idea in strap search is to identify a subset of vertices belonging to a unexpectedly great number of (sufficiently long) paths. To find out those vertices, one has to check a number of sufficiently long paths to be found in $V^*(V, E)$. So, the question is how to develop these paths.

The following procedure provides the necessary set of pathes. First of all, select randomly a subset $\widetilde{\mathfrak{F}} \subset \mathfrak{M}$ of vertices, in $V^*(V, E)$, so that $\mathfrak{f} = |\widetilde{\mathfrak{F}}| = \sqrt{|\mathfrak{M}|}$. Next, all

$\mathfrak{f}^2 = |\mathfrak{M}|$ paths connecting the selected vertices must be build up. Obviously, some of them would be very short, other would as long as d_{V^*}.

Select then longer $V^*(V, E)/2$, and make a list of vertices for each path. Compare the lists and identify the vertices occurring in the great majority of the paths: these are the candidates to comprise a strap. A situation with several straps is a bit more complex, and requires a series of examinations of the compositions of frequently met vertices comprising straps.

Obviously, a dataset might be of severely complex structure; that one shown in Fig. **??** could be approximated with a manifold of rather simple topological character, i. e. a union of two manifolds of genus 0 and genus 1, respectively, or a manifold of the genus 1 (at the right of the figure). The point is that there could be more than one strap in the dataset. Growth of the genus character is not a sole problem in multi-dimensional data analysis; there might take place various loopings, etc. Nonetheless, the proposed method is able to treat such complicatedly organized datasets.

Acknowledgements

This work is partially supported by the grant of the Government of Russian Federation, grant № 14.Y26.31.0004

References

[1] Leskovec J., Rajaraman A., Ullman J. D. Mining of massive datasets. (2014) — Cambridge Univ. Press, 495 p.

[2] Fahad A., Alshatri N., Tari Z., Alamri A., Khalil I., Zomaya A. Y., Foufou S., Bouras A. (2014) A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Trans. on emerging topics in computing, **2**(3): 267 – 279.

[3] Dongkuan Xu, Yingjie Tian (2015) A Comprehensive Survey of Clustering Algorithms. Ann. Data. Sci., **2**(2): 165 – 193.

[4] Comaniciu D., Meer P. (2002) Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Machine Intell., **24**:603 – 619.

[5] Mirkes E. M., Alexandrakis I., Slater K., Tuli R., Gorban A. N. (2014) Computational diagnosis and risk evaluation for canine lymphoma. Computers in Biology and Medicine, **53**:279 – 290.

[6] Girvan M., Newman M. E. J. (2002) Community structure in social and biological networks. PNAS, **99**(12): 7821 – 7826.

[7] Akinduko A. A. Mirkes E. M., Gorban A. N. SOM: Stochastic initialization versus principal components. (2016) Information Sciences, 364-365: 213 – 221.

[8] Lecarme O., Delvare K. (2013) The Book of GIMP: A Complete Guide to Nearly Everything. No Starch Press, 676 pp.

Features of Testing Goodness-of-Fit by Big Data

Stanislav S. Vozhov, Maria A. Semenova and Ekaterina V. Chimitova
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: chimitova@corp.nstu.ru

Abstract

The problems of application of the χ^2 Pearson, Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling goodness-of-fit tests by big data have been considered. The difference of the real test statistic distributions from the corresponding limiting distributions have been shown in the case of testing simple hypotheses by big data, which were obtained with limited accuracy. The power of the χ^2-test has been investigated for various values of data registration accuracy.

Keywords: big data, goodness-of-fit, χ^2 Pearson test, Kolmogorov test, Cramer-von Mises-Smirnov test, Anderson-Darling test.

Introduction

The volumes of observations recorded during system monitoring, in engineering experiments, in mass production, in economical, social, biological or medical research increase every day. This fact makes it necessary to test various statistical hypotheses by big data (goodness-of-fit hypotheses, homogeneity hypotheses, hypothesis of the homogeneity of numerical characteristic and others).

The attempts to apply classical tests (χ^2 test, Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling, Cooper, Watson tests, etc.) for big data usually lead to failure: to rejection of the hypothesis tested, even when it is true. This fact can be explained by the following: the volumes of accumulated data are very large, but the indicators studied are registered with limited accuracy (the observations values are usually obtained with some limited number of digits after the point). The classical tests are intended for samples of continuous random variables. Large volume of measurements obtained with low accuracy results in a set of identical values in big data, which indicates the violation of continuity assumption of observed random variables.

In the case of big data analysis, it is possible either to apply special tests developed and oriented to big data, or to adapt classical statistical tests that have proven in the analysis of samples of identically and continuously distributed random variables. The creation of special tests involves the need to verify these tests and the difficulty of further implementing new methods created in existing software systems with an established structure, which can cause difficulties.

The modification of existing and implemented tests for big data looks to be more promising. However, the possibility of using asymptotic results for the tests requires the limitation of the sizes of extracted and analyzed samples from data. One should consider the influence of the estimation method (and grouping) on the distribution of statistics when testing composite hypotheses with estimation of the distribution parameters by such "digit-grouped" data.

1 Goodness-of-fit tests

Let there is a sample of independent identically distributed random variables $\mathbf{X}_n = \{X_1, X_2, ..., X_n\}$ from the distribution $F(x)$. In this paper, we consider the problem of testing simple goodness-of-fit hypotheses H_0: $F(x) = F_0(x)$ and composite hypotheses, which can be presented as H_0: $F(x) \in \{F_0(x;\theta), \theta \in \theta\}$.

One approach for testing goodness-of-fit hypotheses by complete samples is the application of the nonparametric tests: the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests. The Kolmogorov test statistic is given by

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

where $\hat{F}_n(x)$ is the empirical distribution function. In practice, the statistic is usually used with the Bolshev correction [1] of the form

$$S_K = \frac{6n \cdot D_n + 1}{6\sqrt{n}}. \tag{1}$$

The Cramer-von Mises-Smirnov test statistic can be written by

$$S_\omega = n \int_{-\infty}^{\infty} \left(\hat{F}_n(x) - F_0(x;\theta)\right)^2 dF_0(x;\theta). \tag{2}$$

The Anderson-Darling test statistic can be presented as

$$S_\Omega = n \int_{-\infty}^{\infty} \left(\hat{F}_n(x) - F_0(x;\theta)\right)^2 \frac{dF_0(x;\theta)}{F_0(x;\theta)(1 - F_0(x;\theta))}. \tag{3}$$

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 the considered statistics do not depend on the tested distribution. Statistic S_K belongs to the Kolmogorov distribution, S_ω and S_Ω belong to the $a1$ and the $a2$ distributions, respectively [1]. For composite hypotheses the nonparametric test statistic distributions $G(s|H_0)$ are affected by a number of factors: the form of the tested lifetime distribution $F_0(x;\theta)$, 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 have been proposed in [2], [3] and [4].

2 Testing goodness-of-fit by big data

In a variety of applications of statistical analysis, data can be obtained with some accuracy. For example, the common ruler has the measurement accuracy Δ of one millimeter. As a result of the development of information technologies, the number of

observations in a sample can be very large, while the number of unique sample values can be very small. Under such conditions, the above criteria for testing goodness-of-fit hypothesis can not cope with the task. On Figure 1, you can see the empirical distributions of the Kolmogorov test statistic (1) for testing goodness-of-fit of the standard normal distribution by samples of observations obtained with one decimal place accuracy ($\Delta = 0.1$) and the limiting distribution $K(s)$. We simulated $N = 16600$ samples of size $n = 100, 500, 1500, 3000$.

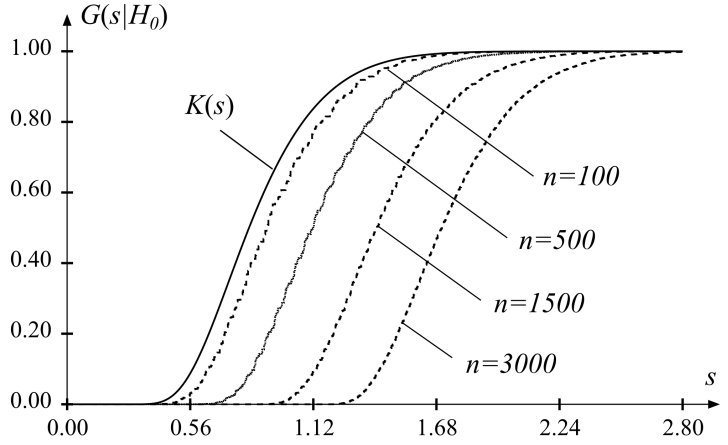


Figure 1: The distributions of the Kolmogorov test statistic, $\Delta = 0.1$

As can be seen from Figure 1, the difference between empirical distributions of the Kolmogorov test statistic and the corresponding limiting distribution $K(s)$ decreases as the sample size grows. Moreover, the difference described is quite large. This deviation of distributions to the right leads to rejection of null hypothesis even if it is true. Similar results were observed for Cramer-von Mises-Smirnov and Anderson-Darling tests.

In fact, the data obtained with some accuracy Δ are the grouped data with the intervals of length Δ. The number of intervals k is defined by the number of unique values in the sample $\mathbf{X}_n$. Let us denote the unique values in the increasing order as $x_1, x_2, ..., x_k$. Then, the boundary points of grouping intervals

$$-\infty = a_0 < a_1 < ... < a_{k-1} < a_k = +\infty$$

are defined as follows:

$$a_i = x_i + \frac{\Delta}{2}, i = \overline{1, k-1}.$$

To test the goodness-of-fit hypothesis by grouped data, the χ^2 Pearson test is usually used. In accordance to the given partition, the number n_i of sample values fallen into the i-th interval is counted, and the probability of falling into the interval $P_i(\theta) = \int\limits_{a_{i-1}}^{a_i} f_0(x;\theta)dx$ corresponding to the theoretical law under H_0 is calculated, $\sum\limits_{i=1}^{k} n_i = n$, $\sum\limits_{i=1}^{k} P_i(\theta) = 1$.

The statistic of χ^2 Pearson test is calculated in accordance to the statement:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - nP_i(\theta))^2}{nP_i(\theta)}. \tag{4}$$

In the case of testing a simple hypothesis in the limit $n \to \infty$, the statistic χ^2 has the χ^2_r-distribution with $r = k - 1$ degrees of freedom if the null hypothesis is true. Similarly, for a composite hypothesis, the statistic χ^2 has the χ^2_r-distribution with $r = k - m - 1$ degrees of freedom, where m is the number of parameters estimated. There are various ways to estimate unknown parameters basing on such samples. Very often, the maximum likelihood method is used. The estimates of unknown parameters are obtained by solving the following optimization problem:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ln L\left(\mathbf{X}_n; \theta\right) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln f_0(X_i; \theta). \tag{5}$$

However, it is well known that the statistic (4) has the χ^2_r-distribution with $r = k - m - 1$ degrees of freedom only if unknown parameters are estimated by the grouped data. So, if unknown parameters are estimates according to (5), then the Pearson's statistic does not belong to the χ^2 distribution, as can be seen from Figure 2. There are the distributions of statistic (4) obtained in the case of testing composite hypothesis of goodness-of-fit of normal distribution by samples of observations, taken with rounding off to a whole ($\Delta = 1$), $k = 7$ grouping intervals were used. Unknown parameters were estimated by initial samples $\mathbf{X}_n$ according to (5).

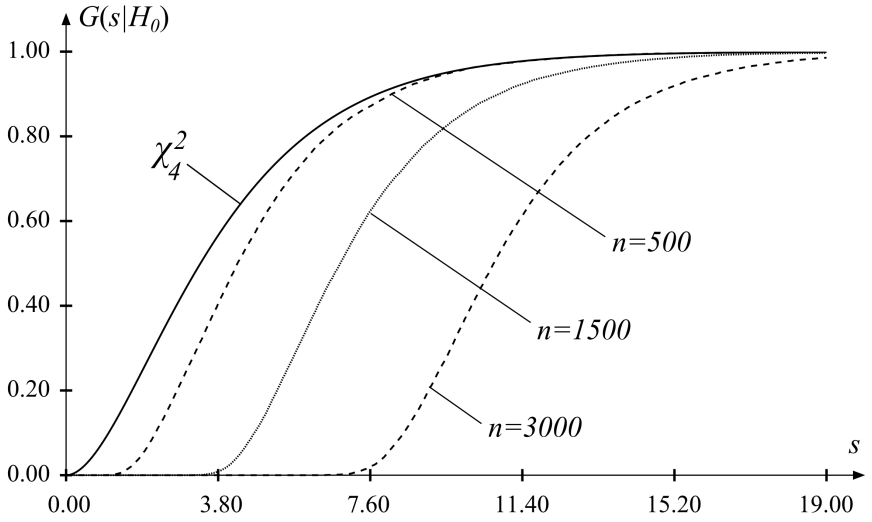


Figure 2: The distributions of the χ^2 test statistic, $\Delta = 1$

When testing composite hypotheses with the χ^2 goodness-of-fit test by data obtained with limited accuracy, unknown parameters of the distribution $F_0\left(x; \theta\right)$ must

be estimated by maximizing the logarithm of the likelihood function:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \ln L\left(\mathbf{X}_n;\theta\right) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{k} n_i \ln\left(F(a_i;\theta) - F(a_{i-1};\theta)\right). \tag{6}$$

In this case, the statistic 4 has the χ^2_r-distribution with $r = k - m - 1$ degrees of freedom.

3 An empirical analysis of the power of χ^2 test

To investigate the power of the χ^2 goodness-of-fit test for different values of Δ and n, we consider the pair of close competing hypotheses:

H_0 : the standard normal distribution against

H_1 : the logistic distribution with the density function

$$f(x;\theta) = e^{-\frac{(x-\theta_1)}{\theta_2}} \Big/ \theta_2 \left(1 + e^{-\frac{(x-\theta_1)}{\theta_2}}\right)^2,$$

parameters $\theta_1 = 0, \theta_2 = \frac{\sqrt{3}}{\pi}$. The estimates of the χ^2 test power were calculated for the significance level $\alpha = 0.05$. We simulated $N = 16600$ samples of observations of size $n = 500, 1500, 3000$.

Let us compare the power estimates for different values of Δ, when testing simple and composite hypotheses.

The estimates of power calculated for $\Delta = 1$ are given in Table 1. In this case, $k = 7$ intervals were used:

$$a_1 = -2.5, a_2 = -1.5, a_3 = -0.5, a_4 = 0.5, a_5 = 1.5, a_6 = 2.5.$$

In the case of composite hypothesis, unknown parameters of normal distribution were estimated according to (6).

Table 1: Power estimates of χ^2_r test, $\Delta = 1$

n	Simple hypothesis	Composite hypothesis
500	0.54	0.62
1500	0.96	0.97
3000	0.99	0.99

If Δ is rather small, that is the measurements are obtained with high accuracy, then the number of grouping intervals k will be large. In [5], it was shown that when testing close competing hypotheses, the power of the χ^2 test decreases with the number of intervals growth. Therefore, it is rational to use a smaller number of intervals $k' < k$. However, there is a question: how to choose new boundary points? The problem is that we have to choose new boundary points only among the values

$a_i = x_i + \frac{\Delta}{2}, i = \overline{1, k-1}$. In this paper, we consider the approximate equiprobable grouping, according to which new boundary points are calculated as follows:

$$a'_i = \left\langle F_0^{-1}\left(\frac{i}{k'}\right)\right\rangle + \frac{\Delta}{2}, i = \overline{1, k'-1},$$

where $\langle\cdot\rangle$ is rounding off to a value with accuracy Δ. For example, if $\Delta = 0.1$, then $\langle 2.74\rangle = 2.7$, $\langle 1.28\rangle = 1.3$.

The estimates of power calculated for $\Delta = 0.1$ and various k' are given in Tables 2 and 3. In the last columns of Tables 2-5, there are the power estimates for the number of grouping intervals k, corresponding to the number of unique values of observations.

Table 2: Power estimates of χ^2 test in the case of simple hypothesis, $\Delta = 0.1$

n	$k' = 2$	$k' = 3$	$k' = 4$	$k' = 5$	$k' = 7$	$k = 61$
500	0.05	0.33	0.37	0.35	0.29	0.38
1500	0.05	0.78	0.85	0.85	0.79	0.87
3000	0.06	0.97	0.99	0.99	0.98	0.99

Table 3: Power estimates of χ^2 test in the case of composite hypothesis, $\Delta = 0.1$

n	$k' = 4$	$k' = 7$	$k' = 11$	$k' = 13$	$k' = 15$	$k' = 20$	$k = 61$
500	0.05	0.09	0.11	0.13	0.16	0.15	0.48
1500	0.05	0.14	0.33	0.35	0.41	0.42	0.92
3000	0.05	0.33	0.66	0.80	0.72	0.84	0.99

From Tables 2 and 3 you can see that using native grouping ($k = 61$) results in a higher power than when using smaller number of intervals. For approximately equiprobable grouping, in the case of simple hypothesis, the optimal number of grouping intervals is $k' = 4$.

The estimates of power calculated for $\Delta = 0.01$ and various k' are given in Tables 4 and 5.

Table 4: Power estimates of χ^2 test in the case of simple hypothesis, $\Delta = 0.01$

n	$k' = 3$	$k' = 4$	$k' = 6$	$k' = 8$	$k' = 10$	$k = 602$
500	0.36	0.37	0.31	0.29	0.27	0.11
1500	0.81	0.85	0.82	0.78	0.74	0.28
3000	0.98	0.99	0.99	0.99	0.98	0.64

As was seen from Tables 2 and 3, when $k = 61$, there is no need to use the smaller number of intervals k' as the test power decreases. However, in the case of $k = 602$ ($\Delta = 0.01$) when testing simple hypothesis, the power of the χ^2 test is much higher

Table 5: Power estimates of χ^2 test in the case of composite hypothesis, $\Delta = 0.01$

n	$k' = 4$	$k' = 6$	$k' = 8$	$k' = 10$	$k = 602$
500	0.05	0.07	0.09	0.11	0.21
1500	0.05	0.13	0.23	0.28	0.45
3000	0.05	0.24	0.43	0.62	0.82

for small k', so it is necessary to use the smaller number of intervals, choosing new boundary point, for example, with approximately equiprobable grouping method.

Concluding remarks

The results of investigation presented in the paper allow us to state some problems in application of the considered goodness-of-fit tests by samples of observations, which were obtained with limited accuracy.

The use of tables of percentage points or corresponding limiting distributions for the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling goodness-of-fit tests can lead to rejection of null hypothesis even if it is true. The reason is that the real statistic distributions in the case of big data are significantly different from the corresponding limiting distributions. So, the application of these popular tests for big data obtained with limited accuracy can lead to incorrect results. In future, it would be interesting to develop the algorithm of p-value estimation that takes into account the deviation of statistic distributions from the corresponding limiting distributions for these tests.

Another approach to testing goodness-of-fit by big data is to use the χ^2 Pearson test. In this case, it is important to remember, that the boundary points must be chosen only among the values $a_i, i = \overline{1,k}$, defined in Section 2 of this paper. The investigation of the power of χ^2 test has shown that if measurements are obtained with high accuracy (the number of unique values in a sample is rather large), then it is rational to use the approximately equiprobable grouping method.

Acknowledgements

This research has been supported by the Ministry of Education and Science of the Russian Federation (project 1.1009.2017/4.6)

References

[1] Bolshev, L.N. and Smirnov, N.V. (1983). *Tables of Mathematical Statistics*. Moscow: Science, (in Russian).

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

[3] Lemeshko, B.Yu. and 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**, 799-812.

[4] Lemeshko, B.Yu., Lemeshko, S.B., Nikulin, M.S. and 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**, 1358–1373.

[5] Lemeshko, B.Yu., Chimitova E.V. (2003). On the choice of the number of intervals in χ^2-type goodness-of-fit tests. *Zavod. Lab. Diagn. Mater.*, **69**, 61–67.

The Online Marketplace Selection for Searching and Placing Advertisements

Vladimir Timofeev and Olga Kravchenko
Novosibirsk State Technical University, Novosibirsk, Russian Federation
e-mail: v.timofeev@corp.nstu.ru, o.p.kravchenko.2012@stud.nstu.ru

Abstract

This paper addresses the issue of selecting the online marketplaces. The selection was carried out between 17 sites on 9 criteria. The online marketplace selection algorithm based on the analytic heirarchy process is implemented in Visual C Sharp as the result of this work. Consistency increasing methods are examined and added to the designed algorithm.

Keywords: online marketplace, selection algorithm, analytic hierarchy process, consistency ratio.

Introduction

The Internet allows people to perform a significant part of the activity at the monitor screen. There is no longer any need to go shopping investigating shelves in search of the right product; to wander the streets examining each notice board to find a house for sale; to ask firms' employees about available vacancies; all of this can now be done sitting at home at the computer, only by selecting an online marketplace and viewing ads on the topic of interest.

At such a moment the question that arises is which Internet site to choose in order to find the necessary advertising information quickly. A similar question can also arise in enterprises in case they need to place any advertising information. But how to do this, if the Internet is full of all kinds of ad sites. How to choose an advertising platform with a sufficient number of ads and visitors. After all, one can place an advertisement on ten online marketplaces, without getting the desired result, or view a dozen sites, without finding the right one. It is for this purpose that you need to rank online marketplaces according to certain criteria corresponding to the reason for the site search. Due to the lack of information about their comparison in its pure form, both the data of web analytics services and the experts' opinions are used to rank the online marketplaces. But the use of the method of expert evaluations is complicated by the fact that the obtained judgments cannot always be consistent, which can lead to incorrect results. Therefore, in this article, the problem of increasing the consistency of judgments is considered. For this, the existing methods of increasing the consistency ratio were investigated and the online marketplace selection algorithm was proposed.

1 Online marketplace selection criteria and their significance

The evaluation of any site is based on certain criteria, according to which its preferences are determined. One of such criteria is number of visitors per unit time (Cr4). The unit of time is most often a day or a month. Information on this criterion is provided both by the Internet sites themselves, and by the analytics services. World rank (Cr1) and Russia rank(Kr2) reflect the relevance of the site. Category rank (Cr3) is a rarer criterion, monitored by statistical services. There are different categories: art, business and industry, education, games, news and media. Internet sites belong to the category of purchases.

Also it is necessary to consider behavioral factors - a set of visitors' actions on a certain site. These indicators include: daily time on site (Cr5), daily pageviews (Cr6) and bounce rate (Cr7). Sites with high behavioral indicators tend to win competition.

In addition, one of the most unique criteria for an online marketplace is number of advertisements (for Novosibirsk (Cr8) and Russia (Cr9)).

Online-marketplace criteria are conveniently presented in the form of a tree (see fig.1), reflecting the structure of the constructed hierarchy.

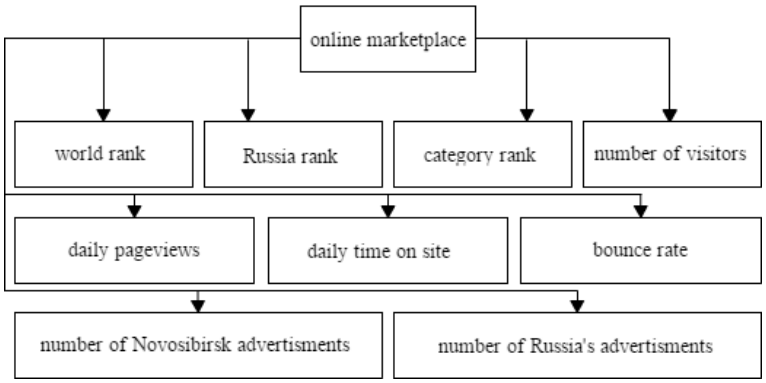


Figure 1: Criteria tree

It should be noted that, on the one hand, the criteria in fig.1 differ in units of measure, for example, daily time on site is measured in seconds, and bounce rate in percentage; and on the other hand they are multidirectional: for world rank, the minimum value is the best, and for the number of visitors - the maximum one.

Moving on to the problem of evaluating the significance of criteria. Due to a significant difference in the criteria, a pairwise comparison method based on the Saati scale was chosen for their comparison.

Since the methodology of Saati involves the construction of pairwise comparison matrices [1], for the criteria illustrated in fig.1 different matrices were constructed by the experts, one of them is presented in tab.1. To estimate their consistency, the consistency ratio was calculated as follows:

$$CR = \frac{CI}{RI} \tag{1}$$

where CI is the consistency index, RI is the random index calculated as

$$CI = \frac{\lambda - n}{n-1}, RI = \frac{1.98 * (n-2)}{n} \tag{2}$$

where λ is the maximum eigenvalue of a pairwise comparison matrix, n is the number of criteria.

It reflects the proportionality of experts' judgments, the higher the value, the worse the consistency, which indicates their controversy.

Table 1: The Inconsistent Pairwise Comparison Matrix

	$Cr1$	$Cr2$	$Cr3$	$Cr8$	$Cr9$
Cr1	1	0.13	0.25	0.17	0.13
Cr2	8	1	0.13	0.25	0.17
Cr3	4	8	1	0.50	0.25
Cr8	6	4	2	1	0.50
Cr9	8	6	4	2	1

For the matrix shown in tab.1, CR = 0.1866, which indicates its inconsistency. The use of inconsistent matrices for ranking is undesirable, since this can lead to incorrect results, thus, methods of increasing consistency are examined.

2 Study of the consistency of judgments improving methods

To improve consistency ratio, methods using an already constructed inconsistent or slightly inconsistent matrix can be used. Xu and Wei method is one of them. This method is to change the original matrix, taking into account the whole set of expert judgments, until a predetermined consistency ratio is achieved [2].

The Cao method also changes all elements until the desired consistency ratio is achieved. In this method, a deviation matrix is constructed using the γ variable. The closer the value of this variable to 1, the closer the resulting matrix to the original one [3].

While the previous methods change all elements of the matrix, the Ergu, Kou, Peng, and Shi method changes only one element at one iteration, which allows one to save most of the expert's initial judgments. This method is to identify the inconsistent matrix element and to change its and the corresponding element's values [4].

Before using the described methods, studies were conducted on laboriousness, increasing consistency and preserving the initial information. To study them, matrices

from 3x3 to 9x9 were chosen, since a 2x2 pairwise comparison matrix is always absolutely consistent. The tabl.2 illustrates an investigation of the number of iterations, the number of changed elements, and the percent change of original comparison information with an increase of the number of alternatives. The 2nd, 5th and 8th columns show, that the number of iterations in the Ergu, Kou, Peng and Shi method grows from one to three, while the Xu and Wei and Cao methods improve the consistency ratio in just one iteration. But these methods change almost the entire matrix, except for single elements, as seen from the 3rd and 9th columns, while the Ergu method changes the maximum of three elements. Despite the fact that only a few elements have changed in the Ergu method, the obtained matrices differ from the original ones (by almost 31 percent). While the completely modified matrix by the Cao method differs from the original by only 6 percent.

Table 2: Investigations of the number of iterations, changed elements and preservation of the original comparison information

	Xu and Wei			Ergu, Kou, Peng and Shi			Cao		
1	2	3	4	5	6	7	8	9	10
3	1	3	12.28	1	1	40.73	1	6	5.23
4	1	6	16.34	1	1	40.73	1	6	5.23
5	1	10	22.81	2	1	36.90	1	10	5.38
6	1	15	19.27	2	2	20.83	1	15	5.88
7	1	25	20.48	3	3	37.11	1	25	6.13
8	1	25	22.72	3	3	28.41	1	25	6.88
9	1	33	22.26	3	3	22.45	1	33	6.52

CR can also be improved in a fundamentally different way - by decomposition and aggregation. The idea of a hierarchy can be used by dividing the entire set of data into clusters according to their relative importance [5]. This method is to compare clusters, their alternatives and to multiply the corresponding priority vectors. An example of this method applied to the online marketplace selection problem is shown in fig.2.

The consistency ratios study for all four methods are presented in tabl.3. In this table CR (2nd, 4th, 6th and 8th columns) and the percentage of CR increase (3rd, 5th, 7th and 9th columns) are shown. In the 2nd and 4th columns one can see that the method of constructing clusters gives the most consistent matrices, the method of Xu and Wei leads to almost the same value of the consistency ratio for any matrix. The Ergu, Kou, Peng, and Shi method for 3x3 matrix is the only one that gives an absolutely consistent matrix, but in general, the percentage of CR increase is lower than the remaining methods, except for the Cao method, which improves consistency by approximately 30 percent, as seen in the 5th column. Low percantage of consistency ratio increase of the Cao menthod can be explained by the fact that in order to preserve the maximum possible amount of original comparison data, the

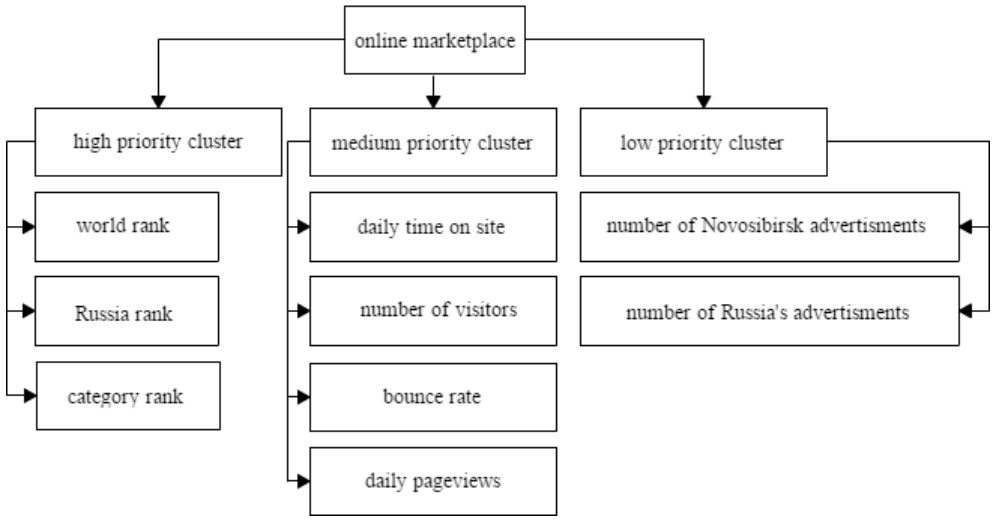


Figure 2: Cluster tree

matrix changes only until the desired CR is achieved.

Table 3: Study of the change in the consistency ratio

	Decomposition and Aggregation		Xu and Wei		Ergu, Kou, Peng and Shi		Cao	
1	2	3	4	5	6	7	8	9
3	-	-	0.0254	75.24	0.0701	31.82	0	100
4	0	100	0.0291	76.27	0.0854	30.27	0.0272	77.80
5	1.41E-4	99.89	0.0286	76.51	0.0882	27.51	0.0447	64.07
6	1.41E-4	99.88	0.0296	75.58	0.0866	28.49	0.0591	51.23
7	0.0014	98.84	0.0289	76.10	0.0859	28.92	0.0415	65.63
8	0.0010	99.13	0.0286	76.37	0.0873	27.84	0.0724	40.14
9	0.0010	99.16	0.0299	76.21	0.0903	28.11	0.0949	24.45

According to the results of the study, the Cao method was chosen as the preferred consistency increasing method. Therefore, this method was applied to the matrix, used in the online marketplace selection problem. The results of pairwise comparisons of 9 criteria modified by the Cao method and the initial ones (in parentheses), as well as the priority vector, are shown in tab.4.

For the modified matrix, the consistency ratio was estimated (1), CR = 0.0252, which is less than the original consistency ratio by 52 percent, furthermore 90 percent of the original information was retained. Although the values of the priority vector were changed, the order of the criteria remained the same.

Table 4: Modified online marketplace criteria matrix by the Cao method

Cr	Cr1	Cr2	Cr3	Cr4	Cr5	Cr6	Cr7	Cr8	Cr9	Weight
Cr1	1	0.41	0.22	0.10	0.15	0.15	0.15	0.09	0.11	0.0150
		(0.33)	(0.2)	(0.11)	(0.14)	(0.14)	(0.14)	(0.11)	(0.13)	(0.0150)
Cr2	2.44	1	0.29	0.11	0.19	0.19	0.19	0.11	0.14	0.0216
	(3)		(0.25)	(0.13)	(0.17)	(0.17)	(0.17)	(0.13)	(0.14)	(0.0216)
Cr3	4.56	3.57	1	0.30	0.40	0.40	0.40	0.30	0.44	0.0563
	(5)	(4)		(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.5)	(0.0569)
Cr4	10.51	8.73	3.29	1	3.49	3.49	3.49	0.97	1.09	0.2256
	(9)	(8)	(3)		(4)	(4)	(4)	(1)	(1)	(0.2251)
Cr5	6.57	5.30	2.47	0.29	1	1	1	0.28	0.38	0.0875
	(7)	(6)	(3)	(0.25)				(0.25)	(0.33)	(0.0874)
Cr6	6.57	5.30	2.47	0.29	1	1	1	0.28	0.38	0.0875
	(7)	(6)	(3)	(0.25)				(0.25)	(0.33)	(0.0874)
Cr7	6.57	5.30	2.47	0.29	1	1	1	0.28	0.38	0.0875
	(7)	(6)	(3)	(0.25)				(0.25)	(0.33)	(0.0874)
Cr8	10.71	8.93	3.36	1.03	3.56	3.56	3.56	1	1.78	0.2433
	(9)	(8)	(3)	(1)	(4)	(4)	(4)		(2)	(0.2433)
Cr9	9.03	7.37	2.29	0.92	2.64	2.64	2.64	0.56	1	0.1759
	(8)	(7)	(2)	(1)	(3)	(3)	(3)	(0.5)		(0.1756)

3 The automation of the hierarchy analysis process as a way of ranking online marketplaces

On the basis of the criteria comparison results, the ranking of 17 online marketplaces providing their services in Novosibirsk and across Russia, presented in tab.5, was performed according to the criteria described in the first section using the hierarchy analysis process.

It should be noted, that this process consists of constructing a hierarchical structure, determining the priorities of its elements by paired comparisons, and synthesizing priorities on the hierarchy, as a result of which the priorities of alternative solutions relative to the main goal are calculated [6]. The best alternative is an alternative with a maximum priority value.

The hierarchy analysis process provides a distributed and ideal way of ranking. For the distributed method, the eigenvector of the constructed pairwise comparison matrix is normalized by the sum of vector's elements, and for the ideal one – by a maximum.

Since all approaches used in this article are independent algorithms, the study of which requires time and skill, they were combined in a single software that automates all stages of ranking. For this, the online marketplace horizontal ranking algorithm for searching and placing advertisements was developed having the following form.

Step 1. Define both the set of online marketplaces and criteria

Step 2. Evaluate the significance of the criteria by constructing a pairwise criteria comparison matrix.

Step 3. Check the consistency of the constructed pair-wise comparison matrix and increase it if necessary by one of the methods described in the second section.

Step 4. Construct the Perron vector by a power method on the basis of the constructed pairwise criteria comparison matrix.

Step 5. Construct pairwise online marketplace comparison matrix for each of the selected criteria on the basis of obtained online marketplace data using the Saati fundamental scale.

Step 6. Construct two eigenvectors – normalized and idealized priority vectors for each pairwise online marketplace comparison matrix using the power method.

Step 7. Obtain global vectors by multiplying the local elements of online marketplaces priority and the corresponding criteria priorities.

Step 8. Arrange the elements of global vectors in descending order, having received a list of online marketplaces in the order of their preference.

This algorithm was implemented in C Sharp. The distributed and idealized ranking results of the 17 selected online marketplaces based on the pairwise criteria comparison matrix modified by the Cao method (see tab.4) are presented in tab.5.

The distributed and idealized results are slightly different, but in both cases the best online marketplaces are Avito, "Iz ruk v ruki" and "NGS ob"yavleniya", and the two worst ones are Adiso and Resurso.

Table 5: Results of the ranking algorithm

Online Marketplace	URL	Weight (D)	Weight(I)
Resurso	resurso.ru	0.0250	0.0881
Doska ob"yavleniy	do.ru	0.0676	0.2929
UBU	ubu.ru/novosibirsk	0.0350	0.1821
Barakhla.net	ns.barahla.net	0.0544	0.2486
Sindom	novosibirsk.sindom.ru	0.0418	0.2074
Avito	avito.ru/novosibirsk	0.2680	0.9444
Acoola	acoola.ru	0.0434	0.2136
Kupi proday	kupiprodai.ru	0.0498	0.2248
Doski.ru	novosibirsk.doski.ru	0.0362	0.1833
Iz ruk v ruki	novosibirsk.irr.ru	0.0826	0.3546
NGS ob"yavleniya	do.ngs.ru	0.0685	0.3084
Dorus	novosibirsk.dorus.ru	0.0437	0.2131
Kupi.ru	novosibirsk.qp.ru	0.0459	0.2334
1000dosok.ru	www.1000dosok.ru	0.0343	0.1730
Besplatnyye ob"yavleniya	besplatnyeobyavleniya.ru	0.0365	0.1684
Adiso	adiso.ru	0.0310	0.1503
Gde	gde.ru	0.0362	0.1915

Conclusion

17 online marketplaces providing services in Novosibirsk were selected for online marketplace ranking. The selection criteria and the method for assessing their significance were also presented.

Four methods of increasing the consistency of judgments were examined and their advantages and disadvantages were revealed. According to the results of the study, the Cao method was chosen as a consistency increasing method.

According to the results of the ranking of the 17 selected online marketplaces, Avito is the most preferred, "Iz ruk v ruki" and "NGS ob"yavleniya" are in the second and third places, thus, the goal of this paper has been achieved.

This paper, as well as the developed software, would be useful to users and enterprises wishing to quickly select the most effective online marketplaces for placing advertising information, and also to choose a method for increasing the consistency of expert judgments.

References

[1] Saaty T.L. (2015). On the measurement of intangibles. A principal eigenvector approach to relative measurement derived from paired comparisons. *Cloud of Science*. Vol. **2**, pp. 1-39.

[2] Xu Z., Wei C. (1999). A consistency improving method in the analytic hierarchy process. *European Journal of Operational Research*. Vol. **116**, pp. 443-449.

[3] Cao D., Leung L.C., Law J.S. (2008). Modifying inconsistent comparison matrix in analytic hierarchy process: A heuristic approach. *Decision Support Systems*. Vol. **44**, pp. 944-953.

[4] Ergu D., Kou J., Peng Y., Shi Y. (2011). A simple method to improve the consistency ratio of the pair-wise comparison matrix in ANP. *European Journal of Operational Research*. Vol. **213**, pp. 246-259.

[5] Saaty T.L. (1993). *Decision making with the analytic hierarchy process*. Radio i svyaz', Moscow.

An Argumentation based Statistical Support Tool

Yiannis Kiouvrekis [1], Petros Stefaneas[1] and Aggeliki Kokkinaki[2]
[1] *National Technical University of Athens,*
Department of Mathematics,
Herron Polytechniou 9, 15780 Zografou, Greece
[2] *Department of management and MIS*
University of Nicosia, Nicosia, Cyprus
e-mail: yiannisq@central.ntua.gr, petros@math.ntua.gr,
kokkinaki.a@unic.ac.cy

Abstract

This paper identifies the need for an information system that assists decision making for the selection of appropriate statistical methods. The target group of users of the proposed system includes professional and/or scientists who are involved in data and statistical analysis but do not necessarily possess thorough mathematical training that could help them avoid shortcomings in their selection of methods. The proposed system employees non-monotonic logic, Argumentation Logic and relies on Gorgias-B system to guide the user through a dynamically readjusted linear path of argumentation to the desired proper solution.

Keywords: AMSA, Gorgias, Logic, Non monotonic logic, statistical decision support tool

Introduction

In this paper we will present a new methodology in the field of decision making process for statical analysis and data analysis. Nowadays, more than ever before, there is an intense need for data analysis and application of statistical tests. This need is encountered across various disciplines; it is driven by the phenomenal development of applications relying on Big Data, Business Intelligence and Machine Learning algorithms, as well as requirements of basic and applied research in a wide variety of scientific fields.

A recurrent problem in such endeavours is the improper selection of statistical methodology; this is often attributed to the fact that characteristics of the data set under examination as well as the conditions of performing a statistical analysis influence what is considered to be proper statistical methodology [1]. Knowledge about such special conditions is often confined within the mathematical community, whereas it is not often employed by those performing statistical analysis in other scientific fields, such as applied sciences, engineering, social sciences, humanities, health sciences etc.

Professionals and researchers in this fields are misguided in their selection of statistical analysis methods and often introduce mistakes in their results due to the discussed inefficiency[3]. As a result significant implications may occur: scientific research results may be affected and/or important decisions in professional settings

may be subjected to incorrect results. The scope of the problem is wide, its implications significant and the required mathematical supervision simply unavailable.
This article introduces a decision-making tool for proper statistical control using argumentation logic as examined from the perspectives of Artificial Intelligence (AI)
The main challenge is to develop a software that will **behave like a scientist** so that

1. the information can be easily obtained and faithfully represented

2. changes of the specifications could be easily transmitted to the software

The final software would also be able to provide information on the reasons why the scientist should use a test suggested as appropriate.
The structure of the paper is the following: in section 1 the basic characteristics of the misuse and abuse of statistics in non-math research are presented, section 2 contains the basic notions of the field of non monotonic logic, the topic of argumentation logic and Gorgias software; in section 3 we present a case study for our tool.

1 Misuse and Abuse of statistics in research

In this section we describe some of the fundamental errors which may occur in a research. First of all errors in the statistical design may be observed. We know that each statistical test has certain assumptions such as type of data, whereas the phenomenon of using the wrong test occurs often among researchers. Second we observe errors which occurs in the description of the data.
In [2] we read that 90% of the published articles avoid to mention or discussed the appropriate assumptions. Furthermore a great amount of articles fails to report in a proper way or does not report at all which statistical test has been used for the data analysis of the paper [2],[3],[4].
Typical examples of misuse of statistical methodologies are outlined in the sequel. Discussing the appropriate assumptions of a statistical test, it is pointed out that the basic assumption for parametric tests is the assumption of normality. If data are not normally distributed, either non-parametric analytical techniques should be employed or data need to be transformed to a normal distribution.
Another issue is to distinguish when we must use parametric and when non-parametric test. In several papers [2] we notice the inappropriate use of statistical test like t-test etc.
Concluding out list of common mistakes one cannot refer to ANOVA and the striking small percentage of non-mathematicians who use the assumptions of ANOVA test in a sound way. Having defined the problem, we now proceed to the suggested solution. The following section explains why the identified problem can be advised by the employment of non-monotonic logic.

2 Non monotonic Logic

Why non standard logic?

The main goal of our paper is to explain to the reader why it is necessary to use a non monotonic logic instead of a standard logic. It is very important to understand that our software will be used from non mathematicians. This implies that available knowledge about statistical tests is incomplete. It is very difficult for researchers without expertise in mathematics to have the appropriate mathematical supervision[1].
Regarding the previous section our goal is also to help the non mathematician in a smooth way. This means that we need to construct a software which will be based on commonsense reasoning about mathematics. It is critical for us that the user will be able to reach to plausible conclusions from heir knowledge.
Another critical question is, whether the choice of the assumptions is blind or not. Most of the statistical knowledge is given by means of general rules which specify typical properties of statistical tests. For instance, "I will run ANOVA" means: I will run an ANOVA test, but there can be exceptions in several cases .
We must see the knowledge of the scientists as a knowledge base. If this is possible then we can use non monotonic logic. Non monotonic reasoning deals with the problem of deriving plausible, but not infallible,conclusions, from a knowledge base. Moreover since the conclusions are not certain, it must be possible to retract some of them if new information shows that they are wrong.
For example the closed-word assumption (CWA), in a formal system of logic used for knowledge representation, is the presumption that a statement that is true is also known to be true. Therefore, conversely, what is not currently known to be true, is false. If a software includes a representation of a sentence which means that something is not known, the logic under the software should be non-monotonic. Learning something that was not known leads to the removal of the object-option specifying that this piece of knowledge is not known. It is admitted that when an action is performed, some facts change and some do not. How do we tell which are which, how do we integrate those facts to the resulting situation in a smooth way without altering the initial conditions?
The standard logic as propositional logic and first order logic are monotonic logics, but what does this mean? A Logical system or a Logic consists of a proof system and semantics. In standard mathematical logic, the logic as a proof theory has the property of **monotony**.

$$\text{If } \Gamma \vdash \varphi \text{ then } \Gamma \cup \Gamma' \vdash \varphi \tag{1}$$

A non-monotonic logic is a Logic whose consequence relation is not monotonic.
Our choice is a specific branch of the AI area, the Argumentation Logic [9] and more concrete we will base our work on Gorgias-B the support tool for developing argumentation software [10],[11] . In the bibliography there are several examples which confirm that this argumentation deals with a huge numbers of real life applications such as deciding about an automatic freight process [12], smarter electricity [13], de-

cision support services [14], evaluating debates on social networks [15] etc
The next step of our work is to define the **argumentation theory for statistical analysis**, we will review some basic theory of argumentation, the basic definitions as well as the semantics of an argumentation theory. After that we will explain the implementation process from the theoretical framework to Gorgias System combining java, R and Gorgias-B.

3 The Case Study

An ANOVA case

Grocerybasket.com is an online grocery and food store established in 2011 in central Europe. Last year alone, Grocerybasket sold more than 18000 products from 1000 brands to over 3500000 customers out of whom about 30% place their orders through their mobile phones. According to "'The Retailer" (April 2017), Ernst and Young's publication in consumer products and retail sector, almost one in six European consumers (16%) bought groceries online, compared to 13% in 2013. This growth was due to the surging popularity of new business models and new players entering the market. EU grocery retailers invest on their online presence and customers experience, following major US companies like Amazon and Google in which recommender systems allegedly contribute to almost 35% of sales.
To face the stiff competition, Grocerybasket sets as its Unique Selling Point (USP) the customer convenience and invests on an updated website and new recommender system with two innovative features, namely "Smart Basket" and the "Did you forget something". Based on the past transactions of each customer, an individual customer profile is developed and patterns in the consumer's behavior can be traced to form the basket of items most likely to be ordered next. The "Smart Basket" option offers customers a list of products that are likely to be ordered at a given time based on past purchases. The "Did you forget something?" option relies on the comparison between the current shopping list of customers who do not use "Smart Basket" and the list automatically derived and suggests items that may have been missed. The real important question, however, is how do customers rate these features.
The Department of Digital Strategy of Grocerybasket launches an online survey with 3000 customers. The first group of customers (1000) will use the new system with the "Smart Basket' option activated; the second group of customers (1000) will use the "Did you forget something" option and the third group will use the legacy version of the company website (without the recommendation modules). At the end of their interaction with the system (recommenders or conventional) they are asked to rate their satisfaction using a 5-point Likert scale. Grocerybasket received 670 responses from the first group, 710 from the second and 820 from the third group. The Department of Digital Strategy expects to detect any significant group differences in customer satisfaction using the ANOVA (Analysis of Covariance) method, with the 'group' (first, second, third) as independent variable and the 'costumer' satisfaction as dependent variable. From a statistical perspective, however, this is not acceptable

because the dependent variable is ordinal and not continuous.
Questions on the proper amount of advanced multimedia elements used in the website are addressed by examining whether there are differences between the five classes of customers' satisfaction (1= not satisfied at all up to 5 = very satisfied) and the speed of the participants' network. Grocerybasket customers have typical network speed from 17 to 27 Mbps, while a few reside in regions where the speed network ranges from 67 to 81 Mbps. Developers plan to run ANOVA; this time the independent variable is the 'customer satisfaction level' as expressed by survey responders and the dependent variable is the 'network speed'. Running ANOVA in the sample, as described above, is improper; The sample contains outliers; namely, those with network speed ranging from 67Mbps to 81 Mbps. Thus, before running ANOVA, outliers must be excluded; moreover, it should have been confirmed that data are normally distributed in each of the five classes of responders.
This examples outline how statistical analysis can be easily sidetracked; unintentional errors that are saliently introduced in the statistical analysis may lead to wrong decisions with significant implications.
Furthermore, the assumptions of the one-way ANOVA test are at least 5.

1. Our dependent variable should be measured at the interval or ratio level

2. We should have independence of observations.

3. There should be no significant outliers.

4. Our dependent variable should be approximately normally distributed for each category of the independent variable.

5. There needs to be homogeneity of variances.

The figure 1 shows how should be the path of the solution when the scientist want everything to go well. But in real-world data one or more of these assumptions are often violated. Our goal is to construct a software that will enable the path to the solution to seem linear - figure 2. This will be achieved through the Argumentation theory and Gorgias.

Conclusions

Based on the presented arguments, it becomes clear that it is important to develop a system that has the capacity to assist scientists/professionals undertaking statistical analysis of data and enable them to select the appropriate statistical methodology. It is expected that the interaction between the system and its users will contribute to the improvement of the systems behaviour which will eventually reach the point to behave like a scientist without deep mathematical knowledge yet wide exposure to data analysis in particulars fields of study.

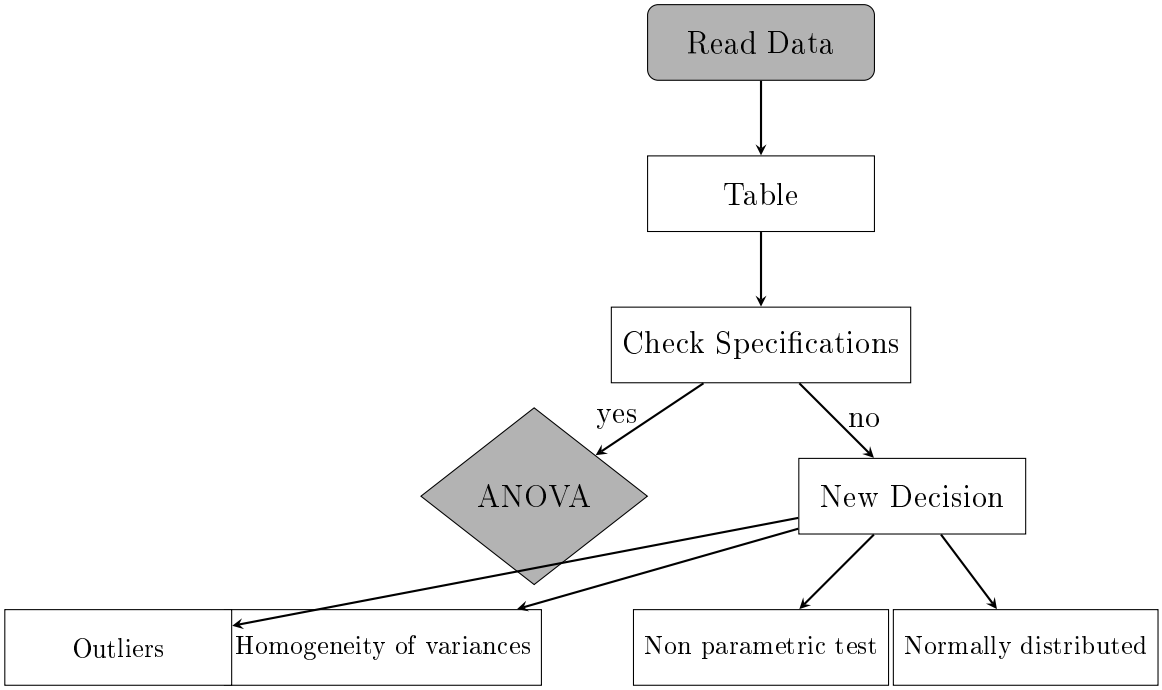


Figure 1: Standard ANOVA tree decision

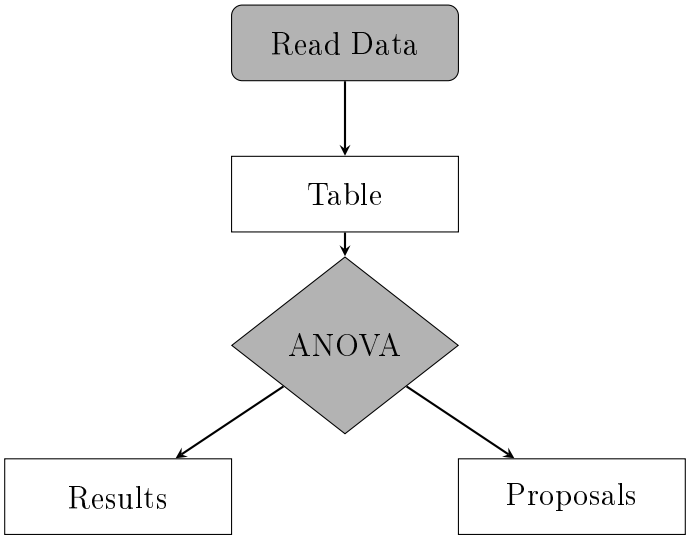


Figure 2: AI ANOVA tree decision.

Acknowledgements

We thank Anastasia Chatzigiannidi for comments that greatly improved the manuscript. We would also like to express our gratitude to the Vladimir Timofeev, Dean, Dr.Sc, Professor of NSTU, for giving us the opportunity to present our work.

References

[1] Strasak AM, Zaman Q, Pfeiffer KP, Gobel G, Ulmer H. Statistical errors in medical research–a review of commonpitfalls. Swiss Med Wkly 2007;137:44-9.

[2] Williams JL, Hathaway CA, Kloster KL, Layne BH. Low power, type II errors, and other statistical problems in recent cardiovascular research. Am JPhysiol 1997;273:H487-93.

[3] Feinstein AR. Clinical biostatistics. XXV. A survey of the statistical procedures in general medical journals. Clin Pharmacol Ther 1974;15:97-107.

[4] Welch GE, 2nd, Gabbe SG. Statistics usage in the American Journal of Obstetrics and Gynecology: has anything changed? Am J Obstet Gynecol 2002;186:584-6. http://dx.doi. org/10.1067/mob.2002.122144.

[5] Matthew S. Thiese et al (2015). The misuse and abuse of statistics in biomedical research. Biochemia Medica 25, 5–11. doi: 10.11613/BM.2015.001.

[6] Spanoudakis N., Kakas A.C., Moraitis P..: Conflicts Resolution with the SoDA Methodology. In the 2nd International Workshop on Conflict Resolution in Decision Making (COREDEMA 2016), held in conjunction with ECAI 2016, The Hague, Holland, 29 Aug, 2016

[7] Bench-Capon, T.J.M., Dunne, P.E.: Argumentation in artificial intelligence. Artif. Intell. 171(10-15), 619–641 (2007), http://dx.doi.org/10.1016/j.artint.2007.05.001

[8] Rahwan, I., Simari, G.R.: Argumentation in Artificial Intelligence. Springer Publishing Company, Incorporated, 1st edn. (2009)

[9] A. C. Kakas, F. Toni, and P. Mancarella. Argumentation for propositional logic and nonmonotonic reasoning. In Proceedings of the 29th Italian Conference on Computational Logic, Torino, Italy, June 16-18, 2014., pages 272–286, 2014.

[10] Spanoudakis N., Constantinou E., Kakas A.C.. Modeling Data Access Legislation with Gorgias. In the 30th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2017), Special Track on Applications of Argumentation (APPARG2017), Arras, France, June 27-30, 2017

[11] Karafili E., Kakas A.C., Spanoudakis N.I., Lupu E.C.. Argumentation-based Security for Social Good. In the AAAI 2017 Spring Symposium on AI for Social Good (AISOC17), Stanford University, March 27-29, 2017

[12] Chow, H.K.H., Siu, W., Chan, C., Chan, H.C.B.: An argumentation-oriented multi-agent system for automating the freight planning process. Expert Syst. Appl. 40(10), 3858–3871 (2013), http://dx.doi.org/10.1016/j.eswa.2012.12.042

[13] Makriyiannis, M., Lung, T., Craven, R., Toni, F., Kelly, J.: Combinations of Intelligent Methods and Applications: Proc. of the 4th International Workshop, CIMA 2014, Limassol, Cyprus, November 2014 (at ICTAI 2014), chap. Smarter Electricity and Argumentation Theory, pp. 79–95. Springer Int. Publishing, Cham (2016), http://dx.doi.org/10.1007/978-3-319-26860-6_5

[14] Fox, J., Glasspool, D., Patkar, V., Austin, M., Black, L., South, M., Robertson, D., Vincent, C.: Delivering clinical decision support services: There is nothing as practical as a good the- ory. Journal of Biomedical Informatics 43(5), 831–843 (2010), http://dx.doi.org/10.1016/j.jbi.2010.06.002

[15] Toni, F., Torroni, P.: Theorie and Applications of Formal Argumentation: First International Workshop, TAFA 2011. Barcelona, Spain, July 16-17, 2011, Revised Selected Papers, chap. Bottom-Up Argumentation, pp. 249–262. Springer Berlin Heidelberg, Berlin, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-29184-5_16

Applying Ideas of Experimental Design to LTS Estimation Parameters Scheme for Big Data Analysis

Ekaterina A. Khailenko
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: ekavka@yandex.ru

Abstract

Problem of robust estimation regression models' parameters is investigated. Applicable for big data analysis modification of method of Least Trimmed squares using experimental design is proposed. This modification is applied for real task of predicting the cost of residential real estate in Novosibirsk.

Keywords: big data, regression model, experimental design, method of least squares, robust methods, method of least trimmed squares.

Introduction

In practice, researchers often deal with the problems of constructing dependencies between input factors describing the operating conditions and the output data characterizing the result of this operation. To solve these kinds of problems, researchers use regression analysis methods, but it is well known [1] that these methods of estimating regression model parameters give correct results when performing a number of constraints [1]. But in practice these constraints can not be executed, for example, gross errors of observations (outliers) may be in sample. It was the cause that robust method of least trimmed squares is offered to use [3].

Moreover, it is known that it is possible to improve the accuracy of parameters estimation by using the optimal experiment plans. Previously in work [4], algorithm of constructing estimated subsets for LTS method based on the optimality criterion for the experimental design was proposed. As was shown by computational experiments, estimates of unknown parameters of regression models obtained using the classical LTS method are less accurate than estimates obtained using proposed modification. However this modification needs high computational resource for applying it for big data analysis. It was the cause that new modification for this method is offered, which can be apply for analysis of this kind of data.

1 Problem definition

Regression equation is considered

$$y = X\theta + e \tag{1}$$

where $X = \begin{pmatrix} f_1(x_{11}) & f_2(x_{12}) & \dots & f_m(x_{1m}) \\ f_1(x_{21}) & f_2(x_{22}) & \dots & f_m(x_{2m}) \\ \dots & \dots & \dots & \dots \\ f_1(x_{n1}) & f_2(x_{n2}) & \dots & f_m(x_{nm}) \end{pmatrix}$ - matrix of experimental plan, which has full column rank, i.e. $rg(X) = m$, m - number of regressors, n - number of observations, $f_1(x), ..., f_m(x)$ - vector of known real functions, x_{ij} - set values of input factors in n observations, $y = (y_1, ..., y_n)^T$ - vector of response values, $\theta = (\theta_0, ..., \theta_m)^T$ - vector of unknown parameters, $e = (e_1, ..., e_n)^T$ - vector of errors of observations, with respect to which it is assumed that the standard assumptions of the classical least-squares method are satisfied [1]:

$$E[e] = 0, D[e] = \sigma^2 I, \sigma^2 < \infty (I - unitmatrix)$$

Let also the measurements be according to optimal plan of experiment [6] $\xi^* = \begin{Bmatrix} x_1 & x_2 & \dots & x_s \\ p_1 & p_2 & \dots & p_s \end{Bmatrix}$ where $\sum_{i=1}^{n}(p_i) = 1$, $p_i = \frac{n_i}{n}$, s - number of points in spectrum of the plan and n_i - number of repeated observations in the i-th point in the spectrum of the plan.

The problem is that, based on the available initial data (response values and input factors), to take the best estimates of the vector of unknown regression model's parameters, while ensuring stability with respect robust to outliers and saving the quality of the initially specified experiment plan as much as possible.

2 Modification of Least trimmed squares method for big data analysis

The algorithm LTS method (FAST-LTS) for estimation regression model's parameters for big data analysis is proposed in work [3]. But this algorithm does not use information about observation when estimated subsets are constructing. As noted earlier in [4], the proposed algorithm considers the informativeness of each observation when estimation subsets for LTS-scheme, treating it as an independent plan of experiment. For the criterion of the D-optimality of the plan of experiment, which minimizes the determinant of the dispersion matrix, the optimality criterion minimizes the generalized variance of all estimates of the regression model. According to the equivalence theorem [6], the D-optimality condition of the plan can be written as follows:

$$\lambda(x)d(x, \xi^*) = m, x \in \xi^*$$

where $\lambda = \frac{1}{\sigma^2(x)}$ - function of effectiveness of plan, $d(x, \xi^*) = f^T(x)M^{-1}f(x)$ - variance of function of response in the point, M - information matrix. However, due to the fact that there are outliers in the sample, equality (2) may not perform at every point of the spectrum of the plan. Therefore, the criterion for adding observation to the estimated subset is the minimum of the following difference:

$$\varphi = max\,(\lambda(x)d(x, \xi^*)) - min\,(\lambda(x)d(x, \xi^*))$$

The following algorithm to solve the problem is proposed by author:

1. The FAST-LTS method is run before convergence, where the size of the estimated subset is pre-defined.

2. At each point of the spectrum of the initial plan ξ_0, the observations are sorted in order of increasing residuals.

3. The number of points included in the estimated subset to which the minimum residuals correspond:
$$n_h = m + \eta h$$
where a specific value η is chosen in advance. This representation allows us to take into account two boundary cases. When $\eta = 1$ there are only that observations in estimated subset, which corresponded to minimum residuals; when $\eta = 0$ estimated subset is formed using only design of experiment algorithms.

4. The new estimated subset for each point of the spectrum of the plan is recorded for $h \cdot n_i$ observations,$i = 1, .., s$, n_i - the number of repeated observations at each point of the spectrum. As a result, a plan is obtained that contains s points:
$$\xi^* = \left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_s \\ \frac{n_1}{n} & \frac{n_2}{n} & \dots & \frac{n_s}{n} \end{array} \right\}$$
The counter value k is set to n_h.

5. While $k < h$, the following sequence of actions is performed:

   a) the point is find for which
$$x^* = argmax(f^T(x)M^{-1}(\xi)f(x))$$
where ξ - plan, which has $k+1$ points, where k points are from ξ_k and point x is from plan ξ_0 , which is not exist in ξ_k;

   b) point x is added to estimated subset
$$\xi_{k+1} = \left(1 - \frac{1}{k}\right)\xi_k + \frac{1}{k}\xi(x^*)$$
.

   c) increase the value of the counter by one $k = k + 1$.

3 Results of investigations

The practical task was to predict the cost of residential real estate in Novosibirsk. The empirical base of an earlier study [5] was used, containing information on objects of real estate in Novosibirsk. Only two-room apartments were analyzed. After removing the missing values, the sample size was 2465 apartments. The following factors have been identified that have an impact on real estate prices:

-total square x_1,

-the material from which the house is built, at $x = 1$ the house is brick, at $x = 0$ - the panel.

In order to avoid the effect of scaling data, the logarithm of the price of apartment was taken as the cost.

The following model is taken for investigation:

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2}, i = 1, ..., n$$

where y_i - logarithm of the coast on apartments, $n = 2465$.

Figures 2 -3 show the dependence of the logarithm of the cost of apartment on the area and material of construction, respectively.

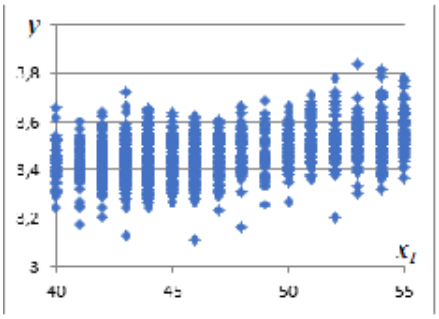


Figure 1: dependence of the logarithm of the cost of apartment on the area

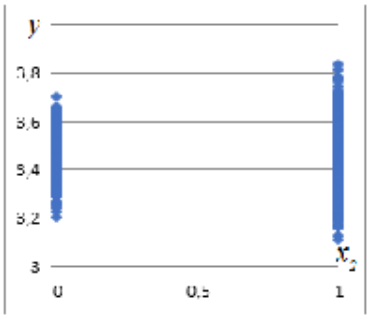


Figure 2: dependence of the logarithm of the cost of apartment on the material

According to Figures 2-3, it can be seen that the measurement area can be considered as an plan of experiment. Plan of experiment and the number of repeated observations at each point of the spectrum are presented in Table 1. The number of points in spectrum of plan is $s = 32$.

As an indicator of the accuracy of estimating the parameters of the regression dependence, the MAD (Mean Absolute Derivation) is taken, which is calculated by the formula [2]:

$$MAD = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Table 1: Plan of experiment

x_{1i}	40	40	41	41	42	42	43	43	44	44	45	45	46	46	47	47
x_{2i}	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
n_i	45	8	44	10	113	38	116	149	244	383	85	180	73	103	41	84
x_{1i}	48	48	49	49	50	50	51	51	52	52	53	53	54	54	55	55
x_{2i}	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
n_i	34	42	27	8	46	45	48	43	37	48	33	95	54	98	61	21

where n is the number of observations, in this case $n = 2465$, $\hat{y}$ – estimated vector of response.

A number of computational experiments were carried out and it was obtained that the most accurate results of the estimation are achieved with the size of the estimated subset is $h = 0.8n = 1965$ observations. Results of estimation by least square method (LS), FAST-LTS and LTS using experimental design and classical LTS are shown in table 2.

Table 2: Results of estimation regression model parameters

Method	MAD	φ	Time
LS	0.0313	21.2353	1
LTS	0.03601	13.1526	32.29
LTS using experimental design	0.0314	7.0971	65.751
FAST-LTS using experimental design	0.0376	6.4827	51.322

As can be seen from the table 2, the FAST-LTS method, using experiment planning, showed the most accurate evaluation results with a shorter execution time of the algorithm. In addition, it should be noted that the observations included in the evaluation subset are more homogeneous and maximally informative.

The results of constructing the forecast values of the logarithm of the apartment coast from the apartment area are shown in Figure 3.

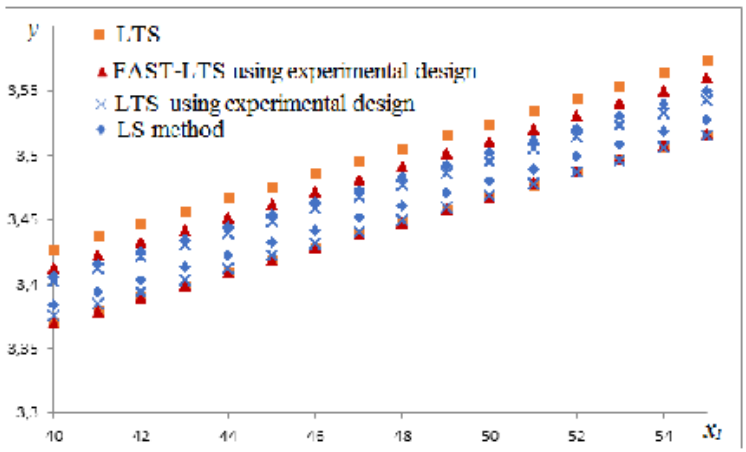


Figure 3: Diagram of range changes the number of students for all educational programs

The figure shows that the forecast, constructed using OLS estimates, has a smaller price spread for each area, depending on the material of the house construction. The methods of LTS and LTS based on the planning of the experiment showed close results and fairly close evaluation results and have a large enough price spread for each area, depending on the construction material. The FAST-LTS method, based on the planning of the experiment, showed a smaller price spread compared to the previous robust methods.

Conclusions

In this work the algorithm for estimating unknown parameters of regression models has been proposed. Results of proposed algorithm investigation are discussed. The advantage of the algorithm is that the estimates obtained are robust to the appearance of outliers, observations in estimated subset are homogeneous and maximally informative. This algorithm is applicable for big data analysis.

References

[1] Aivazyan S.A., Enukov I.S., Meshalkin L.D. (1985) Applied Statistics. M. Finance and Statistics. - 488p. (in Russian).

[2] Ph. Hampel, E. Ronchetti, P. Rousseeuw, V. Shtael (1989) Robust statistics. Approach based on the influence function. M. Mir. - 512p.(in Russian).

[3] Peter J. Rousseeuw, Katrien Van Driessen. (1999) Computing LTS Regression for Large Data Sets. Mimeo. *Dept. Mathematics, University of Antwerp* 21p.

[4] Timofeev V.S., Vostretsova E.A. (2009) Application of algorithms of planning an experiment in the scheme of LTS-estimation *Scientific Bulletin of NSTU* Vol. **1(34)**, pp. 95-105 (in Russian).

[5] Timofeev V., Timofeeva A., Kolesnikov M. (2014) Spatial concentration of objects as a factor in locally weighted models *12th International conference on actual problems of electronic instrument engineering, APEIE-2014- proceedings* Vol. **1**, pp. 567-570.

[6] Fedorov V.V. (1971) Theory of experimental design. M. Scince. 312p. (in Russian)

Robust Principal Component Regression on Compositional Covariates with Application to Educational Monitoring

Anastasiia Yu. Timofeeva
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: a.timofeeva@corp.nstu.ru

Abstract

Many economic data are presented in the form of shares. In particular, educational monitoring describes the specialization of higher education institutions with the help of proportions of students who study in different areas of training. The specialization impact to the indicators of activity of Russian universities is of great interest. To investigate this issue, the data of monitoring the effectiveness of HEIs are used, conducted by the Ministry of Education and Science in 2015. Regression estimation using the standard least squares method is impossible because there is the collinearity problem for compositional explanatory variables. For solving this problem several approaches may be used, including special transformations of compositional data and the principal component regression estimation. The article analyzes to what extent each approach provides robust estimation results with the least standard errors, and gives recommendations on their applicability. The obtained empirical results allow to describe impact of specialties to activity indicators of universities.

Keywords: principal component regression, compositional data, robustness, cross-validation, university, monitoring, indicator.

Introduction

Modern reform of the system of higher education in Russia is aimed at reducing the number of universities by optimizing the activities of inefficient universities, closing them, mergers and acquisitions, and supporting the leading universities. Such transformations are often made without taking into account the specifics of the activity of HEIs and their profile orientation, since they do not analyze in detail how the structure of the student contingent in the areas of training affects the performance indicators of universities.

At the same time, the program for the development of the education system emphasizes the priority of educational programs for the development of science, technology and technology of the Russian Federation, the need to upgrade the skills of engineering and technical personnel and increase the costs associated with training in educational programs on technical (engineering) training. In this regard, it is interesting to what extent the regulation of the structure of training directions (priority of technical orientation) will affect the performance indicators of universities.

1 Empirical data

To investigate this issue data of effectiveness monitoring of the educational institutions of higher education are used, conducted by the Ministry of Education and Science in 2015. The indicators are divided into the following groups. The number of indicators in each group is indicated in brackets.

1. Educational activity (15)
2. Research activity (16)
3. International activities (13)
4. Financial and economic activities (4)
5. Infrastructure (8)
6. Employment of graduates (1)
7. Staff (5)

Data downloaded from the web-pages of each individual institution. A total of 601 universities are represented in the sample, which provided information on the section "IV. The role of the organization in the system of personnel training for the region."

Data on the structure of the student of universities in the areas of training are contain 28 groups of directions. To combine similar specialties, the multidimensional scaling was used. As a measure of dissimilarity of directions Jacquard distance [1] was used. As a result, 8 generalized groups of specialties, or branches of science, were differentiated. They are presented in Table 1. In the second column, the first two digits of the code of the training directions are given, which are assigned to the corresponding branches of science.

Table 1: Grouped branches of science

Branches of science	Codes	The fraction of zeros, %
Mathematical, natural sciences, computer science, communication	01, 02, 09, 21, 23	33.5
Social, humanities, pedagogy, culture and art	03, 04, 05, 07	15.3
Healthcare	06	85.2
Economy, tourism	08, 10	14.3
Agriculture, Forestry	11, 25	79.8
Geodesy, exploration	12, 13	73.9
Technical	14, 15, 19, 20, 22, 24, 26, 27, 28	43.7
Defense	16, 17, 18	90.7

From 62 initial indicators of the activity of universities were selected those that are most correlated with the structure of the student contingent in the branches of science. For this purpose regressions of each efficiency indicator on the share of stu-

dents studying in seven selected areas are constructed. The latter direction (Defense) was not included in the model to avoid full collinearity. Preliminary processing of data included the removal of zero and missing values in performance indicators. In addition, these data were transformed to common logarithms. The significance of the models is verified by the F-statistic. Indicators corresponding to models that were not significant at 95 % confidence level are excluded. The quality of the models was evaluated by the coefficient of determination, the values of which are given in Figures 1. Employment indicator weakly correlates with speciality structure so it is not considered.

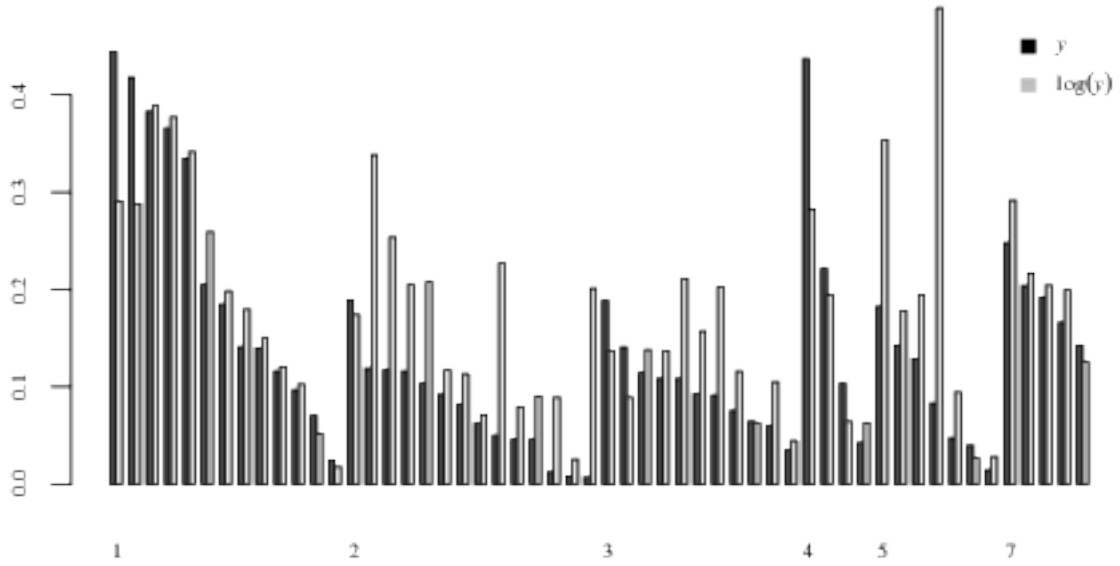


Figure 1: The coefficients of determination of models of university activity indicators on speciality structure

Of each group of indicators, one is selected, which has a strong correlation with the specialization of the university and is of research interest. Thus, the following output variables were further considered.

y_1 is the average score of the Unified State Examination received, with the exception of those taken for the target set and based on privileges.

y_2 is the total volume of research and development (in logarithms).

y_3 is the share of income from revenue-generating activities in total income.

y_4 is the total area of educational and laboratory facilities per student, leased (in logarithms).

y_5 is the number of scientific and pedagogical workers who have a scientific degree of a candidate and doctor of science per 100 students (in logarithms).

Preliminary analysis does not allow to give a conclusion on the direction in which different specialization of higher education institutions affects certain indicators. For the correct estimation of parameters of the regression with compositional covariates, special methods are required, which will be discussed below.

2 Methods of data analysis

Figure 1 clearly shows that for most indicators of research activities, the logarithm of the response significantly improves the quality of the regression model. This suggests that the distribution of the output variables is heavy-tailed and skewed to the right. The same concerns a number of indicators characterizing international activity and infrastructure. For a correct analysis of such data it is necessary to use robust estimation methods.

The use of compositional data as explanatory variables leads to the fact that the sum of proportions for any university is always 1. Therefore, full collinearity is observed, and the raw data can not be directly used as regressors in the model. Two approaches are considered to solve this problem. The first one is principal component regression. The second one is compositional data transformation. Further, these approaches are investigated within the scope of the considered practical problem.

2.1 Principal component regression

One of the approaches to regression analysis, used in conditions where a correlation is observed between input factors, is the construction of principal component regression [2]. The idea is to use the principal component method to decompose the input variables into orthogonal factors.

There are n observations of m input variables, so a data matrix X contains n rows and m columns. The columns in X are normalized with zero mean and unit variance. Then $R = X^T X$ is the correlation matrix for X. The matrix X^T denotes the transpose of X. The matrix R is a real symmetric matrix and its factorization into a canonical form is

$$R = \Lambda \Phi \Lambda^T \tag{1}$$

where an orthogonal matrix Λ contains the eigenvectors of R , and Φ is a diagonal matrix which entries are the eigenvalues of R . The eigenvalues of diagonal Φ imply the variances of the corresponding principal components. The eigenvectors in Λ are column vectors and represent the loadings of each variable on the corresponding principal component.

If the number of principal components equal to the number of variables, the decomposition (1) perfectly reproduces the correlation matrix R. By reduction m variables in q-dimensional sub-space ($q < m$) the correlation matrix is represented as

$$\hat{R}_q = \Lambda_q \Phi_q \Lambda_q^T$$

where Λ_q denotes the matrix Λ with the first q columns, i.e. the loadings on the first q principal component, and Φ_q is a diagonal matrix which entries are the first q eigenvalues of R. The principal components are sorted in order of decreasing eigenvalues so the first principal components keep the most important information from the data set.

Principal component scores F_q in q-dimensional sub-space are found by multiplying the original data matrix X by loading matrix Λ_q: $F_q = X\Lambda_q$.

If we use the matrix of principal components F_q as the input matrix in the model, the problem of collinearity is weakened because of the orthogonality of columns of the matrix F_q. In order to go back from the q -dimensional space of principal components to the original m-dimensional space, it is necessary to multiply the resulting vector of estimates by the matrix Λ_q. Then the final expression for the principal component regression estimates has the form

$$\hat{\beta} = \Lambda_q (F_q^T F_q)^{-1} F_q^T y \tag{2}$$

where y response vector of dimension $n \times 1$.

The choice of the number of components of q can be made on the basis of cross-validation or the Mallows's Cp criteria. However, it must be taken into account that the components are ordered in descending order of the percentage of the explained variance, that is, the quality of the prediction of the input variables. At the same time, the task is to predict the response values, so that not the best components in this sense can be selected.

Bair E. [2] proposed a technique called supervised principal components. Its essence lies in the fact that before selecting the principal components, input factors are selected in terms of correlation with the response. Since in my work the task does not lie in the method of the principal components, but in constructing the regression, it is important for me that the vector of parameter estimates is of dimension m. For this reason, some of the input factors can not be deleted, even if they are weakly correlated with the response. But one can use correlation with the response for selecting the main components. I call this approach supervised principal component regression (SPCR), the algorithm is the following.

Step 1. Compute all principal components of the data matrix.

Step 2. Compute (univariate) standard regression coefficients for each principal component on response.

Step 3. Form a reduced matrix F_q consisting of only those principal components whose univariate coefficient exceeds a threshold θ in absolute value (θ is estimated by cross-validation).

Step 4. Form a reduced loading matrix Λ_q consisting of only those columns whose principal components were chosen in step 3.

Step 5. Use F_q and Λ_q to estimate the principal component regression model.

2.2 Compositional data transformation

There are various kinds of data transformations, which weaken the correlation between them. The most popular is the isometric log-ratio transformation [3]. However, to use it, the compositional data must not contain zero values. For the case with the specialization of higher education institutions this condition is violated. As can be seen from Table 1, the share of universities that do not train in any of the directions is very large, up to 90 %.

One solution to this problem is to replace the zero values. It is usually done if it is assumed that zero is observed with inaccuracy, and in fact the true fractions deviate somewhat from zero. The situation is different when the specialization of the university is considered. If the university does not prepare for the direction "Defense", then adding to it a non-zero percentage of students studying in this direction distorts the initial ideas about the specialization of the university. Therefore, it is suggested to use approaches to weakening the problem of collinearity of input data, which allow for the presence of zero fractions and which do not assume their replacement.

Tsagris M. [4] proposed the use of a data based power transformation to perform regression analysis with compositional covariates. The big advantage of this approach is that zeros are handled naturally and thus no zero imputation technique prior to the analysis is necessary. The α-transformation is a more general than the isometric log-ratio transformation [3] and is a data based power transformation involving one free power parameter, similarly to the Box-Cox transformation.

Tsagris M. [4] suggested a way to choose the value of α which leads to the optimal results. The results show that prediction can be more accurate when one uses a transformation other than the isometric [3] or the additive log-ratio [5]. However, their works focus mainly on prediction and not on inference regarding the regression coefficients.

2.3 Robust estimation methods

To increase robustness of principal component regression estimates, it is suggested to use M-estimates of the parameter β on the basis of the Huber loss function [6] with the unit constant. For this, principal component regression is estimated using a weighted least squares method. The weight matrix W has the diagonal form with $W_{ii} = w(e_i/\sigma_e)$,

$$w(z) = \begin{cases} 1, z \leq 1, \\ 1/z, z > 1, \end{cases}$$

where $e_i = |y_i - \hat{y}_i|$, $\hat{y}_i$ is the prognostic value of the dependent variable obtained from principal component regression, $\hat{y} = X\hat{\beta}$, σ_e is the mean square deviation whose robust estimate is calculated as $\hat{\sigma_e} = \mathsf{med}(e_i)/0.67449$, $i = \overline{1,n}$. As the initial approximation $\hat{\beta}$ we can choose the estimate (2). The iteration process continues until the parameter estimates at the neighboring iterations differ at most by a certain small quantity. Vector $\hat{\beta}$ obtained after completion of the process is robust estimate of β.

Thus, eight approaches to constructing principal component regression have been implemented. The four main are

1. PCR - on the basis of cross-validation, the number of components ordered in descending order of percentage of the explained variance is chosen,

2. SPCR - based on cross-validation, a threshold value of univariate standard regression coefficients for principal component on response is chosen, those components are selected whose coefficients exceed the threshold value,

3. αPCR - PCR with preliminary α-transformation of compositional data,

4. αSPCR - SPCR with preliminary α-transformation of compositional data, and four more included robust estimation methods: PCR$_{rob}$, SPCR$_{rob}$, αPCR$_{rob}$, αSPCR$_{rob}$. Further, I have tested the proposed approaches when solving the real problem of estimating the influence of the structure of students on specialties on the selected indicators of the effectiveness of universities.

3 Empirical results

The choice of the optimal number of principal components was performed by means of Monte-Carlo cross-validation. The entire data set has been divided into equal parts: the training set and the check set. Such a division was carried out randomly and repeated 1000 times to obtain the stable results by subsequent averaging. To study the quality of the principal component regression estimation, the following statistic was used

$$S = \frac{1}{k}\sum_{i=1}^{k}\frac{|\text{mean}(\hat{\beta}_i)|}{\text{sd}(\hat{\beta}_i)} \tag{3}$$

where mean$(\hat{\beta}_i)$ is the estimate $\hat{\beta}_i$ averaged over all training sets, sd$(\hat{\beta}_i)$ is the standard deviation of the estimate $\hat{\beta}_i$ over all training sets, k is the number of elements of vector $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_k)$. The statistic (3) characterizes the quality of estimates. It grows with increasing absolute values of estimates and decreasing their deviation. The table 2 shows the values of the S-statistic.

Table 2: The quality of estimates of principal component regression

Algorithm	y_1	y_2	y_3	y_4	y_5
PCR	4.42	3.50	5.45	2.22	4.14
SPCR	5.23	3.88	5.45	1.49	4.13
αPCR	3.19	3.24	4.75	1.43	2.80
αSPCR	3.19	3.24	4.75	2.07	2.80
PCR$_{rob}$	6.04	3.79	6.46	2.27	5.08
SPCR$_{rob}$	10.50	4.76	6.02	1.62	5.06
αPCR$_{rob}$	6.63	4.50	4.75	2.10	3.29
αSPCR$_{rob}$	6.49	4.49	5.67	2.01	3.29

From Table 2 it is clear that robust method improves the accuracy of estimates of principal component regression. The best algorithms are SPCR$_{rob}$ and PCR$_{rob}$. Using α-transformation does not lead to the improvement of estimation quality. It should be noted that the choice of the optimal value of a threshold θ is rather complicated. The dependence of the mean prediction error on the check sets from the threshold θ is non-smooth. There is a lot of local minima and maxima. This fact complicates the use of optimization routines to find the best values of the threshold. Therefore, the most suitable method for practical tasks is the robust principal component regression.

Conclusions

The paper considers several approaches to estimation of principal component regression under conditions when the input variables represent the compositional data. Their robust modifications based on Huber M-estimates are proposed. The study is motivated by the solution of the real problem of estimating the influence of the structure of students on specialties on the indicators of the effectiveness of universities. On the data of Russian universities, the work of the proposed algorithms has been tested. It turned out that the use of robust modifications makes it possible to improve the accuracy of estimating regression parameters. At the same time, the use of α-transformation of compositional data does not lead to a significant improvement. This indicates that for practical tasks it is sufficient to use robust principal component regression without transformation of the compositional data.

Acknowledgements

This work was supported by the Grant from the President of the Russian Federation for Young Russian Scientists, No. MK-5385.2016.6.

References

[1] Reina D.G., Toral S.L., Johnson P., Barrero F. (2014). Improving discovery phase of reactive ad hoc routing protocols using Jaccard distance. *The Journal of Supercomputing*. Vol. **67**, pp. 131-152.

[2] Bair E., Hastie T., Paul D., Tibshirani R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*. Vol. **101**, pp. 119-137.

[3] Egozcue J.J., Pawlowsky-Glahn V., Mateu-Figueras G., Barcelo-Vidal C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*. Vol. **35**, pp. 279-300.

[4] Tsagris M. (2015). Regression analysis with compositional data containing zero values. *Chilean Journal of Statistics*. Vol. **6**, pp. 47-57.

[5] Aitchison J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*. Vol. **44**, pp. 139-177.

[6] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (2011). *Robust statistics: the approach based on influence functions*. John Wiley & Sons, New York.

Effect of Factors on Professional Employment Types for Graduates: Testing for Association between the Categorical Variables

Alena A. Borisova, Anastasiia Yu. Timofeeva, Maxim A. Bakaev
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: bborisova2012@yandex.ru, a.timofeeva@corp.nstu.ru, bakaev@corp.nstu.ru

Abstract

Our paper is dedicated to the problem of higher education institutions graduates' professional employment. We discover the factors that affect the possible employment type are: 1) specialized, 2) delayed specialized, 3) non-specialized with some professional experience and 4) non-specialized employment. Since employment type is a nominal variable, we test the hypothesis with the Chi-square test. To exclude false correlation, we also check the conditional independence hypotheses using the Cochran-Mantel-Haenszel test. Besides the standard hypotheses testing procedures, bootstrap is employed, which turns out to be advantageous. The empirical data are from the survey that we performed with graduates of the selected Siberian universities. Finally, we identified and interpreted the factors predicting specialized employment of higher education institution graduates.

Keywords: employment, higher education institution graduates, specialized jobs, chi-square test, Cochran-Mantel-Haenszel test, bootstrap.

Introduction

The current research works in education and employment of graduates [1] identify disproportions on the labor markets and emphasize the necessity to normalize contradictions existing between requirements of industry and the capability of educational system to satisfy them. There's still the need for additional research seeking to clarify the reasons responsible for these disproportions and identifying the factors behind them, as well as justifying the effectiveness of introducing normalizing measures in the interaction between educational institutions and employers.

Nowadays, the social and economical premises for emergence of demand for highly skilled labor are studied intensively [2], and the structural disparities on young specialists' labor market are explored, including appraisal of various measures for increasing their specialized employment [3]. The analysis of approaches towards identification of the factors causing the disparities suggests differences in both selecting the predictors and the resulting value that is used to diagnose the strength of the relation. So, in socio-psychological and economic research they often study the relation between profession/university selection motivation and professional achievements [4] and determine the significant predictors for overall career and life success [5].

Thus, the discovery of the factors determining the specialized employment of graduates is an important problem and in our paper we seek to contribute to the knowledge on how various predictors influence the construction of specialty-oriented educational and career paths.

1 The Empirical Data

To discover the significant predictors for the specialized employment and their effect on the graduates' distribution between the employment types, we collected empirical data on 1264 educational and career paths for university students of "Labor Economy", "HR Management" and "Management" majors. In the longitudinal study we monitored the increase in the professional qualification of the students during their educational term and two years after the graduation. The analysis and the conclusions are based on the results of the career path monitoring for students of 6 universities in the Russian cities of Novosibirsk and Irkutsk.

The model implies that the outcome of the graduates' educational path is the degree of their employment specialization – i.e. correspondence to their major. Based on this criterion, we identified the four employment types: 1) specialized, 2) delayed specialized, 3) non-specialized with some professional experience and 4) non-specialized employment. The independent variables are the groups of predictors that initially aggregated the indices set.

The analysis was performed in three steps. First, we determined the effect strength for each index on the resulting variable – the employment type. Then we constructed frequency distribution of the responders in each significant index for employment types. Finally, we checked the "purity" of the relation between the significant indexes and each employment type.

2 Methods of testing for association between the categorical variables

To identify significant predictors for specialized employment of the graduates, we used contingency tables and the Chi-square criterion [6], since the variables were of nominal scale. The hull hypothesis is that there is no association between two variables. The χ^2 statistic is calculated based on the formula

$$\chi^2 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(p_{ij} - d_{ij})^2}{p_{ij}} \tag{1}$$

where d_{ij} is observed frequency associated to ith row and jth column of $n_1 \times n_2$ contingency table, p_{ij} is expected frequency calculated under the assumption of independence of categorical variables. The statistic (1) has a chi-square distribution with $(n_1 - 1)(n_2 - 1)$ degrees of freedom.

However, it should be noted that the popular Chi-square test is asymptotic in nature and is useful when the cell frequencies are "not too small". In [6], the authors explore the accuracy of the Chi-square tests through an extensive simulation study and then propose their bootstrap versions that appear to work better than the asymptotic Chi-square tests. The bootstrap tests are useful even for small-cell frequencies as they maintain the nominal level quite accurately.

For hypotheses testing the quantitative and ordinal variables were transformed into nominal by grouping (categorization). The breakdown into groups was carried out in such a way that each group received an approximately equal number of the analyzed objects (graduates). For this purpose, quartiles of the empirical distribution of quantitative characteristics subject to categorization were used as the boundaries of intervals. In addition to the standard hypothesis testing procedures, a bootstrap is used, which turns out to be preferable [6].

The two-dimensional contingency tables and the Chi-square statistics calculated for them only allow reasoning about effect of a single predictor on specialized employment. However, the nature of a graduate's employment is shaped under influence of many factors, some of which can be also interrelated. Thus, analysis of paired relations only can lead to a wrong understanding of the patterns for specialized employment, due to false correlation problem. In order to avoid such incorrect conclusions, the Cochran-Mantel-Haenszel chi-squared test was used [7]. The null hypothesis is that two nominal variables are conditionally independent, in each stratum, assuming that there is no three-way interaction. In this case the odds ratios for all partial tables are equal to 1. The test statistic is calculated based on the formula

$$\chi^2_{CMH} = \sum_{k=1}^{K} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(p_{ijk} - d_{ijk})^2}{p_{ijk}} \tag{2}$$

where d_{ijk} is observed frequency associated to ith row and jth column of $n_1 \times n_2$ contingency table for kth stratum, p_{ij} is expected frequency calculated under the assumption of independence of categorical variables for kth stratum, K is number of strata. The statistic (2) has a chi-square distribution with $(n_1 n_2 K - n_1 - n_1 - K + 2)$ degrees of freedom.

The strata were identified based on several features describing the education: university, major, graduation year, qualification level, and region. With the Cochran-Mantel-Haenszel criterion, we found partial correlations between each of the significant predictors for specialized employment and a feature from the considered set, with the effect of the other feature removed. On overall, only 65 % of the indexes pre-selected for research were recognized as significant predictors of specialized employment.

3 Results

Let us consider the nature and direction of the relation between the groups of predictors with the employment types described earlier.

3.1 The Environment Factors

The diagnostics of the educational market parameters' effect on the graduates' employment type identified the significant influence of the following factors. Here and further on we show in brackets the values for statistics (1) and minimal values for statistics (2).

- University ($\chi^2 = 52.3$,$\chi^2_{CMH} = 18.9$). Some universities are able to achieve notably higher frequency of specialized employment, and each fifth of their graduates who has worked in some other field later does come back to specialized job. It can be said that university's reputation and the specialized demand for its graduates are strong signals for entrants and their initial professional qualification.

- The broadness of specialization ($\chi^2 = 38.3$,$\chi^2_{CMH} = 3.5$). Broader professional education allows more frequent implementation of specialized and delayed specialized employment types, while narrower one prevents specialized employment. However, if the effect of the university is removed, the effect for the broadness of specialization is no longer significant.

- The year of graduation ($\chi^2 = 18.2$,$\chi^2_{CMH} = 15.6$).

The analysis of the current labor market condition on the frequency of specialized employment let us identify the following significant predictors.

- The graduates' professional qualification ($\chi^2 = 27.3$,$\chi^2_{CMH} = 25.5$). Its lower level causes higher frequency of non-specialized employments and increases the workers' mobility on the market. More than 70 % of those who work outside of their majoring field say that the main reason for their choice of the job was inability "to win the competition for good jobs" due to insufficient competitiveness of their professional potential.

- Excess labor supply by the market in "best qualification" and "good work experience" ($\chi^2 = 30.2$,$\chi^2_{CMH} = 28.7$). Almost every second graduate refuses to "fight" for specialized employment if the state of the market implies excessive supply of labor. This may also explain the significance of the interest towards the performed job factor ($\chi^2 = 35.7$,$\chi^2_{CMH} = 34.7$).

So, the satisfaction of the requests for specialized employment is delayed, it requires additional increase in professional qualification, and the employer has to justify spending on a young specialist's professional development. At the same time, non-specialized employment allows leveling the effect of the predictors and contributes to higher value of employment competitiveness index, and thus ensures higher priority of non-specialized employment types by the graduates.

3.2 The Behavioral Factors

Important goals for students planning the initial steps of their career path are increasing their professional value on the labor market ($\chi^2 = 35.4$,$\chi^2_{CMH} = 33.4$) and their professional potential ($\chi^2 = 46.4$,$\chi^2_{CMH} = 45.6$). Domination of the "obtaining specialized experience" goal ensures higher frequency of specialized employment for graduates. If selection of the first workplace is driven by the desire to obtain just any kind of work experience, then the frequency of non-specialized employment grows. The analysis of such characteristics as gaining independence, higher self-reliance, and material self-dependence for senior students did not find their significant effect on employment type.

The activity of graduates on labor market was analyzed as the frequency of changing employers ($\chi^2 = 66.8$,$\chi^2_{CMH} = 59.3$) and work positions ($\chi^2 = 73.3$,$\chi^2_{CMH} = 68.1$). Higher changeability of employers and work positions is typical for non-specialized employment with some professional work experience. The delayed specialized employment is described by lower mobility on the labor market, and this low activity (corresponding to changing 1-2 companies and no more than 2 work positions during the observed 2-year period) generally causes greater frequency of specialized employment. Thus, organizational measures related to introduction in a new work position become highly important. Extended socialization period of today's youth also requires more attention to aiding young specialists' adaptation and securing them in a company.

3.3 The Organizational Factors

The significant relation between the employment type and the demand for the professional potential in the current business environment was found (the first workplace $\chi^2 = 297.9, \chi^2_{CMH} = 294.4$, the current workplace $\chi^2 = 349, \chi^2_{CMH} = 339.9$), the growth ($\chi^2 = 306.5$,$\chi^2_{CMH} = 295.5$). Securing on the specialized labor market causes higher actualization of the potential: so, for 74.6 % of the graduates who got a specialized work position after graduating, the demand for the potential was over 60 %, and after two years already 90.2 % of young specialists attained this potential level. This is clearly seen in Figure 1 which shows the contingency table of the employment type and the demand for the professional potential in the current workplace. We also found the opposite trend: moving towards the non-specialized types of employment cause a notable decrease in the demand for the potential. The searching behavior aimed towards application of professional training is accompanied by the positive dynamics of the potential use. The refusal of specialized work and choosing another kind of job significantly decreases the degree of the potential engagement, even if it was formed previously.

So, the choice of the first workplace shapes further graduate's behavior on the market, as specialized and non-specialized employments cause lower mobility and moderate potential growth rate, while searching models, on the contrary, lead to higher dynamics of the demand. Significant predictors that describe the employer company are its size in the number of current employees ($\chi^2 = 24.4$,$\chi^2_{CMH} = 23.4$)

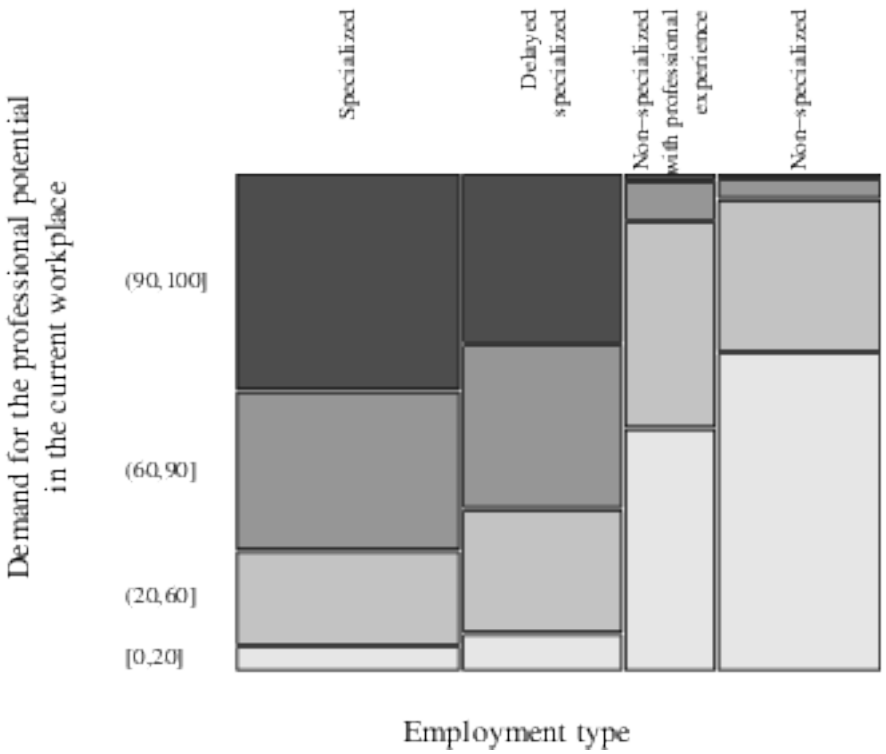


Figure 1: The contingency table of the employment type and the demand for the professional potential in the current workplace

and the degree of personnel turnover ($\chi^2 = 19.8$,$\chi^2_{CMH} = 19.7$). Employment with larger companies (more than 500 workers) allows the specialized employment type more often. We also observed disparity in the graduates' distribution within searching jobs: most of them were in smaller companies, and every fourth – in large ones (more than 1000 workers). Possibly, the choice of a larger enterprise with longer history is justified by search for higher stability and decreasing the chance of further movements.

Higher securement of young specialists in non-specialized jobs is particularly caused by low personnel turnover in the organization. So, every second graduate in a non-specialized job reported the low turnover, while only one-third of the ones employed in specialized jobs testify stability, security and are not afraid of being fired. The testing of the other potentially predictive characteristics of organizations (e.g. the type of industry, the legal ownership type) did not find a significant relation with the employment type.

The wage level does have significant effect on the employment type: important are wages after probationary period in the first workplace ($\chi^2 = 19.7$,$\chi^2_{CMH} = 18.6$) and during adaptation in subsequent jobs ($\chi^2 = 16.5$,$\chi^2_{CMH} = 16.2$). Supply of higher wages stimulates the searching behavior and causes better security of graduates in specialized jobs. Despite that specialized employment allows higher income, the effect of the other limiting factors that we considered before (e.g. higher market entry barrier, security from being fired, etc.) negates this important advantage. A

relatively rapid adaptation of the graduates was noted, but there was a contradiction: the specialized employment does not provide a significant benefit in the position adaptation speed. On the contrary, non-specialized employment and non-specialized one with some professional work experience offer faster reach of the normal return on labor. Possibly, the essence and contents of the job as well as available adaptation programs for non-specialized positions allow graduates without due qualification to learn the nature of their trade quicker.

The duration of working in organization ($\chi^2 = 34.2$,$\chi^2_{CMH} = 28.1$) affects the employment type: longer interaction with an employer are more often reported for specialized employment in both the first (72 %) and the second jobs (69.4 %). The searching models are generally accompanied by short-time cooperation. The pace of workplace change in delayed specialized employment is 1.5 times higher than for non-specialized one with some professional work experience, but this difference is smoothened in the subsequent jobs.

We diagnosed organizational limitations for the specialized employment – the conditions for career advancement ($\chi^2 = 18.3$,$\chi^2_{CMH} = 17.5$). The limitations in vertical promotion decrease the probability of staying in non-specialized jobs, as higher difficulty of getting a specialized promotion causes higher security in non-specialized employment and ensures mobility of young specialists between companies. Domination of the "closed" HR management policy on the youth labor market leads to ageing of personnel, prevents renewal of task fulfillment technologies, and ultimately contributes to erosion of intellectual potential and depreciation of return on professional education.

3.4 The Professional Qualification Factors

Additional training ($\chi^2 = 10.1$,$\chi^2_{CMH} = 12.7$) mostly affects specialized and delayed specialized employments. We also noted that in the majority of specialized employment cases the education was self-funded. The total spending on non-specialized trainings was significantly higher than for growing the professional potential. Enthusiasm for the job ($\chi^2 = 19$,$\chi^2_{CMH} = 17.1$) was more often reported for specialized activities. More than 75 % of the graduates employed in specialized jobs testified having emotional satisfaction and the plans for professional growth. The delayed specialized type had higher degree of satisfaction due to selection of career path.

The successful creation of the potential ($\chi^2 = 17.6$,$\chi^2_{CMH} = 13.8$) that was mostly diagnosed during the education period was a significant predictor for the specialized employment. Good education twice as often allows getting specialized jobs and is a solid reason for deciding to return to specialized path. The students who had a non-specialized job during their education period, more seldom got excellent grades for their final thesis works. Possibly, the decrease in motivation in the professional career happens even before graduation and the interest in making a good thesis project is lost.

Conclusion

The results reported in our current paper can be used for: planning individual career paths; improving interaction between educational institutions and employers – e.g. in forecasting a perspective graduate model; instrumental support for selection of interns and attracting specialists; performing measures for adaptation and job introduction; shaping the HR politics. Monitoring the direction and strength of the specialized employment predictors' effect can be used in justification of preventive measures aimed on securing the young workers in the chosen professional direction during early career stages.

Acknowledgements

The reported study was funded by RHSF/RFBR according to the research project No. 17-32-01087 a2.

References

[1] Lisovsky V.T. (1996). *Sociology of Youth*. SPbSU, St. Petersburg.

[2] Gimpelson V.E., Kapelyushnikov R.I., Lukyanova A.L. (2007). *Demand for labor and qualifications in industry: between deficit and surplus*. Moscow State University, Higher School of Economics, Moscow.

[3] Serova L., Fedorova E. (2014). Employment of graduates: survey of monitoring in 2013. *Employment Service*. No. **1**, pp. 44-47.

[4] Rosenbaum J.E., Kariya T., Settersten R., Maier T. (1990). Market and network theories of the transition from high school to work: their application to industrialized societies. *Annual Review of Sociology*. Vol. **16**, pp. 263-299.

[5] Springett N.R. (2009). Course satisfaction and occupational egoidentity among undergraduates. *Higher Education*. Vol. **15**, pp. 323-331.

[6] Lin J. J., Chang C. H., Pal N. (2015). A revisit to contingency table and tests of independence: bootstrap is preferred to Chi-Square approximations as well as Fisher's exact test. *Journal of biopharmaceutical statistics*. Vol. **25**, pp. 438-458.

[7] Mantel N., Haenszel W. (1989). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute*. Vol. **22**, pp. 719-748.

About One Criterion of Verifying the Independence of Observations *

Khrushchev S.E.[1], Logachov A.V.[2] and Logachova O.M.[3]
[1,2,3] *Novosibirsk State University of Economics and Management, Novosibirsk, Russia*
[2,3] *Siberian State University of Geosystems and Technologies, Novosibirsk, Russia*
[3] *Sobolev Institute of Mathematics, Novosibirsk, Russia*
e-mail: s.e.hrushchev@edu.nsuem.ru, omboldovskaya@mail.ru

Abstract

Some criterions of verifying the independence of observations are considered in this paper. The distribution of samples' elements is unknown.

Keywords: Sample, criterion, regression.

Introduction

We are interesting in the criterion that allows us to check the hypothesis on the homogeneity and independence of sampling elements (further, samples possessing these two properties will be called simple) of a small volume consisting of random variables having the continuous distribution. Various series criterions are used to check the simplicity of the sample most often in cases where nothing is known about the distribution of its elements. Note that these criterions are asymptotic, therefore, the existence of moments of a certain order of random variables, included in the sample, are required except conditions for sample size, see, for example [1, chapter 11].

The criterion is obtained is independent of the distribution of random variables and can be used for small-volume samples.

$\hat{y}(x)$ denotes the function that defines equation of selective pairwise linear regression, e_i denotes residuals of regression, α is level of significance (probability of making an error of the 1st kind), $\mathbb{Z}^+$ is the set of non-negative integers.

The paper is constructed according to the following plan: Review of well-known criterion is done and notations are introduced in section 1; The main result is proved (Theorem 2.2) in section 2; The received criterion is tested on several samples, conclusions are made in section 3.

1 Main results

A sequence of random variables $\{X_n\},\ n \in \mathbb{Z}^+$ is considered.

We introduce the notation

$$N_{\min} := \min\{n \geq 1 : X_n < X_0\},\ N_{\max} := \min\{n \geq 1 : X_n > X_0\},$$

where $\min \varnothing = \infty$.

*This work was supported by NSUEM grant N 276-2017

Lemma 1. *Let random variables $X_0, X_1, \ldots, X_n, \ldots$ are independent copies of a random variable X, which has a continuous distribution function. Then*

$$\mathbf{P}(N = n) = \frac{2}{n} - \frac{2}{n+1}, \quad n \geq 2, \tag{1}$$

where $N := \max\{N_{\min}, N_{\max}\}$.

Proof. It's obvious that $\mathbf{P}(N = 1) = 0$. By virtue of incompatibility of events $\{\omega : N_{\min} = n\}$ and $\{\omega : N_{\max} = n\}$ for $n \geq 2$ we have

$$\mathbf{P}(N = n) = \mathbf{P}(\{N_{\min} = n\} \cup \{N_{\max} = n\}) = \mathbf{P}(N_{\min} = n) + \mathbf{P}(N_{\max} = n). \tag{2}$$

Due to the fact that random variables $X_0, X_1, \ldots, X_n, \ldots$ are independent and identically distributed for any $n \geq 1$ we have

$$\mathbf{P}(N_{\min} > n) = \mathbf{P}(\min_{0 \leq k \leq n} X_k = X_0) = \frac{1}{n+1}.$$

Therefore

$$\mathbf{P}(N_{\min} = n) = \mathbf{P}(N_{\min} > n-1) - \mathbf{P}(N_{\min} > n) = \frac{1}{n} - \frac{1}{n+1}. \tag{3}$$

Similarly the following equality is obtained, see also [2, P. 30]

$$\mathbf{P}(N_{\max} = n) = \frac{1}{n} - \frac{1}{n+1}. \tag{4}$$

Formula (1) is follows from the equalities (2)–(4). □

Remark 1. *From Lemma 1 it is obviously that*

$$\mathbf{P}(N > n) = \frac{2}{n+1}, \quad n \geq 2.$$

Let $X_0, X_1, \ldots, X_n$ is a sample.

Theorem 1. *Let the level of significance α is given and the hypothesis*

$$H_0 : \textit{the sample is simple}$$

is verified against a competing hypothesis

$$H_1 : \textit{the sample is not simple.}$$

Then the following criterion is valid:

$$\textit{if } N \leq \frac{2}{\alpha}, \textit{ then the hypothesis } H_0 \textit{ is accepted,}$$

$$\textit{if } N > \frac{2}{\alpha}, \textit{ then the hypothesis } H_0 \textit{ is rejected.}$$

Proof. It is obviously follows from the Remark 1.□
Denote

$$N^-_{\min} := \max\left\{n < \left[\frac{n}{2}\right] : X_n < X_{\left[\frac{n}{2}\right]}\right\},\ N^-_{\max} := \max\left\{n < \left[\frac{n}{2}\right] : X_n > X_{\left[\frac{n}{2}\right]}\right\},$$

$$N^+_{\min} := \max\left\{n > \left[\frac{n}{2}\right] : X_n < X_{\left[\frac{n}{2}\right]}\right\},\ N^+_{\max} := \max\left\{n > \left[\frac{n}{2}\right] : X_n > X_{\left[\frac{n}{2}\right]}\right\}.$$

Theorem 2. *Let the level of significance* $0 < \alpha \leq 4/9$ *is given and and the hypothesis*

$$H_0 : \textit{the sample is simple}$$

is verified against a competing hypothesis

$$H_1 : \textit{the sample is not simple.}$$

Then the following criterion is valid:

$$\textit{if } \min\{N^-, N^+\} \leq \frac{2}{\sqrt{\alpha}}, \textit{ then the hypothesis } H_0 \textit{ is accepted,}$$

$$\textit{if } \min\{N^-, N^+\} > \frac{2}{\sqrt{\alpha}}, \textit{ then the hypothesis } H_0 \textit{ is rejected,}$$

where $N^- := \min\left\{\left[\frac{n}{2}\right] - N^-_{\min}, \left[\frac{n}{2}\right] - N^-_{\max}\right\}$, $N^+ := \min\left\{N^+_{\min} - \left[\frac{n}{2}\right], N^+_{\max} - \left[\frac{n}{2}\right]\right\}$.

Proof. Let the hypothesis H_0 is valid, then the random variables N^- and N^+ are independent and it follows from the remark 1 that

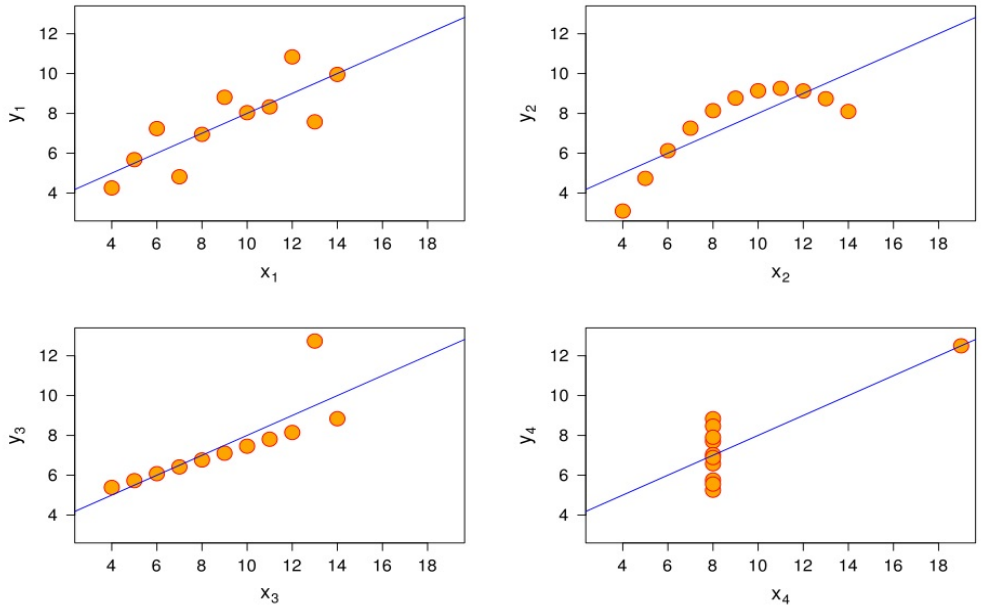
$$\mathbf{P}\left(\min\{N^-, N^+\} > \frac{2}{\sqrt{\alpha}}\right) = \mathbf{P}\left(\min\{N^-, N^+\} > \left[\frac{2}{\sqrt{\alpha}}\right]\right) =$$

$$= \left(\frac{2}{\left[\frac{2}{\sqrt{\alpha}}\right] + 1}\right)^2 < \alpha.\square$$

2 An example of applying the criterion

Let's test the received criterion on remainders of the linear regression for samples from the Anscombe Quartet [3].

I		**II**		**III**		**IV**	
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89



Correlation fields for samples I-IV are presented in figures above. In all four cases, the selective linear regression equation will have the view

$$\hat{y} = 0,5x + 3,$$

and coefficient of determination is $R^2 = 0,816$.

We give remainders of regressions ordered by increasing of the explanatory variable in the following table.

Numb.	e_0	e_1	e_2	e_3	e_4	e_5
I	$-0,74$	$0,18$	$1,24$	$-1,68$	$-0,05$	$1,31$
II	$-1,9$	$-0,76$	$0,13$	$0,76$	$1,14$	$1,27$
III	$0,39$	$0,23$	$0,08$	$-0,08$	$-0,23$	$-0,39$
IV	$-0,42$	$-1,24$	$0,71$	$1,84$	$1,47$	$0,04$

Numb.	e_6	e_7	e_8	e_9	e_{10}
I	$0,04$	$-0,17$	$1,84$	$-1,92$	$-0,04$
II	$1,14$	$0,76$	$0,13$	$-0,76$	$-1,9$
III	$-0,54$	$-0,69$	$-0,85$	$3,24$	$-1,16$
IV	$-1,75$	$-1,44$	$0,91$	$-0,11$	0

A visual analysis of graphs above suggests that residuals of regressions will be independent only for cases I, IV.

The criterion from the Theorem 1 accepts hypotheses that samples I and IV are simple on any level of significance $\alpha < 2/3$, the Theorem 2 allows us accepts hypotheses that samples I and IV are simple on any level of significance $\alpha < 4/9$.

The criterion from the Theorem 1 allows us to reject the hypothesis, that the sample II is simple on any level of significance $\alpha \geq 0,2$, the Theorem 2 allows us to reject the hypothesis, that the sample II is simple on any level of significance $\alpha \geq 0,16$.

The criterion from the Theorem 1 allows us to reject the hypothesis, that the sample III is simple on any level of significance $\alpha \geq 0,25$, the Theorem 2 allows us to reject the hypothesis, that the sample III is simple on any level of significance $\alpha \geq 0,22$.

Note that series criterions can not be applied because of the small sample size.

References

[1] Aivazian S.A. Applied statistics, M. – 1983.

[2] Feller W. Introduction to probability theory and its applications, M., Vol. 2. – 1966.

[3] Anscombe F.J. Graphs in statistical analysis //The American Statistician. – 1973. – Vol. 27. – N 1. – pp. 17-21.

On Improving Statistical Estimation by Utilizing Collateral Information ("Guesses"): a Case of the Probability Estimation

Yu.G.Dmitriev, F.P.Tarasenko, P.F.Tarasenko
National Research Tomsk State University, Russia
e-mail: dmit@mail.tsu.ru, ftara@ich.tsu.ru, ptara@mail.tsu.ru

Abstract

The statistical estimation of any functional of a probability distribution function may be improved by merging it with any additional, collateral, side information ("prior guesses") about the estimated characteristic. This paper considers such a possibility in case of estimating a probability of a certain stochastic event.

Keywords: Nonparametric estimation of distribution parameters, joining of heterogeneous statistical data, optimal statistical procedures, adaptive procedures.

Introduction

Cognition of any aspect of surrounding us reality (either objects or their behavior) begins from obtaining (and recording) results of our practical interactions (observations and/or measurements) with the item of interest. Then this obtained "raw" information (the experimental data) is purposely processed and transformed in models of the matter considered. These models systemically represent essential (i.e. important to us) characteristics of the matter.

If we are dealing with an object X of stochastic nature, all information about it is exhaustively condensed in its probability distribution function $F(x)$: in this case any numerical characteristic of interest, J, is a quantitative image of certain distribution quality that may be designated by certain functional of the distribution: $J = J(F)$. Mises [1] has initiated "the substitution approach" to statistics by suggestion to construct a statistical estimate of any functional by substituting the empirical distribution function F_n (x) into the functional: $\hat{J} = J(Fn)$. Many other statisticians (including Siberian ones, e.g. [2] - [10]) extended this idea to substituting of various nonparametric estimates of other presentations of the distribution (cumulative, density, and/or their derivatives) represented in an estimated functional. Quality of the obtained estimate may usually be characterized by its mean square error (MSE), $MSE(\hat{J}) = M(\hat{J}-J)^2$, and statisticians make special efforts to minimize it by varying changeable parameters in the resulting formula of estimate.

However, the accuracy of estimation still may be improved further if we could join to it some additional information about the estimated parameter, obtained from other sources, besides the sample. Such a possibility stems from the law of Nature saying that any two somehow interconnected items contain certain information of each other; and this information may also be used purposefully.

Thus a problem emerges: How to combine information contained in a sample with the collateral information from other sources, in order to improve a quality of estimation? What are conditions of enhancing the estimate by making such a combination? The authors suggested [11] some *ad hock* answers to those questions - in case of estimating a linear functional of distribution. In this paper we present some results of applying these methods to estimation of a certain stochastic event B probability, $P(B)$.

1 Statement of the problem and possible solutions

Let $X_1, \ldots, X_n$ be a sample of i.i.d.r.v.'s of size n from $F(x)$, and $P = P(B)$ be a probability of an event B that may happen in this experiment. The nonparametric estimate of P is the relative frequency of B: $\hat{P} = \hat{P}(B) = n^{-1}\sum_{i=1}^{n} I_B(X_i)$; here $I_B(\cdot)$ is indicative function. Let p_j, $j = 1, \ldots, m$, be additional prior guesses of estimated probability P. (The term "prior guess" has been coined by Ferguson [5]). The problem is: how to combine the prior guesses with $\hat{P}$ to improve a quality of estimation?

One possibility is to set up a linear combination of all of them, at least two different approaches can be used:

$$\hat{P}_1 = \hat{P} - \sum_{j=1}^{m} \lambda_j(\hat{P} - p_j), \qquad or \qquad \hat{P}_2 = \hat{P} - \tfrac{1}{m}\sum_{j=1}^{m} \lambda_j(\hat{P} - p_j),$$

where every weight λ_j is defined separately by minimizing MSE (mean square error) $M[\hat{P} - \lambda_j(\hat{P} - p_j) - P]^2$, and is equal to $\lambda_j = (1 + n\Delta_j^2/\sigma^2)^{-1}$, where $\sigma^2 = P(1-P)$, and $\Delta_j = P - p_j$ is a deviation of a guess p_j from real value of P. Coefficients λ_j may be referred to as optimal for the case of utilizing the single prior guess p_j, but they are not optimal in the sense of $MSE(\hat{P}_q)$, $q = 1, 2$, so we will refer to estimates $\hat{P}_q$ as quasi-optimal estimates.

However, optimal coefficients λ_j contain unknown value of P. The use of their estimates is suggesting themselves for obtaining an *adaptive* estimate of P. Let us consider two options of estimating λ_j's. The first option is to use $\hat{\lambda}_{j,1} = (1+n\hat{\Delta}_j^2/\hat{\sigma}^2)^{-1}$, where $\hat{\Delta}_j = \hat{P} - p_j$, and $\hat{\sigma}^2 = \hat{P}(1-\hat{P})$. The second option is to set $\hat{\lambda}_{j,2} = (1 + n\hat{\Delta}_j^2/\sigma_j^2)^{-1}$, where $\sigma_j^2 = p_j(1-p_j)$. Then we obtain four adaptive estimates

$$\hat{P}_{ql} = \hat{P} - \sum_{j=1}^{m} \hat{\theta}_{jql}(\hat{P} - p_j),$$

where $q, l = 1, 2$, $\hat{\theta}_{j1l} = \hat{\lambda}_{j,l}$ and $\hat{\theta}_{j2l} = \frac{1}{m}\hat{\lambda}_{j,l}$.

All the introduced estimates are biased, and the question arises: when the adaptive estimates are better than usual one $\hat{P}$ (in the sense of MSE, $S_{ql}^2 = M[\hat{P}_{ql} - P]^2$)? That MSEs may be calculated for different values of n, P and $p_1, \ldots, p_m$ through the binomial distribution $Bi(n, P)$ of value $n\hat{P}$: $\pi_i(P) = P(n\hat{P} = i) = C_n^i P^i (1-P)^{n-i}$, $i = 0, \ldots, n$, so that in expectations we can use $\hat{\theta}_{jql} = \hat{\theta}_{jql}(i)$ and

$$S_{ql(1)}^2 = MSE\left(\hat{P}_{ql}\right) = \sum_{i=0}^{n} \left(\frac{i}{n} - \sum_{j=1}^{m} \hat{\theta}_{jql}(i)\left(\frac{i}{n} - p_j\right) - P\right)^2 \pi_i(P). \quad (1)$$

For the quasi-optimal estimates we have

$$S^2_{q(1)} = MSE(\hat{P}_q) = \left[\left(1 - \sum\nolimits_{j=1}^{m} \theta_{jq}\right)^2 + \left(\sum\nolimits_{j=1}^{m} \theta_{jq}\Delta_j\right)^2 n/\sigma^2\right]\sigma^2/n, \qquad (2)$$

where $q = 1, 2$, $\theta_{j1} = \lambda_j$ and $\theta_{j2} = \frac{1}{m}\lambda_j$.

2 An iterative joining of guesses

In search of hopefully more effective (in MSE-sense) addition of collateral information to our basic estimate, the iterative weighted addition turned out to be very promising. The idea is: to add guesses p_j, $j = 1, \ldots, m$ to previously accumulated information sequentially, minimizing MSE at each step:

$$\hat{P}_k = \hat{P}_{k-1} - \lambda_{k,n}(\hat{P}_{k-1} - p_k) = (1 - \lambda_{k,n})\hat{P}_{k-1} + \lambda_{k,n}p_k, k = 1, \ldots, m, \qquad (3)$$

where $\hat{P}_0 = \hat{P}$, and weights $\lambda_{k,n}$ are found from the condition of minimizing MSE

$$S^2_k = M[\hat{P}_k - P]^2 = M[\hat{P}_{k-1} - \lambda_{k,n}(\hat{P}_{k-1} - p_k) - P]^2,$$

$$\lambda_{k,n} = \frac{M[(\hat{P}_{k-1} - P)(\hat{P}_{k-1} - p_k)]}{M(\hat{P}_{k-1} - p_k)^2}. \qquad (4)$$

In such a procedure, MSE at the k-th step equals to

$$S^2_k = M(\hat{P}_{k-1} - P)^2 - \frac{[M(\hat{P}_{k-1} - P)^2 + \Delta_k M(\hat{P}_{k-1} - p_k)]^2}{M(\hat{P}_{k-1} - P + \Delta_k)^2}. \qquad (5)$$

This means that $S^2 \geq S^2_1 \geq \ldots \geq S^2_m$ where $S^2 = M(\hat{P} - P)^2 = P(1 - P)/n$, i.e. each added guess may only to decrease MSE by the second term in (5). If the k-th guess is a true value ($p_k = P$), then $\Delta_k = 0$, $\lambda_{k,n} = 1$, and $S^2_k = 0$.

After pooling all available guesses, the ultimate formula for $\hat{P}_m$ may be represented (by using (3)) as a linear combination of all data, with a particular set of item's weights:

$$\hat{P}_m = \left(1 - \sum\nolimits_{j=1}^{m} \theta^{(m)}_{j,n}\right)\hat{P} + \sum\nolimits_{j=1}^{m} \theta^{(m)}_{j,n} p_j = \hat{P} - \sum\nolimits_{j=1}^{m} \theta^{(m)}_{j,n}\hat{\Delta}_j, \qquad (6)$$

where $\theta^{(m)}_{j,n} = \lambda_{j,n}\prod_{i=j+1}^{m}(1 - \lambda_{i,n}), \quad j = 1, \ldots, m-1, \quad \theta^{(m)}_{m,n} = \lambda_{m,n}$.

The corresponding adaptive version can be written as

$$\tilde{P}_m = \hat{P} - \sum\nolimits_{j=1}^{m} \hat{\theta}^{(m)}_{j,n}\hat{\Delta}_j, \qquad (7)$$

$$\hat{\theta}^{(m)}_{j,n} = \hat{\lambda}_{j,n}\prod\nolimits_{i=j+1}^{m}(1 - \hat{\lambda}_{i,n}), \quad j = 1, \ldots, m-1, \quad \hat{\theta}^{(m)}_{m,n} = \hat{\lambda}_{m,n}.$$

Let us reserve notation $S^2_{1(2)} = MSE(\hat{P}_m)$ for MSE of (6) and notation $S^2_{2(2)} = MSE(\tilde{P}_m)$ for MSE of (7).

3 Estimation based on averaging guesses

The next idea is trying to construct a linear combination of specifically weighted guesses, $p = \sum_{j=1}^{m} v_{jn} p_j$. The weights v_{jn} should be non-negative $v_{jn} \geq 0$, normalized $\sum_{j=1}^{m} v_{jn} = 1$, and should be chosen according to the principle: the less a normalized deviation $|b_{jn}|$ is ($b_{jn} = \sqrt{n}(P - p_j)/\sqrt{P(1-P)}$), the larger weight v_{jn} is it granted. Values of v_{jn} can be assigned through a specific functions $\psi_j(b_{jn})$. For example:

$$v_{jn} = \psi_{j1}(b_{jn}) = \frac{\exp\{-b_{jn}^2\}}{\sum_{j=1}^{m} \exp\{-b_{jn}^2\}} \qquad or \qquad v_{jn} = \psi_{j2}(b_{jn}) = \frac{\exp\{-|b_{jn}|\}}{\sum_{j=1}^{m} \exp\{-|b_{jn}|\}}.$$

With these matched weights, v_{jn}, the optimal (on λ_{mn}) estimate of P has a form of

$$\hat{P}_m = \hat{P} - \lambda_{mn} \left(\hat{P} - \sum\nolimits_{j=1}^{m} v_{jn} p_j \right) = \hat{P} - \lambda_{mn} \sum\nolimits_{j=1}^{m} v_{jn} \hat{\Delta}_j, \tag{8}$$

where $\lambda_{mn} = [1 + (\sum_{j=1}^{m} v_{jn} b_{jn})^2]^{-1}$.

Its adaptive counterpart may be built by substituting in (8), instead of unknown λ_{mn} and v_{jn}, their sample estimates, $\hat{\lambda}_{mn} = [1 + (\sum_{j=1}^{m} \hat{v}_{jn} \hat{b}_{jnl})^2]^{-1}$, and $\hat{v}_{jn} = \psi_j(\hat{b}_{jnl})$, where $\hat{b}_{jn1} = \sqrt{n}(\hat{P} - p_j)/\sqrt{\hat{P}(1-\hat{P})}$ or $\hat{b}_{jn2} = \sqrt{n}(P - p_j)/\sqrt{p_j(1-p_j)}$, so that

$$\tilde{P}_m = \hat{P} - \hat{\lambda}_{mn} \sum\nolimits_{j=1}^{m} \hat{v}_{jn} \hat{\Delta}_j.$$

For the later usage let us introduce notation $S^2_{q(3)} = MSE(\hat{P}_m)$ for MSE of (6), where index $q = 1, 2$ corresponds to weighting function ψ_{jq}. The notation $S^2_{ql(3)} = MSE(\tilde{P}_m)$ for MSE of (7), where index $l = 1, 2$ selects one of the relative deviation estimate $\hat{b}_{jnl}$.

4 Comparison of MSEs of the estimates

To make numeric comparison of estimates from section 1, it is convenient to use relative characteristics $E_{q(1)} = S^2_{q(1)}/S^2$ $(q = 1, 2)$ to measure enhancement for quasi-optimal estimates and $E_{ql(1)} = S^2_{ql(1)}/S^2$ $(q, l = 1, 2)$ for adaptive estimates, where $S^2 = \sigma^2/n$ is variation of the regular unbiased empirical estimate (relative frequency). Analogously, for estimates from section 2 we will use ratios $E_{1(2)} = S^2_{1(2)}/S^2$ and $E_{2(2)} = S^2_{2(2)}/S^2$. Estimates from section 3 will be described by $E_{q(3)} = S^2_{q(3)}/S^2$ $(q = 1, 2)$ and $E_{ql(3)} = S^2_{ql(3)}/S^2$ $(q, l = 1, 2)$. To simplify illustrations, let us consider only "uniform" prior guesses $p_j = (2j - 1)/2m$, $j = 1, \ldots, m$.

In this section we present selected results of numeric comparison of the above mentioned characteristics. Figures 1 through 6 present relations between MSEs from secttion 1 for various values of P, n , m.

Left plot on Fig.1 shows that the gain in MSE is growing in larger diapason of values of P. in the right plot of Fig.1 it is seen that both adaptive estimates have

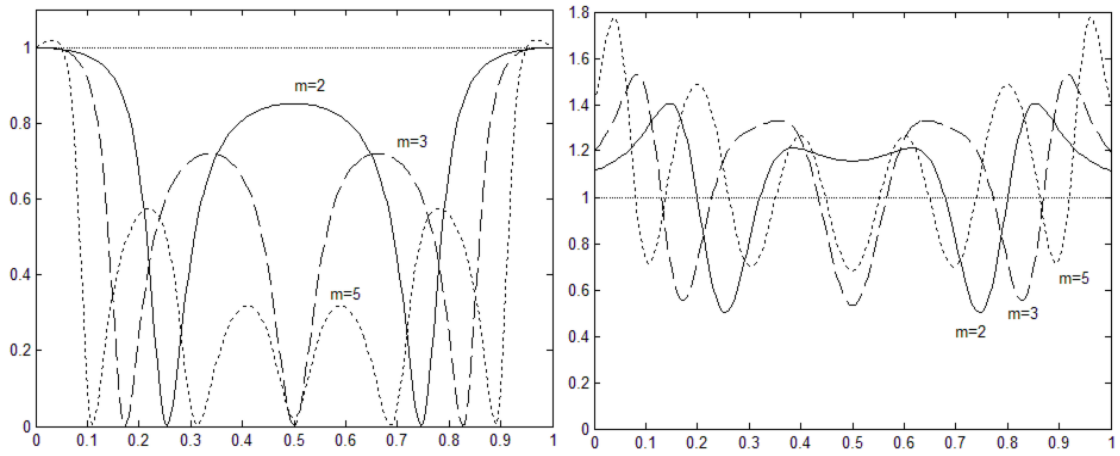


Figure 1: Dependence of $E_{1(1)}$ (left plot) and MSE $E_{11(1)}$ (right plot) on P for $m = 2, 3, 5$ under $n = 100$

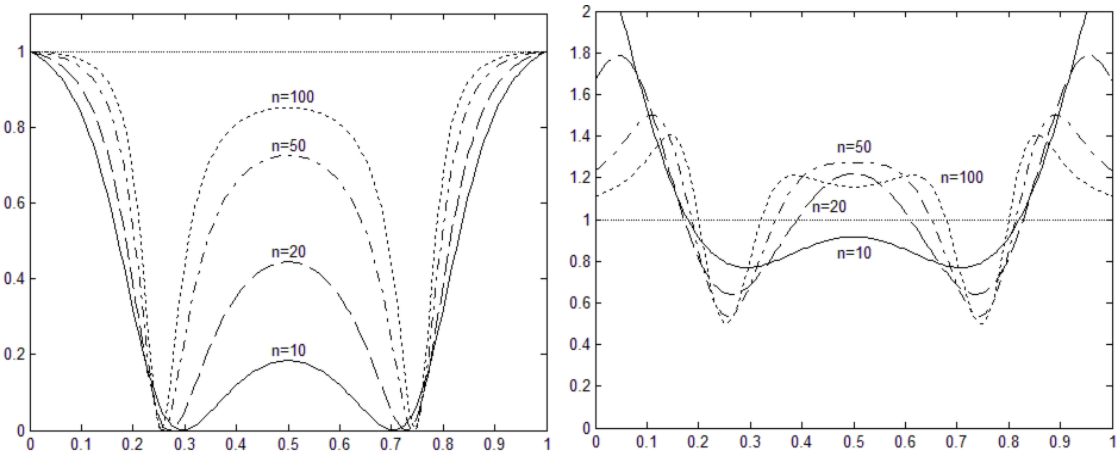


Figure 2: Dependence of $E_{1(1)}$ (left plot) and MSE $E_{11(1)}$ (right plot) on P for $n = 10, 20, 50, 100$ under $m = 2$

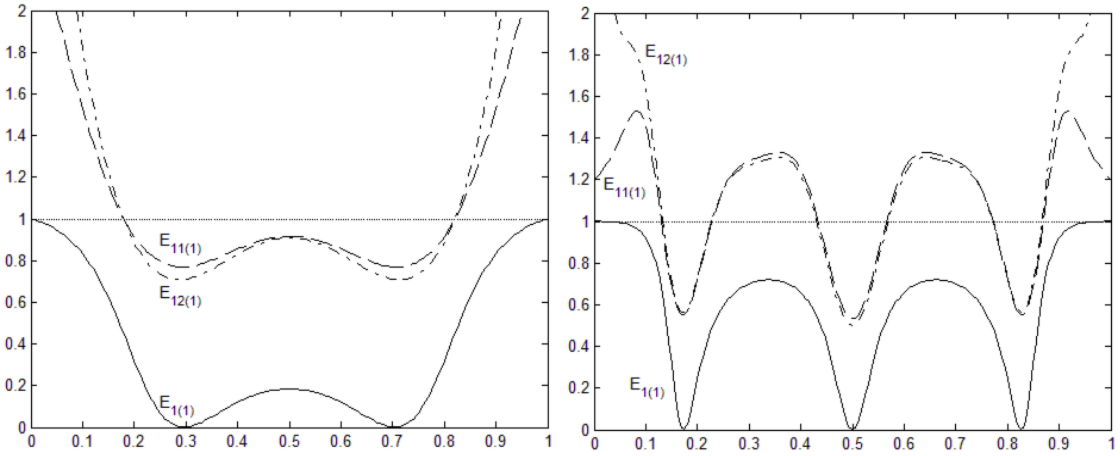


Figure 3: Dependence of $E_{1(1)}$, $E_{11(1)}$, $E_{12(1)}$ on P for $m = 2$, $n = 10$ (left plot) and for $m = 3$, $n = 100$ (right plot)

lessened MSE. Left plot on Fig.2 demonstrates that MSEs are worsened if the real value is $P = 0,5$. Right plot of Fig.2 shows that the gain in MSE for adaptive estimate is growing with larger n, but at the same time the interval of improvement is shrinking. From the left plot of Fig.3 it is seen that the second adaptive estimate is better than the first one, under small n. But under large n both adaptive estimates become to be practically equivalent (see right plot on Fig.3).

Detailed study of properties of these estimates gave the following results.

All of them *do improve* the quality (MSE) of estimate, but do that to a different extent for different number m of guesses and different relations between their values. In case of $m = 1$ they are identical (this case was considered in [10]).

If a true value of estimated parameter exists among guesses ($\Delta_j = 0$), then corresponding λ_j equals to 1, and this finalizes the estimation.

Unfortunately, those improvements require knowing of P that enters into the definition of Δ's. Trying to make the method to be applicable in practice, we suggested to modify it by substituting into algorithms instead of the $\Delta_j = P - p_j$ their estimates $\hat{\Delta}_j = \hat{P} - p_j$ This resulted in weakening improving properties of estimation by merging guesses: improvements (lessening MSE) still take place (although became slack) only in an interval around a guess, but outside this interval MSE is even worsened. It may be seen in Fig.4, presenting effects of merging one guess ($p_1 = 0.5$) ($m = 1$) into estimation of unknown probability under various sizes of a sample: level 1 line is MSE-ratio of relative frequency, lower line is for optimal, upper one – for adaptive merging of a guess). It is seen how the contribution of a guess decreases with the growing the size n of a sample.

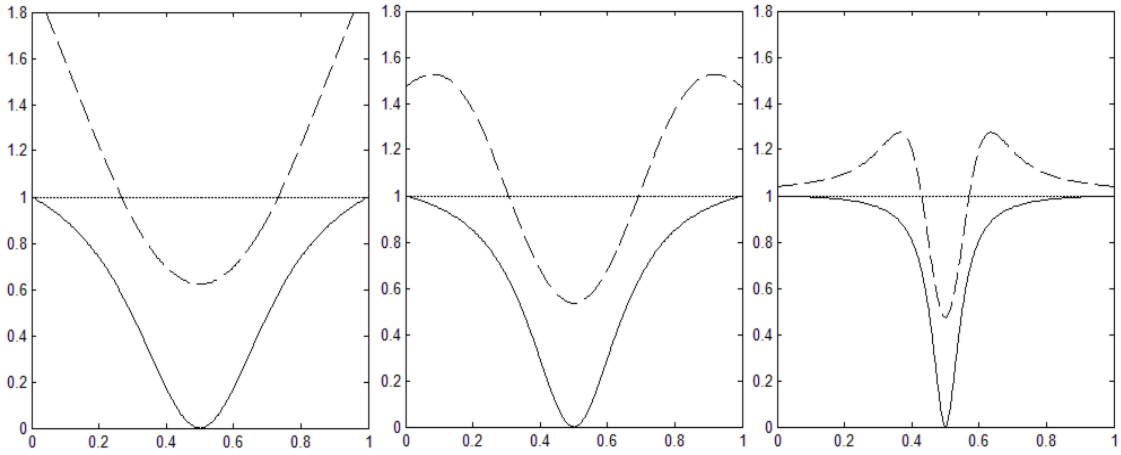


Figure 4: Dependence of $E_{1(1)}$ (optimal estimate, solid line) and $E_{11(1)}$ (adaptive estimate, dashed line) on P under $m = 1$ ($p_1 = 0.5$), for $n = 5, 10, 100$ (left to right)

But under a larger number of guesses the behavior of MSE becomes to be even more quaint [6]. For instance, already in case of two guesses, the all estimates behave quite differently (see Fig.5).

Note (it will be important in further studies of the phenomenon of collecting of dissimilar information) that the "best" behavior is demonstrated by the optimal joining *the sum* of guesses ($E_{11(1)}$) and usage of average guesses ($E_{11(3)}$). The fall of

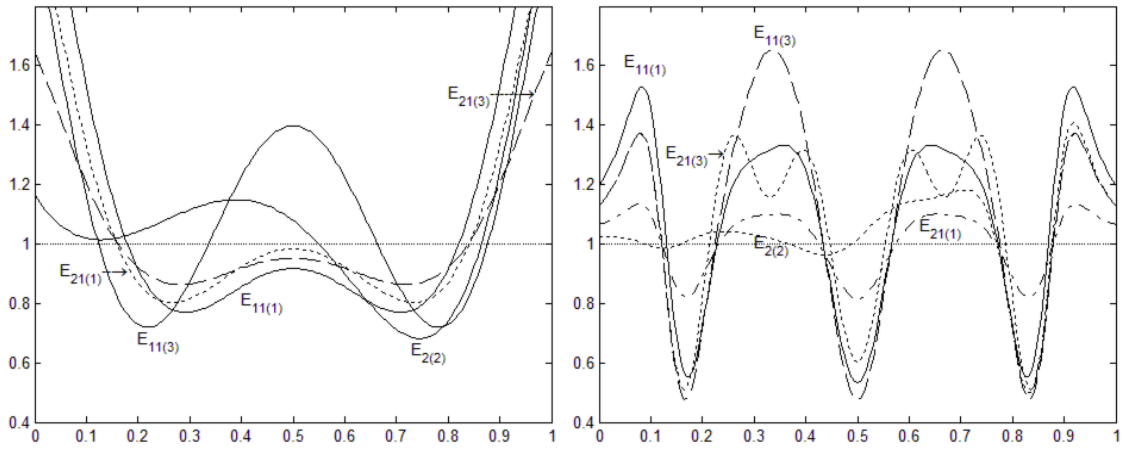


Figure 5: Dependence of $E_{11(1)}$, $E_{21(1)}$, $E_{2(2)}$, $E_{11(3)}$, $E_{21(3)}$ on P for $m = 2$, $n = 10$ (left plot) and for $m = 3$, $n = 100$ (right plot)

significance of guesses with the growth of a sample size n is also evident.

Conclusion

The idea of combining several prior guesses $(p_i, i = 1, \ldots, m)$ of the evaluated probability (P) with its statistical estimate $(\hat{P} = \hat{P}(B) = n^{-1} \sum_{i=1}^{n} I_B(X_i))$ with intention of improving quality (in minimal MSE sense) of ultimate estimate, has stumbled upon the existence of many various ways of pooling the inhomogeneous information. Comparative consideration of several (linear) combinations gave the following conclusions:

The optimal (in minimum MSE sense) *unique* linear combination of main and additional information is possible only for $m = 1$, and the optimal coefficient is $\lambda_i = (1 + n\Delta_i^2/\sigma^2)^{-1}$.

Then several instantaneous and one sequential (iterative) additions of side information (prior guesses) were studied on their qualities.

All combinations give (certain but different) improvements to qualities of estimation, but all of them demand knowing exact value of estimated (unknown) probability (for defining Δ_is and σ).

Then "adaptive" estimates of a probability were considered, where instead of unknown parameters Δ_i and σ their estimates were substituted. The result strongly depends on certain combination of the involved factors (n, m, Δ_i, σ) and type of chosen formula, but the common features are following: a) adaptive joining guesses to statistical estimate make sense only in a narrow region of their exactness; joining a guess from outside this region only worsens MSE of the estimation; b) naturally, importance of joining guesses rises with their exactness, and lessens with exactness of basic estimate (i.e. with growing n).

It still is actual to consider other (non-linear?) combinations of guesses in the problem of estimation of parameters, including the estimation of a probability.

References

[1] Mises R.V. (1964). *Mathematical Theory of Probability and Statistics.* NY, Acad. Press.

[2] Dmitriev Yu.G., Tarasenko F.P. (1974). On Statistical Estimation of Probability Density Functionals. *Theory of Probability and Its Applications.* Vol. **XIX**, **2**, pp. 104 – 109 (In Russian).

[3] Dmitriev Yu.G., Koshkin G.M., Shulenin V.P., Simakhin V.A., Tarasenko F.P. (1974) *Nonparametric Estimation of Functionals by Stationary Samples.* Tomsk, TSU Publ. (In Russian).

[4] Tarasenko F.P. (1976) *Nonparametric Statistics.* Tomsk, TSU Publ. (In Russian)

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

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

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

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

[9] Dmitriev Yu.G., Tarassenko P.F. (2015) On Adaptive Estimation Using a Prior Guess. *Proceedings the International Workshop, Applied Methods of Statistical Analysis. Nonparametric Approach.* Novosibirsk, Russia. Novosibirsk, 14-15 September, 2015. NSTU publisher, pp. 49–55.

[10] Dmitriev Yu.G., Koshevaya T.O. (2015) Combined Estimators of Probability. *Izvestia vuzov, Physics.* Vol. **58**. **11/2**, pp, 242–248. (In Russian)

[11] Dmitriev Yu.G., Tarasenko F.P., Tarasenko P.F. (2016) On Making Use of Presumed Values of a Linear Functional in its Statistical Estimation *Third Conference of the International Society for Nonparametric Statistics (ISNPS), Avignon, Palace of the Popes, 11-16 June 2016. Book of Abstracts.* p. 48.

Oil Pipeline Pressure Measurements Forecasting and Correction

Agafonov E.
Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
e-mail: agafonov@gmx.de

Antropov N.
Siberian Federal University, Krasnoyarsk, Russia
e-mail: nikita.antropov.92@mail.ru

Abstract

The article is devoted to adaptive modeling of oil pipeline pressure distribution. Hydraulic head parametric description together with nonparametric residuals correction model form a hybrid model of pressure distribution. One proposes and tests model variants for oil pipeline technological segment.

Keywords: adaptive modeling approach, hybrid model, hydraulic head, error correction, pressure sensor.

Introduction

Measurement error is considered to be the main parameter of any measurement device or sensor. This is a random value describing residual between true value and a measured one. Measuring equipment manufacturers usually mention error characteristics in operating instructions or technical passports, splitting it into *random error* and *systematic error*.

We are sure that real measurement errors rarely correspond to those in supplied documentation. Indeed, errors can demonstrate nonstationary behavior, i.e. they constantly vary during operation. Some possible reasons of such drift are changing of operation conditions (factors of surroundings); drifting physical parameters of sensor (internal factors); disturbances in communication channels; partial or total breakdown of sensor.

Taking these factors in consideration one should be faced with the problem of *measurements correction*. Consequently we suppose that correction algorithms development is an important step of the problem solution. It is evident to use *adaptive approach* to deal with uncertainties and nonstationary nature of measured data.

In scientific articles and reference literature [1, 2, 3] various approaches for correction are proposed. Some of them assert that sensor should be constructively changed and optimized (sensing element, housing, or topology improvement). Other suggest inclusion of additional passive or active correction elements as parts of the sensor. Alternative approach is soft correction using mathematical modeling (concept of intelligent sensor). It is a concept that we basically utilize in this research.

Object of our interest is an *oil trunk pipeline*. Modern pipeline systems are supplied with numerous kinds of measurement and communication equipment [4]. The

main purpose of the equipment is monitoring and control of technological parameters. One of the most important parameters demanding constant monitoring is *pressure of oil*. Pressure in various components of pipeline, such as in its linear part, have to be traced and analyzed to provide secure process of oil transport.

Pressure sensor is a measuring device with sensing elements and measuring transducers assembled in a common housing. Applying pressure to the sensor leads to changes of its physical state, and, consequently, the output electric signal. Generally, the signal is nonlinear function from pressure, and it strongly depends on other parameters like temperature. This nonlinearity is usually eliminated by special electronic correction circuits [1].

Figure 1 depicts common representation of pressure sensor as a «black box» converter.

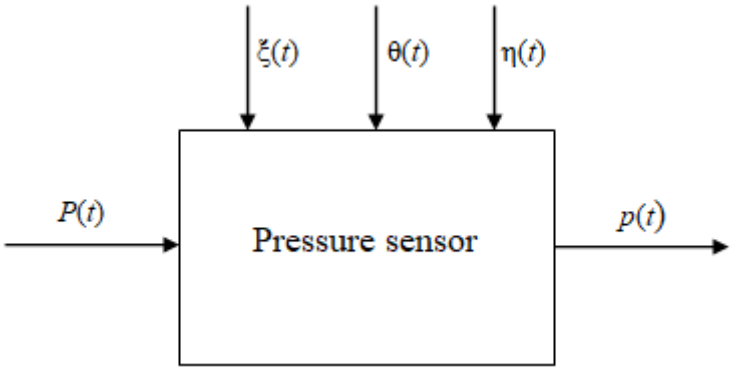


Figure 1: Pressure sensor as an input/output effect converter

In Figure 1 the following notations are used: $P(t)$, $p(t)$ are namely input (pressure), and output (electric signal) variables of the sensor; $\varepsilon(t)$, $\theta(t)$, $\eta(t)$ are random additive, multiplicative, and systematic measurement errors; $\varepsilon(t)$, $\theta(t)$ are assumed to be unbiased with limited standard deviation, and $\eta(t)$ with limited bias and deviation. The equality establishes the relationship between above introduced variables:

$$p(t) = A(P(t)) \cdot \theta(t) + \varepsilon(t) + \eta(t), \tag{1}$$

for unknown operator A.

1 Problem statement: pressure sensor output correction

Linear part of oil pipeline is equipped by set of pressure sensors being the part of monitoring and automation control system. Distance between neighbor sensors normally doesn't exceed 30 km. Variety of sensors' types, brands, and their technical condition leads to scattered random and systematic errors. Having information on physical patterns of pressure distribution along the pipeline, on mutual allocation

of sensors, and long-term measured data, one should develop *pressure distribution model* for certain technological modes with an oil pipeline segment.

2 Principle of pressure distribution in oil pipeline

According to Bernoulli Equation [5], *hydraulic head* along the constant cross-section pipeline without gravity-driven flows for steady-state case is linearly dependent from cross-section coordinate (Figure 2).

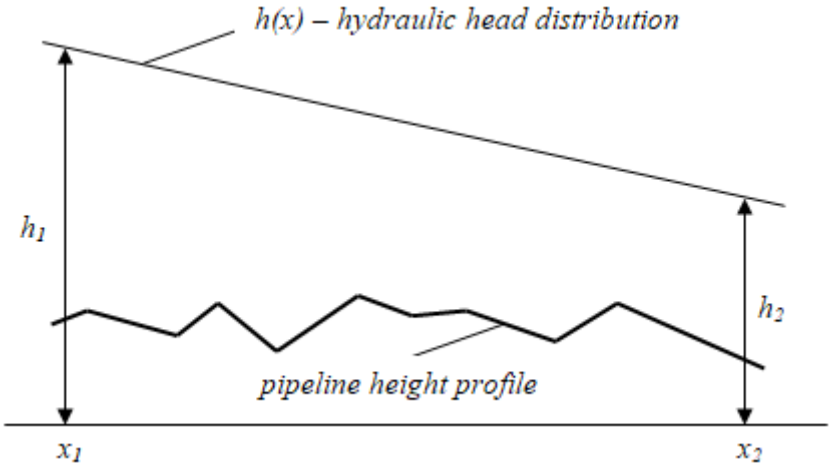


Figure 2: Geometric interpretation of Bernoulli Law for linear pipeline segment

The following notations are used in the figure: h_1, h_2 are hydraulic heads for the initial (x_1) and the final (x_2) cross-sections of the pipeline segment; z_1, z_2 are altitudes of the corresponding extreme cross-sections. The equation gives relation between hydraulic head and pressure:

$$h = \frac{\alpha_k v^2}{2g} + \frac{p}{\rho g} + z, \tag{2}$$

where $\frac{\alpha_k v^2}{2g}$ is high head [m], $\frac{p}{\rho g}$ is piezometric head $[m]$, z is geometric head $[m]$, v is velocity of oil flow $[m/s]$, p is pressure $[Pa]$, ρ is oil density $[kg/m^3]$, and g is acceleration of gravity force $[m/s^2]$.

A line between h_1 and h_2 upper points is called *hydraulic slope line*. Physical meaning of the slope is hydraulic head loss due to friction per unit length of the pipeline [5].

Linearity of hydraulic head distribution corresponds to the ideal stationary flow. It is valid for the case of constant cross-section along the pipeline segment as well as unaltered physical properties of oil. In reality this distribution can have nonlinear form. Local hydraulic resistances, natural drift of oil properties and other factors provoke shifts, kinks and alter slope of the hydraulic head distribution. Nevertheless, proximity of the distribution to linear form has to be used as *a priori information* while

developing model. Moreover, *additional information* in measured pressure samples should be involved while modeling.

Pressure forecasts for oil pipe segment are obtained using *identification* problem solution [6]. Identification approach is versatile, it also repeatedly applicable to oil pipeline problems, as an example for pipeline network modeling [7].

One often reduces identification problem to optimal parameters selection for preselected model structure, using so-called *parametric identification approach*. Correctness of the model structure in this case is the core premise of successful identification. Parametric models [8] demand extensive a priori information, what is rarely found in real-life problems. Thus one implements *nonparametric* or *hybrid models* [9, 10] in order to solve identification problem filling the lack of information about the process, i.e. to combine assumptions of a structure, particular measurement properties, and information contained in the measurements.

3 Hybrid modeling procedure

Let the data set D_s: $V = \{H[x_i], x_i\}, i = 1, ...k$ represent observations of object's state. Information D_p: $\hat{A}(x, \alpha)$ about transform operator $A(x)$ up to parameters α (linear model structure in our case) is also available.

Take the notation $q_\Sigma(x) = A(x) - \hat{A}(x)$ or in the other form: $q_\Pi = \frac{A(x)}{\hat{A}(x)}$. Then *hybrid model* is given by the expression

$$H_\Sigma(x) = \hat{A}(x, \alpha) + \hat{q}(x) \tag{3}$$

for additive error, or

$$H_\Pi(x) = \hat{A}(x, \alpha) \cdot \hat{q}(x) \tag{4}$$

in case of multiplicative measurement error.

Transform $\hat{q}(x)$ is substituted by the nonparametric regression estimate [9]:

$$\hat{q}(x) = \frac{\sum\limits_{i=1}^{k} q_m[x_i] \Phi\left(\frac{x - x_i}{c_n}\right)}{\sum\limits_{i=1}^{k} \Phi\left(\frac{x - x_i}{c_n}\right)}, \tag{5}$$

where $q_m[x_i]$, $m \in \{\Sigma, \Pi\}$ are *discrepancies* dependent on data set D_s and represented as follows:

$$q_\Sigma[x_i] = H[x_i] - \hat{A}[x_i, \alpha], i = 1, ...k. \tag{6}$$

$$q_\Pi[x_i] = \frac{H[x_i]}{\hat{A}[x_i, \alpha]}, i = 1, ...k. \tag{7}$$

Kernel function $\Pi(\cdot)$ and its bandwidth parameter c_n satisfy the conditions of convergence [8].

Variety of discrepancy function evaluation methods raises the problem of uncertainty in selection of an optimal hybrid model. It is evident that additive errors can

be efficiently compensated by the additive hybrid model, and multiplicative errors demand usage of the correspondingly organized hybrid model [9]. Lack of information about error type can be the reason of *ensemble modeling approach* implementation [9, 10, 11]. Below we suggest one example of such ensemble construction [9]:

$$H(x) = \lambda \cdot H_{\Pi}(x) + (1 - \lambda) \cdot H_{\Sigma}(x), \lambda \in [0, 1], \tag{8}$$

where $H_{\Pi}(x)$ and $H_{\Sigma}(x)$ are hybrid models for multiplicative and additive errors respectively; λ is weight of the model $H_{\Pi}(x)$ in ensemble. Ensemble modeling approach helps to effectively deal with low background (a priori) information and eliminate uncertainty of correct model choice.

Models (3), (4), and (8) satisfy statistical convergence conditions, i.e. they are asymptotically unbiased and converge in the mean square [9].

4 Initial data and modeling procedure

Initial date for pipeline pressure distribution model are

- linear parametric structure of the model with respect to hydraulic head $\{D_p : h(x) = ax + b\}$,
- data set of hydraulic head obtained from pressure measurements$D_s : V = \{p_i[t]\}$, using expression (2).

Here index i indicates sensor number $i = 1, ...k$, index t defines discreet time instants when measurements were performed, $t = 1, ...n$. During this research modeling of one oil pipeline in West Siberia has been performed. Length of technological segment was L = 907,8 km. Number of linear sections divided by pump stations was $T = 5$. Amount of pressure sensors was k= 87, number of measurements in data set was n = 1440.

Preliminary stage of modeling involved data preparation. During this stage censoring of the data set was performed, i.e. outliers were detected and removed from the data set. A lot of articles describe data censoring problem and ways how to solve it [12]. We used the following procedure of censoring: calculate mean value m_i and standard deviation σ_i for every$i$-th pressure sensor:

$$m_i = \frac{1}{n} \sum_{i=1}^{n} p_i[t], i = 1, ...k, \tag{9}$$

$$\sigma_i = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i[t] - m_i)^2}, i = 1, ...k, \tag{10}$$

Then remove from the data set measurements with standard deviation greater then 3σ, according to recommendations of mathematical statistics [13]. After that

perform procedure of data set D_s averaging for every position x_i of pressure sensor:

$$p[x_i] = \frac{1}{n}\sum_{i=1}^{n} p_i[t], i = 1, ...k, \tag{11}$$

Resulting pressure data set $D_s : V = p[x_i], x_i, i = 1, ...k$ is transferred to hydraulic head data set using expression (2) for oil with density $\rho = 970\ kg/m^3$ and average flow velocity. Note that oil velocity defines offset of the data and it has no influence on mutual arrangement of sample elements. New data set of hydraulic heads is $D_s : V = h[x_i], x_i, i = 1, ...k$.

All three models (3), (4), and (8) were applied to the problem solution. Models (3) and (4) were implemented using the following procedure [14]:

1. Parametric identification $\alpha = (a, b)$ of the linear model $h(x) = ax + b$ using least squares based on data set D_s;
2. Discrepancies evaluation $q[x]$ according formulas (6), (7);
3. Nonparametric estimation of the discrepancy function $\hat{q}(x)$ using expression (5);
4. Summation (for the model (3)) or multiplication (for the model (4)) of the corresponding parametric and nonparametric models;
5. If needed transformation of hydraulic head modeling results to pressure values.

The procedure should be repeated of every linear segment of oil pipeline. Model (8) is formed as combination of models (3) and (4).

Model (8) demands optimal parameter λ. The *best probe strategy* is suggested to perform the optimization. Certain range of λ is evenly splitted producing a uniform grid. The best λ one can find using LSE criterion searching all the nodes of the grid. Tuning of bandwidth cn in nonparametric estimates (5) one performs cross-validation procedure. Optimization criteria with respect to λ and c_n are

$$\lambda = \frac{1}{k}\sum_{i=1}^{k} \sqrt{\left(h[x_i] - \lambda h_{i,\Pi}[x_i] - (1-\lambda)\, h_{i,\Sigma}[x_i]\right)^2} \to \min_{\lambda} \tag{12}$$

$$c_n = \frac{1}{k}\sum_{i=1}^{k} \sqrt{\left(\hat{q}[x_i, c_n] - q[x_i]\right)^2} \to \min_{c_n} \tag{13}$$

Resulting models (2), (3) as well as ensemble (7) are expressed as follows:

$$h_{j,\Sigma}(x) = a_j x + b_j + \hat{q}_j(x), j = 1, ...T, \tag{14}$$

$$h_{j,\Pi}(x) = (a_j x + b_j) \cdot \hat{q}_j(x), j = 1, ...T, \tag{15}$$

$$h_{j,\lambda}(x) = \lambda \cdot h_{j,\Pi}(x) + (1-\lambda) \cdot h_{j,\Sigma}(x), j = 1, ...T, \tag{16}$$

where $h(x)$ is hydraulic head for point of the pipeline with position x, $a_j = (a_j, b_j)$, $j = 1, ...T$ are j-th linear segment parameters, $\hat{q}_j(x), j = 1, ...T$ nonparametric discrepancy function estimates.

5 Modeling results

The identification problem was repeatedly solved using additive, multiplicative, and ensemble models. To compare results one implemented LSE validation criterion

$$W = \frac{1}{k}\sum_{i=1}^{k}\sqrt{(h[x_i] - h_{i,m}[x_i])^2}, \tag{17}$$

where k is sample size, $h[x_i], i = 1, ...k$ are elements of training sample, $h_{i,m}[x_i]$ are predicted by a model hydraulic heads.

Estimated parameters and optimal LSE values for different model configurations are given in the table 1.

Table 1: Estimated parameters and optimal LSE values

	a_i	b_i	λ_i	W
Nonparametric regression (I)	—	—	—	6.823e-04
Parametric linear model (II)	$a_1 = -0.1741$ $a_2 = -0.1819$ $a_3 = -0.1821$ $a_4 = -0.1721$ $a_5 = -0.1864$	$b_1 = 45.40$ $b_2 = 68.33$ $b_3 = 99.71$ $b_4 = 135.06$ $b_5 = 176.06$	—	19.8128
Hybrid model with with additive discrepancy (III)	$a_1 = -0.1741$ $a_2 = -0.1819$ $a_3 = -0.1821$ $a_4 = -0.1721$ $a_5 = -0.1864$	$b_1 = 45.40$ $b_2 = 68.33$ $b_3 = 99.71$ $b_4 = 135.06$ $b_5 = 176.06$	—	1.486e-04
Hybrid model with with multiplicative discrepancy (IV)	$a_1 = -0.1741$ $a_2 = -0.1819$ $a_3 = -0.1821$ $a_4 = -0.1721$ $a_5 = -0.1864$	$b_1 = 45.40$ $b_2 = 68.33$ $b_3 = 99.71$ $b_4 = 135.06$ $b_5 = 176.06$	—	7.859e-04
Ensemble of models III and IV (V)	$a_1 = -0.1741$ $a_2 = -0.1819$ $a_3 = -0.1821$ $a_4 = -0.1721$ $a_5 = -0.1864$	$b_1 = 45.40$ $b_2 = 68.33$ $b_3 = 99.71$ $b_4 = 135.06$ $b_5 = 176.06$	$\lambda_1 = 6.76e-05$ $\lambda_2 = 2.55e-06$ $\lambda_3 = 5.23e-05$ $\lambda_4 = 2.72e-06$ $\lambda_5 = 3.87e-06$	1.486e-04

Judging on the achieved accuracy the most effective approach for hydraulic head model bacame hybrid model with additive discrepancy (III). Least square error of the model (III) is substantially better than for the nonparametric model (I), and by several orders of magnitude better than for linear regression model (II), that is widely used by oil transport companies. The results approve outstanding features of hybrid

modeling approach. Such models do the best utilizing all the available information about pressure (hydraulic head) distribution in oil pipeline.

Model (III) together with initial data is depicted in figure 3. Points correspond to hydraulic heads transferred from the measured pressure. Line represent model of hydraulic head distribution.

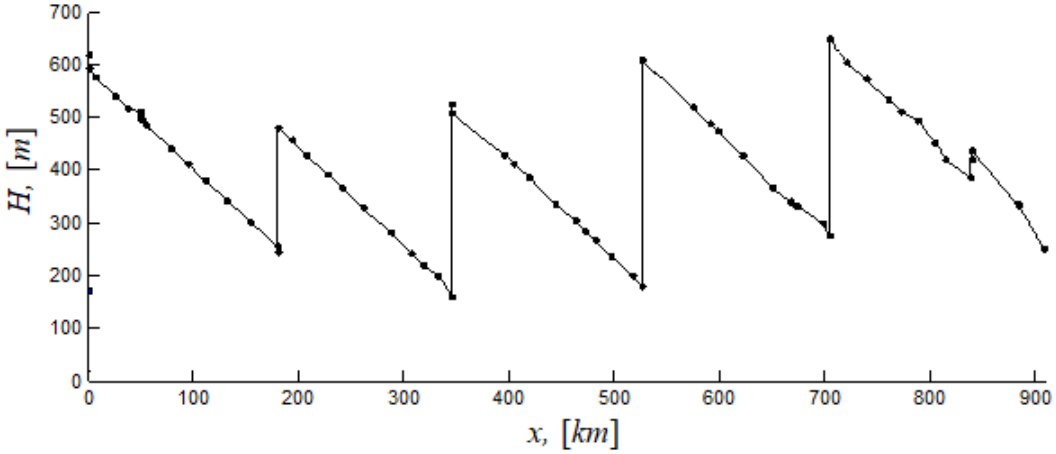


Figure 3: Hybrid model with additive discrepancies (III)

Optimal values of parameter λ for model (V) are close to zero. This fact indicates low efficiency of model (IV) and, as the result, of ensemble models applied to the problem under consideration.

Multiplicative discrepancy hybrid model (IV) demonstrated low efficiency because true errors have additive nature. Another reason is preliminary treatment of data with getting rid of outliers. Being more robust to outliers the multiplicative discrepancy model has here no chance to prove its superiority with respect to the additive one.

Conclusion

Modern trunk oil pipelines' equipment includes diversity of automation and control means including SCADA systems for oil transport process monitoring and control. The most important technological parameter to be continuously tracked is pressure and, of course, hydraulic head that is closely related to pressure. Effectiveness and safety of oil transport process relies upon the quality of pressure monitoring. Thus pressure sensors play the great role and truly indispensable with oil transport systems.

This research is devoted to pressure forecasting problem, and corresponding modeling principles that effectively help to correct measurement errors of pressure sensors; to develop forecasts of pressure for broken or malfunctioning sensors; to extrapolate dependence of pressure or hydraulic head from position of the pipeline cross-section.

References

[1] Zemelman M.A. (1952).*Automatic Error Correction for Sensors.* Publishing House Standards, 199 p.

[2] Bromberg E.M. (1978). *Testing Methods of Measurement Accuracy Improvement.* Moscow: Energy, 176 p.

[3] Gelman M.M. (1974). *Automatic Correction of Systematic Measurement Errors in «Voltage – Code» Transducers.* Moscow: Energy, 88 p.

[4] Zemenkov Yu.D., Shalay V.V., Zemenkova M.Yu. (2015). Expert systems of multivariable predictive control of oil and gas facilities reliability. *Procedia Engineering.* Vol. 113, pp. 312-315.

[5] Lourier M.V. (2003). *Mathematical Modeling of Processes in Oil, Fuels, and Gas Pipelines.* Moscow: Oil and Gas, 335 p.

[6] Eickhoff P. (1975). *Fundamentals of control systems identification.* Moscow: Mir, 683 p.

[7] Agafonov E.D., Antropov N.R. (2014).On Estimation of Kichhoff Equations Solution for Hydraulic Net Modeling. *News of Tula State University.* Vol. 4, pp. 110-117.

[8] Medvedev A.V. (2015). *Fundamentals of the adaptive systems theory: Monograph.* Krasnoyarsk: SibSAU, 526 p.

[9] Lapko A.V., Chentsov S.V., Krohov S. I. , Feldman L.A. (1996). *Training Systems of Data Analysis and Decision Making.* Novosibirsk: Science, 296 p.

[10] Yaping X. , Pengfei C. (2015). The Application of a Presented Hybrid BFGS-Based Method for Data Analysis in Automation System. *Proceedings - 2015 International Conference on Computational Intelligence and Communication Networks, CICN.* Pp. 973-976.

[11] Kirik E.S. (2007). On Iteration Method of Data Censoring in Regression Analysis *Avtomatika i Telemekhanika.* Vol. 4. Pp. 79-91.

[12] Gmurman V.E. (1972). *Probability Theory and Mathematical Statistics.* Moscow: High School, 368 p.

[13] Antropov N.R. , Agafonov E.D. (2015). Adaptive Models of Pressure Sensors Data in Trunk Oil Pipeline *Reshetnev Readings: XIX International Scientific and Practical Conference, SibSAU, Krasnoyarsk.* Vol. 2. Pp. 8-10.

[14] Agafonov E.D. , Antropov N.R. (2016). Pressure Sensors Correction Algorithm for the Linear Segment of Trunk Oil Pipeline. *Testing. Diagnostics.* Vol. 7. Pp. 43-48.

The Fractional Dimension's Processes

E.D.Mihov
Siberian Federal University, Krasnoyarsk, Russia
e-mail: edmihov@mail.ru

Abstract

The problem of inertia-free object's modeling is investigated. A special kind of processes occurring in a space of fractional dimension is considered. It is described that such processes can be not only fractional, but also changing.

Keywords: inertia-free object, identification, fractional dimension.

Introduction

Identification of stochastic objects is often reduced to the identification of static systems with delay. This is due to the fact that some output variables of the object controlled by much longer intervals than the input. For example, several variables measured electrically (in this case, discrete control can be brief), but the other variables are controlled by chemical analysis or physico-mechanical tests (in this case the discrete control ΔT is long, i.e. $\Delta T \gg \Delta t$).

The most common scheme is investigated discrete-continuous process is shown in figure 1:

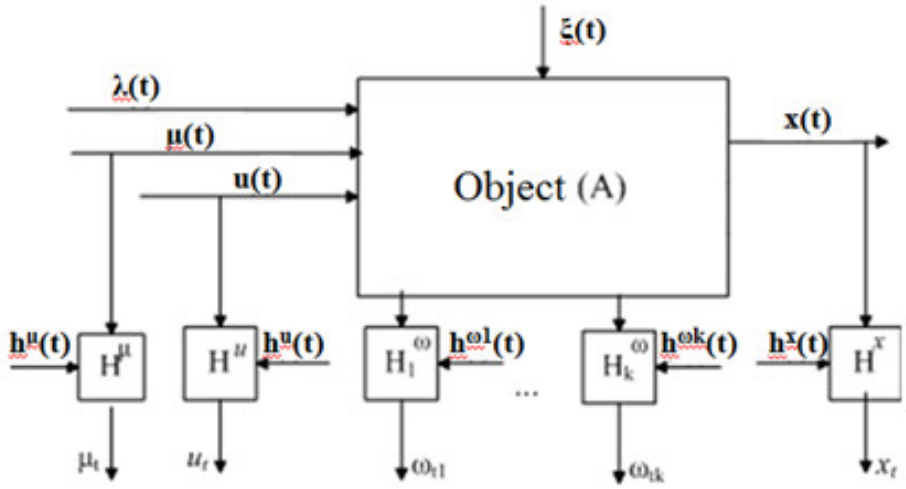


Figure 1: The general scheme of the test process

The notation of the figure 1: A is the investigate object (the process); $\overrightarrow{x(t)}$, $\overrightarrow{q(t)}$ and $\overrightarrow{z(t)}$ are the output vectors of the process; $\overrightarrow{u(t)}$ is the vector control actions; $\overrightarrow{\mu(t)}$ is uncontrolled but the measured input vector of th process; $\overrightarrow{\lambda(t)}$ is vector input unmanaged and measured process variable; $\overrightarrow{\xi(t)}$ is casual influence; $\overrightarrow{\omega^i(t)} : i = \overline{1,k}$ are the process variables controlled in object; t is continuous time $H^\mu, H^u, H^x, H^z, H^q, H^\omega$

are the communication channels corresponding to various variables; including control devices, devices for measurement of observed variables $\mu_t, u_t, x_t, \omega_t$ is measurement $\overrightarrow{\mu(t)}, \overrightarrow{u(t)}, \overrightarrow{x(t)}, \overrightarrow{\omega(t)}$ in discrete time $h^{\mu}(t), h^{u}(t), h^{x}(t); h^{\omega_1}(t) \dots h^{\omega_k}(t)$ are casual hindrances of measurements of the corresponding process variables.

1 About one feature of model operation of "tubular" processes

Let the object be described by the equation:

$$x(u) = f(u_1, u_2, u_3) \tag{1}$$

where three-dimensional vector $\overrightarrow{u} = (u_1, u_2, u_3) \in R^3$ is is the input variable, and $x \in R^1$ is output variable. We define $\hat{x}(u) = \hat{f}(u_1, u_2, u_3, \alpha)$ and assessment parameters α using observations, $(u_i, x_i, i = \overline{1, s}), s$ is sample size. Let us analyze this example from the different points of view. Let us assume that input variables $\overrightarrow{u} = (u_1, u_2, u_3)$ are independent. In this case we can use the traditional algorithm described above. Now we assume that objective components of the vector of input variables have functionally dependence, for example,

$$u_2 = \varphi_1(u_1), u_3 = \varphi_2(u_2) = \varphi_2(\varphi_1(u_1)) \tag{2}$$

Naturally, we do not know about the existence of dependences (2). Otherwise it would be possible to substitute (2) into (1) and to obtain the dependence of x on one variable u_1:

$$x(u) = f(u_1, \varphi_1(u_1), \varphi_2(\varphi_1(u_1))) \tag{3}$$

When u_3 depends u_2 we have

$$x(u) = f(u_1, \varphi_1(u_1), u_3) \tag{4}$$

i.e. x depends on u_1, u_3. Let us emphasize again that input variables are not independent. We do not know about the existence of interrelation between input variables. Now we analyse the most interesting case directly related to the H-process 4. Let us assume that u_3 and u_2 are related to each other stochastically 2. First, if components of vector $\vec{u}$ are independent the process is described by the function of the three variables. If only two components of vector $\vec{u}$ are independent the process is described by the function of two variables. If two variables are related to each other stochastically the process is described by the function of more than two variables but less than three variables. It is possible to assume that we have fractional number of variables. Therefore, we deal with a space of fractional dimension.

Let the process is described (1).

In case of a stochastic dependence between variables $u_2(u_1), u_3(u_1)$ on the available training selections it is possible to calculate a squared error δ of the estimate $\hat{u_{2s}}(u_1), \hat{u_{3s}}(u_1)$. Here $\hat{u_{2s}}(u_1), \hat{u_{3s}}(u_1)$.

$$\delta_{21} = \sum_{i=1}^{s}(u_2 - \hat{u_{2s}}(u_1)^2)/\sigma_{u_2}^2, \delta_{31} = \sum_{i=1}^{s}(u_3 - \hat{u_{3s}}(u_1)^2)/\sigma_{u_3}^2 \tag{5}$$

Here $\hat{u_{2s}}(u_1), \hat{u_{3s}}(u_1)$ there are nonparametric estimates, δ_{ij} is a a squared error of the estimate u_i, based on u_j.

The value of stochastic communication λ between any two variables can be calculated as:

$$\lambda = 1 - \delta \tag{6}$$

In the case strong functional communication $\lambda = 1$, lack of communication corresponds to $\lambda = 0$. In the case of stochastic relation between input variables $0 < \lambda < 1$. In general case, one can interpret such process as function of many variables. For example, this function can be expressed as follows [1].

$$x = \begin{cases} f(t, u_1, \mathbf{u_2}, \mathbf{u_3}, u_4, u_5) - T_1 \\ f(t, u_1, u_2, \mathbf{u_3}, u_4, u_5) - T_2 \\ f(t, u_1, u_2, u_3, u_4, u_5) - T_3 \\ f(t, u_1, u_2, u_3, u_4, u_5, u_6) - T_4 \\ f(t, u_1, u_2, u_3, u_4, u_5, u_6) - T_5 \\ f(t, u_1, u_2, u_3, u_4, u_5, u_6) - T_6 \\ f(t, u_1, \mathbf{u_2}, u_3, u_4, \mathbf{u_5}, u_6) - T_7 \\ f(t, \mathbf{u_1}, \mathbf{u_2}, u_3, u_4, \mathbf{u_5}, \mathbf{u_6}, u_7) - T_8 \end{cases} \tag{7}$$

Variables which have strong impact on x (functional relation) are designated in dark colour ($\mathbf{u_1}$). Less dark colour (u_1) means that this variable has weaker influence on x than than ($\mathbf{u_1}$) (perhaps, strong, stochastic dependence). Variables marked as u_1 and u_1 have weaker influence on x than (u_1). $T_i, i = \overline{1,8}$, are time intervals. Role of each variable may change in real process. Given above relations show that some variables can lose their significance, some variables can restore their significance and some variables can appear for the first time such as u_6, u_7.

To treat function of many variables as a point in many-dimensional space we introduce space of fractional dimension F^λ. The dimension of F^λ can be calculated as:

$$dimF^\lambda = (n+1) - \sum_{i=1}^{n-1}\lambda_{i,i+1} \tag{8}$$

n is the dimension of a vector $\vec{u}$, $\lambda_{i,i+1}$ is the intensity of stochastic relation between u_i and u_{i+1}.

There are other ways to calculate space dimension, for example,

$$dimF^\lambda = (n+1) - \sum_{i=1}^{n-1}\lambda_{1,i} \tag{9}$$

$\lambda_{1,i}$ is the dependence of all components of a vector of u_i on one component u_1.

2 The modelling's of the fractional dimension's processes features

If the input variables have a stochastic dependence, then the invectigated proceeds in fractional dimension and it has a "tubular" structure (H-proccess).

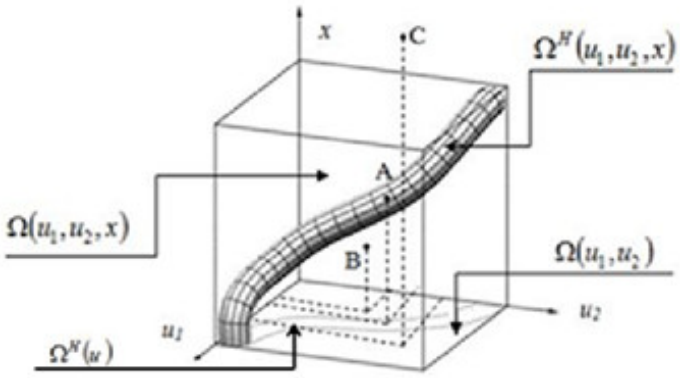


Figure 2: H-proccess

A is the point that, under the input variables $u = (u_1, u_2, ..., u_k) \in \Omega(u)$, corresponds to the output variable $\Omega^H(u_1, u_2, x)$. "B" is a point that $B \notin \Omega^H(u_1, u_2, x)$for input variables $u = (u_1, u_2, ..., u_k) \in \Omega(u)$, although in this case $B \in \Omega(u_1, u_2, X)$. "C" is a point that, for input variables $u = (u_1, u_2, ..., u_k) \in \Omega(u), C \notin \Omega(u_1, u_2, x)$.

During the construction of the parametric model for H-processes, only those elements of the training sample that belong to $\Omega^H(u, x)$ will be used. In the particular case, for a line in space, as many models as possible can be designed. This feature is shown in figure 3.

The difficulties presented in figure 3 can arise in the modeling of the H process by standard methods.

As an example, you can bring "A" (figure 4). At this point, u_1 and $u_2 \in \Omega(u)$, but, "A" $\notin \Omega(x)$. This means that for u_1 or $u_2 \notin \Omega^H(u)$, using the parametric model, we can predict physically unreal output variables [3]. In order to describe such processes, it is necessary to include the indicator function $I(u(t), \mu(t))$ in the existing model:

$$\hat{x}_s(t) = A_s(u(t), \mu(t), \alpha) I(u(t), \mu(t)) \tag{10}$$

here is:

$$I(u(t), \mu(t)) = \begin{cases} 0, & \text{when } A_s(u(t), \mu(t), \alpha) \notin \Omega^H \\ 1, & \text{when } A_s(u(t), \mu(t), \alpha) \in \Omega^H \end{cases} \tag{11}$$

Obviously, in the case when the process is not an H-process, the indicator function always has the value 1, and the model has the form. The H-model includes a standard

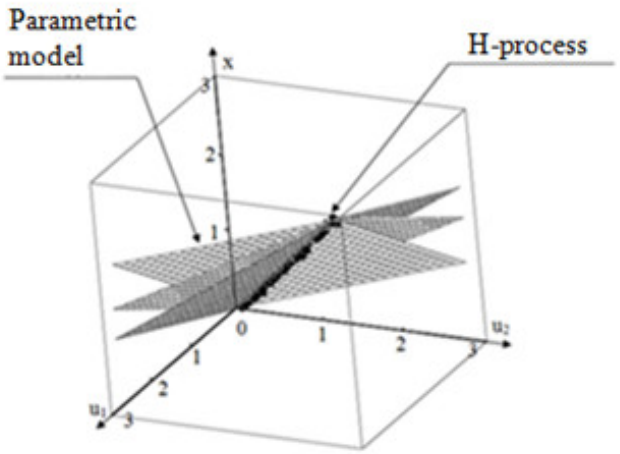


Figure 3: Parametric models of H-process

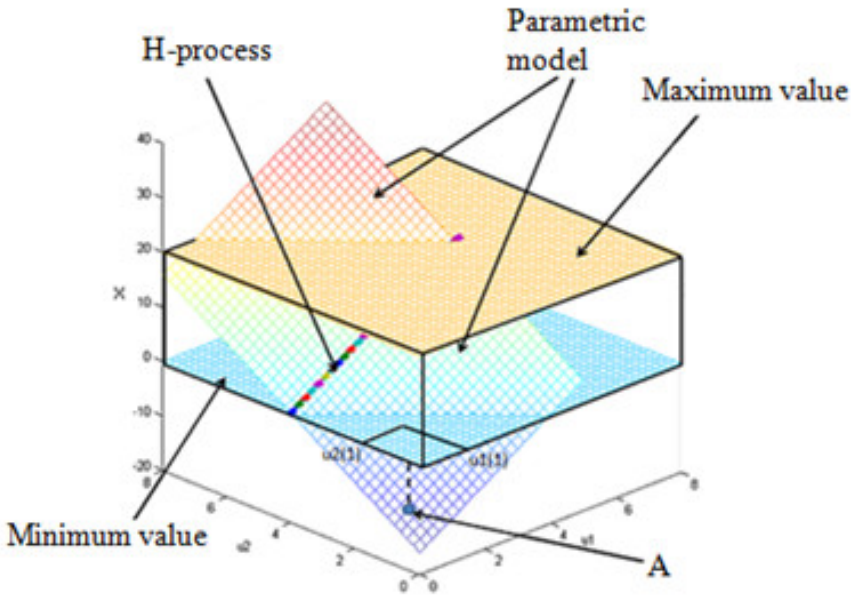


Figure 4: H-process's modeling

identification model. This means that the H-model is a more general type of models.

Conclusions

A method for modeling processes occurring in a space of fractional dimension is proposed. The problems associated with modeling these processes using standard methods are considered. It is considered that the dimensionality of the space in which the process proceeds is change.

Acknowledgements

The research had been conducted with financial support from the Russian Science Foundation grant (project N 16-19-10089).

References

[1] Medvedev A.V. [Some notes on H – models for non-inertis systems with a delay] Vestnik SibGAU. 2014. vol. 54, no. 5, p.24-34. (In Russ.)

[2] Medvedev A.V. [Analysis of the data in the problem identification] Computer analysis of simulation data. V.2. - Minsk: BSU, 1995. p. 201-206.

[3] Medvedev A.V. [H-model for non-inertia systems with delay]. Vestnik SibGAU. 2012, vol. 45, no. 5, p.84-89.(In Russ.)

[4] Mondelbrot B. Fraktalnaya geometriya prirody [Fractal Geometry of Nature]. Moscow – Izhevsk, Institute of Computer Science, NITS "Regular and Chaotic Dynamics"Publ., 2010. 656 p.

Numerical Stochastic Model of the Joint Periodically Correlated Process of Air Temperature and Relative Humidity

Olga V. Sereseva and Alisa M. Medvyatskaya
Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
e-mail: seresseva@mail.ru, medvyatskaya@mail.ru

Abstract

In this paper we study a model of joint non-stationary non-Gaussian periodically correlated time series of air temperature and relative humidity taking into account their real properties. When modeling time the daily periodicity of the parameters of the distributions is taken into account. The probabilities estimations of the various combinations of these meteorological parameters are carried out on the basis of the model.

Keywords: periodically correlated processes, non-Gaussian periodically random processes, Gaussian periodically random processes, correlation functions.

Introduction

Methods of numerical simulation of random processes and fields are used in various fields of science and technology, for example, in Meteorology, Hydrology, Oceanography etc. They are used for the purposes of forecasting, studying of anomalous phenomena, studying the properties of statistical estimates, the construction of dynamic-stochastic models and solution to other important theoretical and applied tasks. The modeling stationary processes and homogeneous spatial and spatial-time Gaussian and non-Gaussian fields are well developed by now. In connection to the modern possibilities of computational engineering in recent time, increasing attention is paid to the problems connected with the use of nonstationary processes and inhomogeneous fields. One of the important types of non-stationary processes are periodically correlated processes. In nature, these processes are quite frequent, for example, the meteorological series are experiencing daily and annual variations as well as rhythm of different time scales is observed in Oceanography. Quite a lot of papers are devoted to studying and usage of periodically correlated processes for solving various problems of statistical meteorology and Oceanography for example the works of Rozhkov V. A., Dragan Ya.P., Trapeznikov Yu.A.[1],[3], etc. Now there are a lot of approaches to modeling of periodically correlated processes. The algorithms based on the simulation of a stationary vector processes (Bokov V. N., Lopatoukhin L. I., Rozhkov V. A., Rumyantseva S. A., Derenok K. V., Ogorodnikov V.A.) [2],[3], for inhomogeneous Markov processes with periodically varying matrix of transition probabilities (Kargapolova.N.A., Ogorodnikov V. A.[4]), on the use of models Poisson point flows

are among them. An important class of models of random processes and fields is approximate spectral model (Mikhailov G. A., Prigarin S. M., Palagin Y. I., etc.). In the works Prigarin S. M. [8] and Palagin Y. I. for constructing inhomogeneous fields the parametric approach was used in which the spectral density of parametrically depends on the spatial (or spatial-time) coordinates.

1 The numerical stochastic model of the periodically correlated process

As well is known [3], the conditions of the periodic correlation of the process are the following expressions:

$$E\xi(t_i + T) = E\xi(t_i),$$

$$D\xi(t_i + T) = D\xi(t_i),$$

$$R_\xi(t_i + T,\ t_j + T) = R_\xi(t_i,\ t_j).$$

The periodically correlated Gaussian scalar random process of discrete argument at time $t_1, \ldots, t_N$, $\Delta t = t_{i+1} - t_i$, $i = 0, \ldots, N$, with the period of $T = p\Delta t$ can be represented in the form of a vector stationary process of the form:

$$\vec{\xi}^T(t_1),\ \vec{\xi}^T(t_2), \ldots,\ \vec{\xi}^T(t_N) \ldots, \tag{1}$$

or $\vec{\xi}_1^T,\ \vec{\xi}_2^T, \ldots,\ \vec{\xi}_N^T \ldots$, where $\vec{\xi}_k^T = \vec{\xi}^T(t_k) = (\xi_1(t_k), \xi_2(t_k), \ldots, \xi_p(t_k))$. As the shown in [1], [2] the periodically correlated Gaussian processes of the form (1) can be modeled using a vector autoregressive process of order n of the form:

$$\vec{\xi}_t = B_1^T[n]\vec{\xi}_{t-1} + \ldots + B_n^T[n]\vec{\xi}_{t-n} + C_n\vec{\varphi}_t, \tag{2}$$

where $B_k^T[n]$- is matrix of size p (the matrix of regression coefficients) and C_n- is lower triangular matrix $C_n C_n^T = Q_n$, where Q_n - is the conditional covariance matrix of the vector $\vec{\xi}_t$ at fixed values of the vectors $\vec{\xi}_{t-1}^T,\ \ldots,\ \vec{\xi}_{t-n}^T$.

As the starting vectors for this process Gaussian vectors are used $\vec{\xi}_1, \ldots,\ \vec{\xi}_n$, which can be represented in the form of vector $\vec{\xi}_{(n)} = (\vec{\xi}_1^T,\ \vec{\xi}_2^T, \ldots,\ \vec{\xi}_n^T)^T$ with block - Toeplitz correlation matrix:

$$R_{(n)} = \left(R_{|i-j|}\right) = (R_k)\,,\ \ i, j = 1, \ldots, n,$$

where $R_k = (r_{k,\mu\nu}), \qquad k = 0, \ldots,\ n-1,\ \mu, \nu = 1, \ldots, p$ - matrix $p \times p$.

A sequence of vectors with a covariance matrix $R_{(n)}$ can be constructed using the conditional distributions method [7]. When performing the known conditions, the process (2) is a vector stationary process and thus the process (1) is a periodically correlated one.

For modeling of non-Gausian processes $\vec{\eta}_t$ with specified one-dimensional distributions $F_i(x)$ of the components η_{it} of a vector $\vec{\eta}_t$, various modifications of the inverse distribution functions method can be used.

In this article we use a well-known modification of the method of inverse distribution functions, based on the normalization of the real data (so-called $F\Phi\Phi F$ method)[5]. The basic idea of the method is that the real data X_i in the phenomenon under consideration with the distribution functions $F_i(x)$ are normalized with the help of transformations $\xi_i = \Phi^{-1}(F_i(X_i))$, using normalized data, correlation function of the process is estimated, then Gaussian process with this correlation function is simulated and, finally, the simulated Gaussian process takes a form of the process with one-dimensional distribution $F_i(x)$ with the help of transformations $\xi_i = F^{-1}(\Phi(X_i))$. Here $\Phi(x)$ is the function of a standard normal distribution.

This method is an approximation because the correlation function is estimated on the normalized data, and Gaussian process is modeled on the basis of this function.

This work is devoted to construction of a numerical stochastic model of joint time series of surface air temperature and relative humidity using the data of real observations.

In this paper we consider the data of air temperature and relative humidity for five years (2012-2016) for March (Republic of South Africa). There are 12 measurements in the day i.e. the measurements were carried out every two hours.

Applying the modeling method of Gaussian periodically correlated processes in combination with the $F\Phi\Phi F$ method two joint scalar time series of surface air temperature and relative humidity are constructed. Normal approximation is used for temperature and distribution is used for relative humidity as a one-dimensional probability β-distributions. The parameters of the distributions (with a period $T = p\Delta t$) are periodically time-dependent and are derived from observational data. Since normal approximation was used for temperature the normalization was carried out by centering and normalization of the real data. For construct the Gaussian series in the present paper we used vector autoregression model of order $n = 5$, and the dimension of the vector $\vec{\xi}_t$ is equal to $2p$. The realization of joint time series of temperature and relative humidity were constructed and model estimates for various characteristics model series were estimated on the basis of the described model. In particular, we have estimated the accuracy of modeling. As an example, on Fig. 1,2 you can see the estimates of normalized real and simulated series of dependency of coefficients correlation $r_{TT}(t, t+h)$ series of normalized temperature at different points in time and similar dependencies of the correlation coefficients $r_{TH}(t, t+h)$ between the normalized temperature and humidity obtained from the joint block-Toeplitz correlation matrix $R_{(n)}$. The shift h is selected so that the time points were removed from each other at a time interval of 1, 2 days. On Fig.3 the same dependence $r_{TT}(t, t+h)$ for $h = 1, 5$, 2,5 days are shown. For model estimators we used 10 000 joint realizations of the series of temperature and humidity.

From the Fig.1, Fig.2, Fig.3 you can see that the real correlation structure of the normalized level is replicated in the model with accuracy up to statistical error estimates on the model samples. However, when you switch from normal data to non-Gausian data on the based transformation $\xi_i = F^{-1}(\Phi(X_i))$ of the correlation structure of the model non-Gaussian series due to the use of the approximate method $F\Phi\Phi F$ is markedly different from the real one. As can be seen on Fig.4 model curves

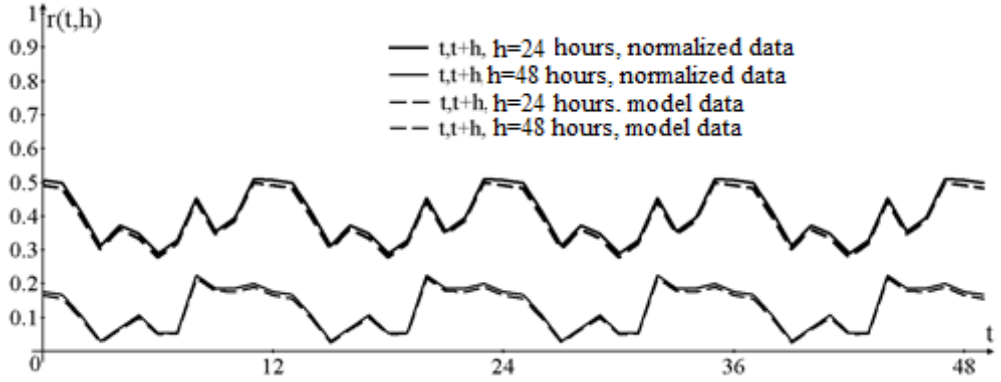


Figure 1: Correlation function $r_{TT}(t, t+h)$ time series normalized data and correlation function calculated with Gaussian model series ($h = 1$, 2 days)

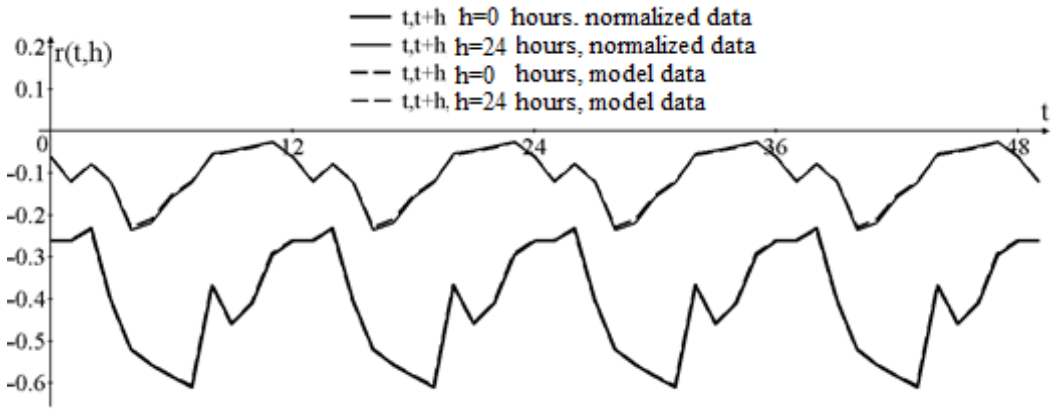


Figure 2: Correlation function $r_{TT}(t, t+h)$ time series normalized data and correlation function calculated with Gaussian model series ($h = 0$, 1 days)

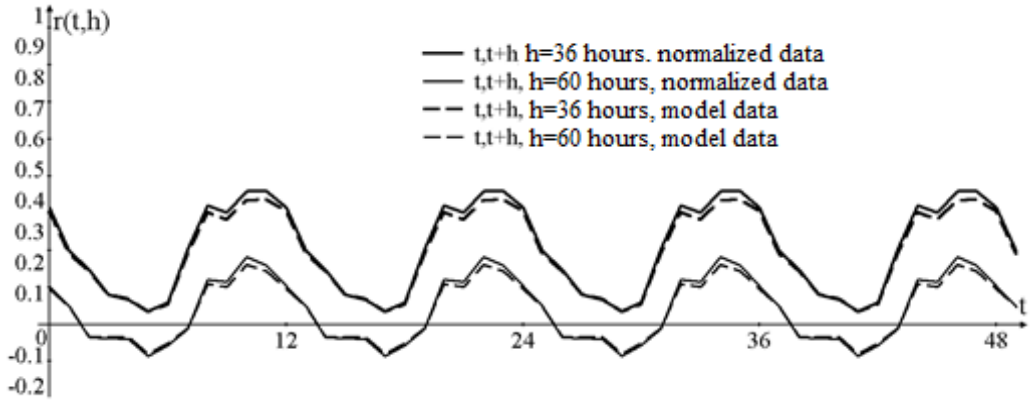


Figure 3: Correlation function $r_{TT}(t, t+h)$ time series normalized data and correlation function calculated with Gaussian model series ($h = 1,5$, 2,5 days)

differ markedly from the real, but periodicity and the character of dependence is preserved. To Refine the model further study is needed of the relations of correlation functions of Gaussian and non-Gaussian series that are implemented in the framework of the used transformations.

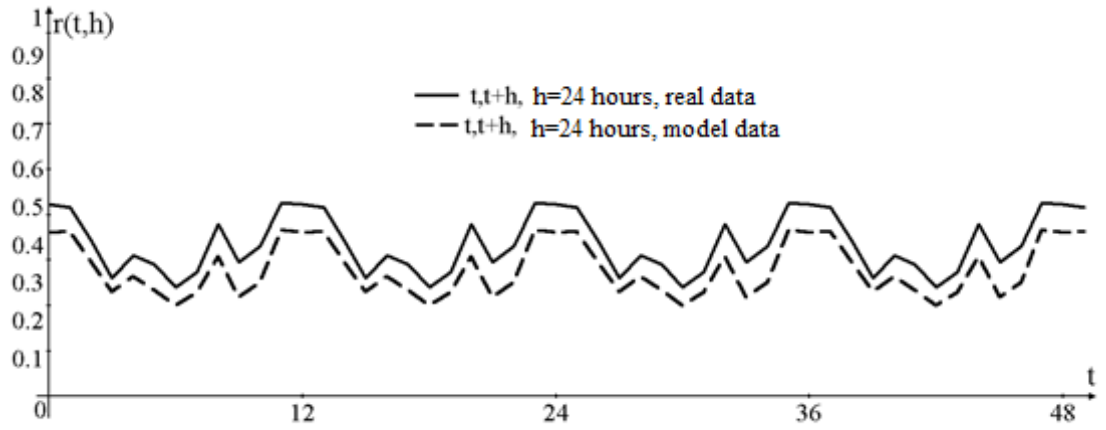


Figure 4: Correlation function $r_{TT}(t, t+h)$ time series real data and correlation function calculated with Gaussian model series ($h = 0, 1$)

On the basis constructed of a numerical stochastic model of series probabilities of certain unfavourable combinations of temperature and relative humidity as functions of time are constructed. On the Fig.5 given the probability of events $A = (T(t) < 70(F),\ H(t) > 90\%)\ \&\ B = (T(t) > 80(F),\ H(t) < 60\%)$. This characteristic is a single point one and depends on the mutual correlation coefficient between temperature and relative humidity at this point. The oscillation is determined by the periodicity of one-dimensional distributions. On the Fig.6 given the probability of events

$$C = ((T(t) < 70(F), T(t+h) < 70(F), T(t+2h) < 70(F))\& \\ (H(t) > 90\%, H(t+h) > 90\%, H(t+2h) > 90\%)),$$

$$D = ((T(t) > 80(F), T(t+h) > 80(F), T(t+2h) > 80(F))\& \\ (H(t) < 60\%, H(t+h) < 60\%, H(t+2h) < 60\%)),$$

which depends on the correlation structure of the joint series of temperature and relative humidity and periodically time-dependent.

Conclusions

In conclusion we should note that the model is constructed using a close modification of the method of the inverse distribution functions therefore reproduces the correlation structure with noticeable errors. However the fidelity of the correlation structure may be acceptable because it realistic reproduces the main features of the daily periodicity in the correlation function. We need to do some research on the

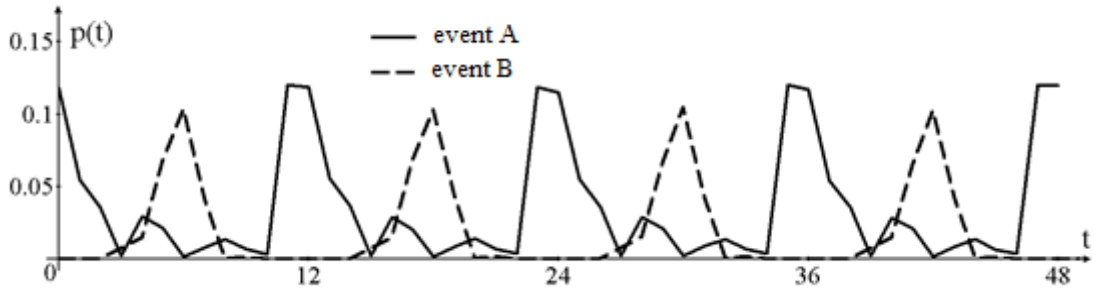


Figure 5: The probability of events A & B from time

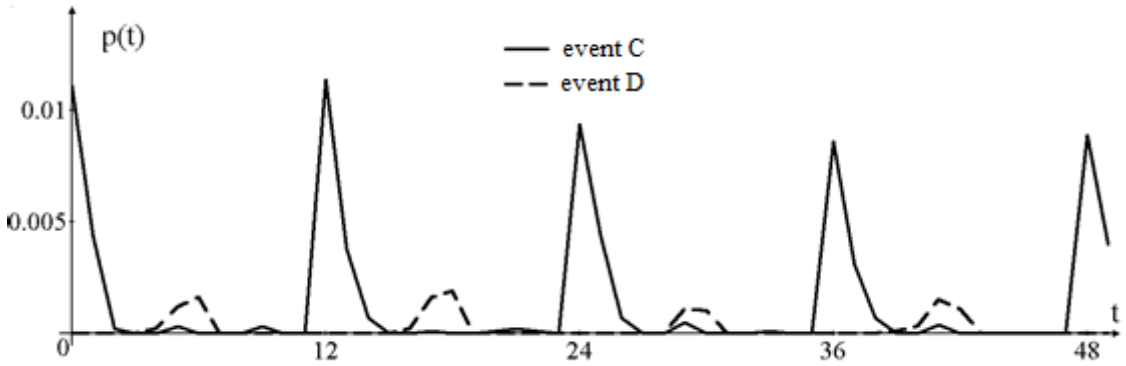


Figure 6: The probability of events C&D from time

verification of the model after refinement of the input parameters of the model. Preliminary estimators showed that the model describes some of the characteristics are quite acceptable. In the future it is planned to develop approaches to increase in the accuracy of the model.

Acknowledgements

The reported study was funded by RFBR and the government of the Novosibirsk region according to the research project No. 17-41-543338 r-mol-a; and RFBR grants No. 15-01-01458-a, 16-31-00123-mol-a, 16-01-00145.

References

[1] Bokov V. N., Lopatukhin L. I., Mikulinskaya S. M., Rozhkov V. A., Rumyantseva S. A. On The Interannual Variability Of The Disturbances. Saint Petersburg: Gidrometeoizdat, 1995, P. 446 – 454. (in Russian)

[2] Derenok K.V., Ogorodnikov V.A. Numerical simulation of significant long-term decreases in air temperature // Russ. J. Num. Anal. Math. Modelling, 2008, Vol. 23, No 3. P. 223 – 277.

[3] Dragan Y. P., Rozhkov V. A., Yavorsky I. N. The Methods of Probabilistic Analysis of Oceanological Rhythmics. L: Gidrometeoizdat, 1987, 320 P. (in Russian)

[4] Kargapolova N.A., Ogorodnikov V.A. Inhomogeneous Markov chains with periodic matrices of transition probabilities and their application to simulation of meteorological processes. // Russ. J. Num. Anal. Math. Modelling, 2012, Vol. 27, No 3. P. 213 – 228

[5] Marchenko A. S., Semochkin A. G. $F\Phi\Phi F$ - The method of modeling time series on observed realizations. // Numerical methods of statistical simulation. – Novosibirsk, 1987, pp. 14-22.

[6] Medvyatskaya A. M., Ogorodnikov V.A. The Approximate Model Of a Periodically Correlated Random Process Based On a Spectral Representation. Abstracts in Advanced Mathematics, Computations and Applications 2015. (in Russian)

[7] Ogorodnikov V. A., Prigarin S. M. Numerical Modeling of Random Processes and Fields: Algorithms and Applications, Utrecht: VSP. The Netherlands, 1996.

[8] Prigarin S.M. Numerical Modeling Of Random Processes And Fields. ICM&MG Publisher, Novosibirsk, 2005. 259 P. (in Russian)

Correlation Structure of the Piecewise Linear Process on the Poisson Flow

Vasily A. Ogorodnikov[1,2] and Olga V. Sereseva[1]
[1] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: ova@osmf.sscc.ru, seresseva@mail.ru

Abstract

The paper is concerned with the investigation of a class of piecewise linear processes on the Poisson flows with independent identically distributed random variables at the nodal Poisson points. An approach to the calculation of the correlation function of the process based on the use of the total probability formula is considered.

Keywords: piecewise linear process, Poisson flow, correlation function.

Introduction

In this paper a piecewise linear process on Poisson flows with independent identically distributed random values at the nodal Poisson points is considered. Processes of these type are a modification of piecewise-constant processes described for example by G.A. Mikhailov in [1]. The results of a study of a nonstationary piecewise linear process on Poisson flows whose values at the nodal Poisson points are a sum of independent identically distributed quantities with an arbitrary one-dimensional distribution with a consecutively increasing number of terms are given in [2],[3]. In these papers exact expressions for some characteristics of this process are given. In particular, the exact dependences of the initial moments of an arbitrary order on time are given. The asymptotic properties of these functions were also investigated.

In [1],[2] a piecewise linear process with independent identically distributed non-Gaussian random variables at Poisson support points was also considered. For this type of process, exact expressions are obtained for the mean value, variance, asymmetry coefficients and kurtosis as functions of time. In this paper we consider an approach to the calculation of the correlation function of this process.

1 Section of the Paper

Let's consider a piecewise linear process of the form [4]

$$Y(t) = (Y_{n+1} - Y_n)\,\frac{t - S_n}{S_{n+1} - S_n} + Y_n = (Y_{n+1} - Y_n)\,Q(t) + Y_n, \qquad (1)$$

where $S_n \le t < S_{n+1}$, $n = 0, 1, \ldots$, $S_0 = 0$, $S_n = \sum_{i=1}^{n} X_i$, $n \ge 1$, and X_i – are independent positive random variables with exponential probability density $f(x) = \lambda \exp(-\lambda x)$, $x \ge 0$, $\lambda > 0$.

The variables Y_n in (1) (Y_n: $Y_0 = \alpha_0$, $Y_n = \alpha_n$, $n \geq 1$) at the Poisson reference points S_n are independent of each other and X_i random variables with an arbitrary one-dimensional distribution $G(y)$.

The formal expression for the correlation function of the piecewise linear process $Y(t)$ (1) has the form:

$$corr(Y(t), Y(t+h)) = \frac{E[Y(t)Y(t+h)] - E[Y(t)]E[Y(t+h)]}{\sqrt{D[Y(t)]}\sqrt{D[Y(t+h)]}},\ t > 0,\ h > 0. \quad (2)$$

In formula (2), it is necessary to know the values of the process $Y(t)$ at the points t and $t+h$.

In order to calculate the mean $E[Y(t)Y(t+h)]$ in (2) we use the total probability formula

$$E[Y(t)Y(t+h)] = \sum_{k=1}^{7} P[B_{nm}^{(k)}(t, t+h)]E[Y(t)Y(t+h) \left| B_{nm}^{(k)}(t, t+h)\right] ,\ n, m = 1, 2, ...,$$

for the following probabilistic events:

$$1a.\ B_{nm}^{(1)}(t, t+h) = \{S_n < t, t < S_{n+1} < t+h, S_{n+1} < S_m < t+h, S_{m+1} > t+h\},$$
$$n = 1, 2, \ldots, \quad m = 1, 2, \ldots \ .$$

$$2.a\ B_n^{(2)}(t, t+h) = \{S_n < t, S_{n+1} > t+h\},\ n = 1, 2, \ldots \ .$$

$$3a.\ B_{n1}^{(3)}(t, t+h) = \{S_n < t, t < S_{n+1} < t+h,\ S_{n+2} > t+h\},\ n = 1, 2, \ldots,\ m = 1.$$

$$4a.\ B_1^{(4)}(t, t+h) = \{S_n > t+h\},\quad n = 1.$$

$$5a.\ B_1^{(5)}(t, t+h) = \{t < S_n < t+h, S_{n+1} > t+h\} ,\quad n = 1.$$

$$6a.\ B_{1m}^{(6)}(t, t+h) = \{t < S_n < t+h, S_n < S_{n+1} < t+h, S_{n+1} < S_m < t+h, S_{m+1} > t+h\},$$
$$n = 1,\ m = 1, 2, \ldots \ .$$

$$7a.\ B_{11}^{(7)}(t, t+h) = \{(t < S_n < t+h, S_n < S_{n+1} < t+h, S_m > t+h, S_{m+1} > S_m)\},$$
$$n = 1\,, \quad m = 1.$$

Let us find the joint density of the distribution of the variables S_n, S_{n+1}, S_m and S_{m+1}, n, $m = 1, 2, \ldots$ ($S_n < S_{n+1} < S_m < S_{m+1}$). Let us represent

$$\begin{aligned} &S_0 = 0, \quad S_n = S_0 + X_1 + X_2 + \ldots + X_n, \\ &S_{n+1} = S_n + X_{n+1}, \\ &S_m = S_{n+1} + X_{n+2} + \ldots + X_{n+m+1}, \\ &S_{m+1} = S_m + X_{n+m+2}, \quad n \geq 1, \ \ m \geq 1. \end{aligned} \quad (3)$$

On the basis of the transformation of the random variables determined by the relations (3), make the transformations of the variables:

$$\begin{aligned} &y_1 = x_1, \\ &y_2 = y_1 + x_2, \\ &y_3 = y_2 + x_3, \\ &y_4 = y_3 + x_4. \end{aligned} \quad (4)$$

The inverse transformation has the form:

$$\begin{aligned} x_1 &= y_1, \\ x_2 &= y_2 - y_1, \\ x_3 &= y_3 - y_2, \\ x_4 &= y_4 - y_3. \end{aligned}$$

The Jacobian of the transformation (4) is equal to 1:

$$J(y_1, y_2, y_3, y_4) = \mathrm{mod} \begin{vmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{vmatrix} = 1.$$

Taking into account the fact that the $S_n = \sum_{i=1}^{n} X_i$, $n \geq 1$, and X_i – are independent positive random variables with exponential probability density $f(x) = \lambda \exp(-\lambda x)$, $x \geq 0$, $\lambda > 0$, the joint distribution density of the random variables S_n, S_{n+1}, S_m, S_{m+1} is determined by the expression:

$$f_{S_n S_{n+1} S_m S_{m+1}}(y_1, y_2, y_3, y_4) = \frac{(\lambda y_1)^{n-1} \lambda e^{-\lambda y_1}}{(n-1)!} \lambda e^{-\lambda(y_2 - y_1)} \frac{\lambda^{m-1}(y_3 - y_2)^{m-1} \lambda e^{-\lambda(y_3 - y_2)}}{(m-1)!} \times$$
$$\times \lambda e^{-\lambda(y_4 - y_3)} J(y_1, y_2, y_3, y_4) = \lambda^4 \frac{(\lambda y_1)^{n-1}}{(n-1)!} \frac{\lambda^{m-1}(y_3 - y_2)^{m-1}}{(m-1)!} e^{-\lambda y_4}.$$

Then the probabilities of events 1a-7a have the form:

$$1b.\ P(B^{(1)}_{nm}(t, t+h)) = \lambda^{n+m+1} e^{-\lambda(t+h)} \frac{t^n}{n!} \frac{h^{m+1}}{(m+1)!}, \quad n = 1, 2, \ldots, \quad m = 1, 2, \ldots,$$

$$2b.\ P(B^{(2)}_{n}(t, t+h)) = \lambda^n \frac{t^n}{n!} e^{-\lambda(t+h)}, \quad n = 1, 2, \ldots,$$

$$3b.\ P(B^{(3)}_{n1}(t, t+h)) = e^{-\lambda(t+h)} h \frac{\lambda^{n+1} t^n}{n!}, \quad n = 1, 2, \ldots, \ m = 1,$$

$$4b.\ P(B^{(4)}_{1}(t, t+h)) = e^{-\lambda(t+h)}, \ \ n = 1,$$

$$5b.\ P(B^{(5)}_{1}(t, t+h)) = \lambda h e^{-\lambda(t+h)}, \ \ n = 1,$$

$$6b.\ P(B^{(6)}_{1m}(t, t+h)) = \frac{\lambda^{m+2} h^{m+2}}{(m+2)!} e^{-\lambda(t+h)}, \ \ n = 1, \ m = 1, 2, \ldots,$$

$$7b.\ P(B^{(7)}_{11}(t, t+h)) = \lambda^2 e^{-\lambda(t+h)} \frac{h^2}{2}, \quad n = 1, \quad m = 1.$$

Conditional probability densities of the variables S_n, S_{n+1}, S_m, S_{m+1} under the conditions 1*a*-7*a* are equal to:

$$1c.\ f(y_1, y_2, y_3, y_4 \,\big|\, B^{(1)}_{nm}(t, t+h)) =$$
$$= f(y_1, y_2, y_3, y_4 \,|\, S_n < t, t < S_{n+1} < t+h, S_{n+1} < S_m < t+h, S_{m+1} > t+h) =$$
$$= \frac{\lambda y_1^{n-1} (y_3 - y_2)^{m-1} e^{-\lambda y_4} nm(m+1)}{e^{-\lambda(t+h)} t^n h^{m+1}}.$$

$$2c.\ f(y_1, y_2 \left| B_n^{(2)}(t, t+h)\right.) = f(y_1, y_2 \,|S_n < t, S_{n+1} > t+h\,) = \frac{\lambda n y_1^{n-1} e^{-\lambda y_2}}{t^n e^{-\lambda(t+h)}}.$$

$$3c.\ f(y_1, y_2, y_3 \left| B_{n1}^{(3)}(t, t+h)\right.) =$$
$$= f(y_1, y_2, y_3 \,|S_n < t, t < S_{n+1} < t+h, S_m > t+h\,) = \frac{\lambda y_1^{n-1} e^{-\lambda y_3} n}{h t^n} e^{\lambda(t+h)}.$$

$$4c.\ f(y_1 \left| B_1^{(4)}(t, t+h)\right.) = f(y_1 \,|S_1 > t+h\,) = \frac{\lambda e^{-\lambda y_1}}{e^{-\lambda(t+h)}}.$$

$$5c.\ f(y_1, y_2 \left| B_1^{(5)}(t, t+h)\right.) = f(y_1, y_2 \,|t < S_n < t+h, S_{n+1} > t+h\,) = \frac{\lambda e^{-\lambda y_2}}{h e^{-\lambda(t+h)}}.$$

$$6c.\ f(y_1, y_2, y_3, y_4 \left| B_{1m}^{(6)}(t, t+h)\right.) =$$
$$= f(y_1, y_2, y_3, y_4 \,|t < S_n < t+h, S_n < S_{n+1} < t+h, S_{n+1} < S_m < t+h, S_{m+1} > t+h\,) =$$
$$= \lambda \frac{(m+2)(m+1)m(y_3 - y_2)^{m-1}}{h^{m+2} e^{-\lambda(t+h)}} e^{-\lambda y_4}.$$

$$7c.\ f(y_1, y_2, y_3 \left| B_{11}^{(7)}(t, t+h)\right.) =$$
$$= f(y_1, y_2, y_3 \,|t < S_n < t+h, t < S_{n+1} < t+h, S_m > t+h\,) = \frac{2\lambda e^{-\lambda y_3}}{e^{-\lambda(t+h)} h^2}.$$

Since

$$\begin{aligned} &E[Y(t)Y(t+h)] = \\ &\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} P[B_{nm}^{(1)}(t, t+h)] E[Y(t)Y(t+h) \left| B_{nm}^{(1)}(t, t+h)\right.] + \\ &+ \sum_{n=1}^{\infty} P[B_n^{(2)}(t, t+h)] E[Y(t)Y(t+h) \left| B_n^{(2)}(t, t+h)\right.] + \\ &+ \sum_{n=1}^{\infty} P[B_{n1}^{(3)}(t, t+h)] E[Y(t)Y(t+h) \left| B_{n1}^{(3)}(t, t+h)\right.] + \\ &\quad + P[B_1^{(4)}(t, t+h)] E[Y(t)Y(t+h) \left| B_1^{(4)}(t, t+h)\right.] + \\ &\quad + P[B_1^{(5)}(t, t+h)] E[Y(t)Y(t+h) \left| B_1^{(5)}(t, t+h)\right.] + \\ &+ \sum_{m=1}^{\infty} P[B_{1m}^{(6)}(t, t+h)] E[Y(t)Y(t+h) \left| B_{1m}^{(6)}(t, t+h)\right.] \\ &\quad + P[B_{11}^{(7)}(t, t+h)] E[Y(t)Y(t+h) \left| B_{11}^{(7)}(t, t+h)\right.], \end{aligned}$$

and Y_k , Y_l, $k \neq l$ are independent identically distributed variables with $EY_k = \mu$, $DY_k = \sigma^2$, then the main terms in this expression for $E[Y(t)Y(t+h)]$ have the form:

$$1d.\ \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} P[B_{nm}^{(1)}(t, t+h)] E[Y(t)Y(t+h) \left| B_{nm}^{(1)}(t, t+h)\right.] =$$

$$= \mu^2 e^{-\lambda(t+h)}(1 - e^{\lambda t})(1 - e^{\lambda h} + \lambda h),$$

$$2d.\ \sum_{n=1}^{\infty} P[B_n^{(2)}(t, t+h)] E[Y(t)Y(t+h) \left| B_n^{(2)}(t, t+h)\right.] =$$

$$= 2\sigma^2\lambda^2 \int_0^t (t - y_1)(e^{-\lambda(t+h-y_1)} - \lambda(t+h-y_1)\Gamma[0, \lambda(t+h-y_1)])\, dy_1 -$$

$$-\sigma^2\lambda^2\int\limits_0^t(t-y_1)\Gamma[0,\lambda(t-y_1)]\,dy_1-\sigma^2\lambda^2\int\limits_0^t(t+h-y_1)\Gamma[0,\lambda(t+h-y_1)]\,dy_1+$$

$$+(\sigma^2+\mu^2)(e^{-\lambda h}-e^{-\lambda(t+h)}),$$

$$3d.\ \sum_{n=1}^{\infty}P[B_{n1}^{(3)}(t,t+h)]E[Y(t)Y(t+h)\,\Big|B_{n1}^{(3)}(t,t+h)]=$$

$$=\sigma^2\lambda^3\int\limits_0^t\left(\int\limits_t^{t+h}\left(\int\limits_{t+h}^{\infty}\frac{t-y_1}{y_2-y_1}\frac{y_3-t-h}{y_3-y_2}e^{\lambda y_1}e^{-\lambda y_3}dy_3\right)dy_2\right)dy_1+\mu^2\lambda h(e^{-\lambda h}-e^{-\lambda(t+h)}),$$

$$4d.\ P[B_1^{(4)}(t,t+h)]E[Y(t)Y(t+h)\,\Big|B_1^{(4)}(t,t+h)]=$$

$$=2\sigma^2\lambda te^{-\lambda(t+h)}-2\sigma^2\lambda^2t(t+h)\Gamma[0,\lambda(t+h)]-2\sigma^2\lambda t\Gamma[0,\lambda(t+h)]-$$

$$-\sigma^2\lambda h\Gamma[0,\lambda(t+h)]+e^{-\lambda(t+h)}(\sigma^2+\mu^2),$$

$$5d.P[B_1^{(5)}(t,t+h)]E[Y(t)Y(t+h)\,\Big|B_1^{(5)}(t,t+h)]=$$

$$=-\sigma^2\int\limits_t^{t+h}\int\limits_{t+h}^{\infty}\frac{t}{y_1}\frac{t+h-y_1}{y_2-y_1}\lambda^2e^{-\lambda y_2}dy_2dy_1+\sigma^2\int\limits_t^{t+h}\frac{\lambda t}{y_1}e^{-\lambda(t+h)}dy_1+\lambda he^{-\lambda(t+h)}\mu^2,$$

$$6d.\ \sum_{m=1}^{\infty}P[B_{1m}^{(6)}(t,t+h)]E[Y(t)Y(t+h)\,\Big|B_{1m}^{(6)}(t,t+h)]=$$

$$=e^{-\lambda(t+h)}(e^{\lambda h}-1-\lambda h-\frac{1}{2}\lambda^2h^2)\mu^2,$$

$$7d.P[B_{11}^{(7)}(t,t+h)]E[Y(t)Y(t+h)\,\Big|B_{11}^{(7)}(t,t+h)]=\lambda^2e^{-\lambda(t+h)\frac{h^2}{2}}\mu^2.$$

Thus, the mean $E[Y(t)Y(t+h)]$ in formula (2) is determined by the expressions 1d-7d described above.

It was shown in [4] that for a piecewise linear process $Y(t)$ of the form (1) that the mathematical expectation is equal to

$$E[Y(t)]\ =\mu, \tag{5}$$

where $\mu=EY_n,\ n=1,2,...$, and variance of the process (1):

$$D[Y(t)]=\sigma^2\left(2(E[Q^2(t)\,]-E[Q(t)])+1\right), \tag{6}$$

where $\sigma^2=D[Y_n],\ n=1,2,...,$

$$E[Q^k(t)]=\frac{1}{k+1}(1-e^{-\lambda t}(\lambda t+1)+(\lambda t)^k(1+k+\lambda t)\Gamma[1-k;\lambda t]),k=1,2,\ldots, \tag{7}$$

where $\Gamma[a,\,z]=\int\limits_z^{\infty}x^{a-1}e^{-x}dx$ is Gamma function.

Thus, the correlation function (2) of the piecewise linear process $Y(t)$ of the form (1) is completely determined by the expressions 1d-7d and by the formulas (5)-(7).

The expression for the correlation function (2) for t=0 takes the form

$$corr(Y(0), Y(h)) = \frac{e^{-\lambda h} - h\lambda\,\Gamma[0, \lambda h]}{\sqrt{2(E[Q^2(h)] - E[Q(h)]) + 1}} \tag{8}$$

Using direct modeling trajectories of the process (1), similar functions were calculated for different values of parameter λ. Numerical calculations that these functions, with accuracy up to a statistical error in estimating the correlation function, coincide with the correlation function (8).

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grants No. 15-01-01458-a, 16-31-00123-mol-a, 16-01-00145, 17-07-00775).

References

[1] Mikhailov G.A. Optimization of weighted Monte Carlo methods // -M: Nauka, 1986 (in Russian), [Engl.transl.: Springer-Verlag, 1992].

[2] Ogorodnikov V.A., SavelievL.Ya., Sereseva O.V. Computation stochastic models of piecewise random processes. // Russian Journal of Numerical Analysis and Mathematical Modelling. 2007. Vol. 22, No. 3. P. 505-514.

[3] L.Ya. Savel'ev, V.A.Ogorodnikov, O.V.Sereseva Stochasticmodel ofpiecewise-linear random process. // Vestnik Syktyvkar university. 2007. V.1. No. 7. P. 67-76. (in Russian)

[4] Sereseva O.V. Nonstationary and asymptotically stationary piecewise linear processes on the Poisson flows and some of their properties // Proceedings of the conference of young scientists. ICMMG SB RAS. 2016. P. 83-92. (in Russian)

[5] Sereseva Olga V. and Ogorodnikov Vasily A. Piecewise Linear Processes on Pouisson Flow // Proceedings of the International Workshop "Applied Methods of Statistical analysis. Nonparametric approach". Novosibirsk&Belokurikha, 2015, pp. 384-391.

Conditional Stochastic Model of Daily Precipitation and River Flow Joint Spatial Field

Nina A. Kargapolova[1,2] and Vasily A. Ogorodnikov[1,2]
[1] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: nkargapolova@gmail.com, ova@osmf.sscc.ru

Abstract

In this paper a stochastic model of river flow and daily precipitation joint fields is suggested. This model is based on real data from weather stations situated in Novosibirsk region and gauging stations on Berd' river. Spatial heterogeneity of precipitation distributions is taken into account.

Keywords: stochastic simulation, daily precipitation, river flow, heterogeneous field.

Introduction

Groundwater inflow, precipitation and snowmelt in the river basin and artificial drainage are usually considered as key factors defining a river flow [2]. It is obvious that it is not always possible to carry out a full-scale experiment and to study on its basis influence of each specific factor on the river flow. By contrast, numerical simulation may help to solve this problem. In this paper precipitation influence on the river flow is considered. One of the approaches to numerical study of such influence is a dynamical-probabilistic approach. In this approach precipitation in a river basin are considered as a random field and entry of water into the river from its water system is simulated on a basis of a deterministic dynamical hydrological model. In this paper another approach is suggested: both precipitation and river flow are considered as a joint random process. Corresponding real data analysis, model description and simulation algorithms are given in this paper.

1 Model description

In this chapter assumptions of daily precipitation and river flow random fields statistical structure that were used for a stochastic model development and model parameters selection methods are described.

The model of daily precipitation and river flow joint random fields is based on real data from 56 weather stations situated in Novosibirsk region. Data about daily precipitation (mm) was collected since 1969 to 1983. Two weather stations (Maslyanino and Stariy Iskitim) are also gauging stations on Berd' river where average daily river flow (m^3/s) was measured. Figure 1 represents a map of Novosibirsk region with marked weather and gauging stations. Figure 2 shows Berd' catchment basin.

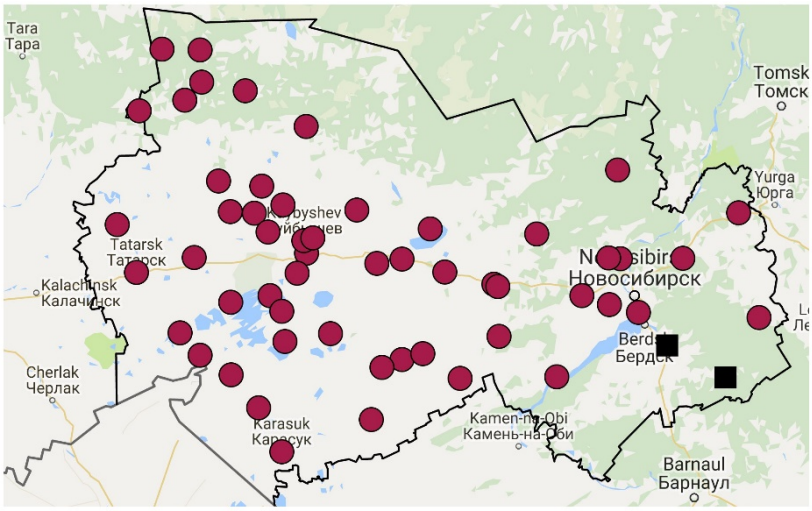


Figure 1: Map of Novosibirsk region. Circles – weather stations, squares – weather and gauging stations

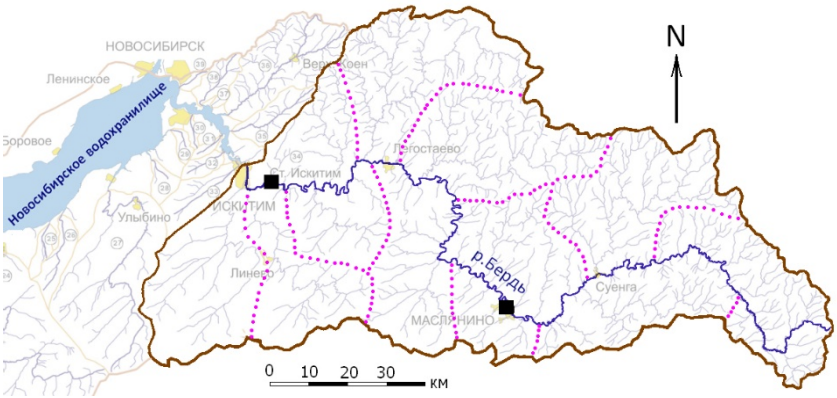


Figure 2: Berd' river catchment basin

Simulation of a precipitation field was done on gauging stations and in nodes of a rectangular grid, covering the catchment basin, with a grid step $\Delta x = \Delta y = 3\ km$ in South-North and West-East directions. Total number of points where simulation of precipitation was done is equal to $N = 1158$ (1156 grid nods and 2 gauging stations). It should be noted that only 2 weather stations lie within the simulation area. In this paper we consider a model in which river flow is simulated in a single point (on one of the gauging stations).

For precipitation simulation, a modification of an approach presented in [3, 5] was used. It is supposed that precipitation χ_i in a point number i, $i = 1, \ldots, N$ may be presented in a form

$$\chi_i = \omega_i \xi_i,$$

where ω_i is a value of a precipitation indicator random field ω (if $\omega_i = 1$ a day is

considered as a wet day (there are non-zero precipitation) in a point i, if $\omega_i = 0$ it's a dry day) and ξ_i is amount of precipitation in a point is i under a condition of precipitation presence. Random fields ω and ξ are independent on each other. These fields are supposed to be homogeneous in sense of their correlation structure. Since real data allow to estimate correlation coefficients of a precipitation field only between weather stations and give no information about correlation coefficients between nods, it was necessary to approximate real data based correlation coefficients with correlation functions of continuous arguments and calculate correlation matrixes C^{ω}, C^{ξ} of fields ω and ξ using these functions. Approximating functions have a form

$$corr\,(x_i, y_i, x_j, y_j) = corr\,(x_j - x_i, y_j - y_i) = corr\,(x, y) = \exp\left(-\left[ax^2 + bxy + by^2\right]^d\right).$$

Parameters of approximating functions are given in Table 1. Elements of correlation matrixes C^{ω}, C^{ξ} are denoted as $c^{\omega}_{ij}, c^{\xi}_{ij}$ $(i, j = 1, \ldots, N)$ respectively.

Table 1: Parameters of approximating correlation functions. July

Field	a	b	c	d
ω	0.0021	-0.0026	0.0010	0.2130
ξ	0.0032	-0.0090	0.0095	0.3010

In [5] it was supposed that fields ω and ξ are homogeneous in sense of their one-dimensional distributions. Here we consider these fields as heterogeneous. One-dimensional distribution of the indicator field ω in each point i is defined with a probability $p_i = P\,(\omega_i = 1)$. Gamma-distribution with density

$$f_i\,(x) = x^{k_i - 1} \exp\left(-\frac{x}{t_i}\right) \Big/ \Gamma\,(k_i)\, t_i^{k_i}$$

is assumed to be a one-dimensional distribution of the field ξ in point i. Parameters p_i, k_i, t_i $(i = 1, \ldots, N)$ were determined on a basis of real data using IDW-interpolation technique. We should note that usage of gamma-distribution as a one-dimensional distribution of precipitation amount and IDW-interpolation for distribution parameters calculation is a common approach in Statistical Meteorology (see, for example [1, 7, 9]).

Type of one-dimensional distribution function of river flow varies on different gauging stations. In this regard in this paper as river flow distribution function we use piecewise-linear approximation of a sample distribution function with exponential tail.

There is a rather tricky question: "How to define correlation coefficients between precipitation in the grid nods and river flow on a gauging station?" We use an approach that was proposed in [8]. Let r^{ω}_i be a correlation coefficient between precipitation indicator in a nod number i and river flow on a gauging station Δt days later. We suppose that r^{ω}_i may be presented in a form

$$r^{\omega}_i = r_{\omega s}\,(\Delta t) \cdot c^{\omega}_{ik},$$

where $r_{\omega s}(\Delta t)$ is a correlation coefficient between precipitation indicator and river flow on a gauging station Δt days later, c_{ik}^{ω} is a defined above spatial correlation coefficient of precipitation indicator in a nod number i and on a gauging station ($k = N-1$ or $k = N$ depending on a gauging station under consideration). Correlation coefficients r_i^{ξ} between precipitation amount and river flow are defined in an analogous manner.

2 Simulation algorithm

It is necessary to simulate a joint random field of precipitation indicators, precipitation amount ad river flow. First two fields are defined in N points, the last one – in a single point. Joint correlation matrix is a $(2N+1) \times (2N+1)$ square matrix, and its dimension doesn't let to simulate field using Choletsky or spectral decomposition of the correlation matrix. Yet method of conditional distributions may be applied for simulation [4, 10]. Since fields ω and ξ are independent, method of conditional distributions simplifies and turns into next simulation algorithm:

1. Using threshold transformation of a Gaussian process, a field ω of precipitation indicators with correlation matrix C^{ω} is simulated.

2. Independently on ω a field ξ of precipitation amount with correlation matrix C^{ξ} is simulated using inverse distribution function method [6].

3. Fields ω and ξ are pointwise multiplied to form a final field of precipitation.

4. When fields ω and ξ are given, a single point conditional field of river flow is simulated.

Acknowledgements

This work was supported by the Russian Foundation for Basis Research (grants No 15-01-01458-a, 16-01-00145-a, 16-31-00123-mol-a, 16-31-00038-mol-a) and the President of the Russian Federation grant (No MK-659.2017.1).

References

[1] Hartkamp A.D., De Beurs K., Stein A., White J.W. (1999). *Interpolation Techniques for Climate Variables*. CIMMYT, Mexico.

[2] Komlev A.M. (2002). *Formation patterns and calculation methods of a river flow*. PSU, Perm. (in Russian)

[3] Marchenko A.S., Ogorodnikov V.A. (1991). *Probabilistic models of dry and wet days sequenses: preprint No 933*. CC SB AS USSR, Novosibirsk. (in Russian)

[4] Ogorodnikov V.A., Prigarin S.M. (1996). *Numerical modelling of random processes and fields: Algorithms and Applications*. VSP, Utrecht.

[5] Ogorodnikov V.A., Sereseva O.V. (2015). Multiplicative nimerical stochastic model of daily precipitation random fields and its application for statistical study of extreme precipitation. *Optics of athmosphere and ocean*. Vol. **28**, pp. 238-245. (in Russian)

[6] Prigarin S.M. (2005). *Methods of numerical simulation of random processes and fields*. ICMaMG SB RAS, Novosibirsk. (in Russian)

[7] Richardson C.W., Wright D.A. (1984). *WGEN: A Model for Generating Daily Weather Variables*. U.S. Department of Agriculture, Agricultural Research Service.

[8] Shlichkov V.A., Ogorodnikov V.A., Sereseva O.V. (2015). Joint numerical stochastic model if daily river flow time-series and spatio-temporal precipitation fields. *Interexpo Geo-Sibir*. Vol. **4**, pp. 150-154. (in Russian)

[9] Sluiter R. (2009). *Interpolation methods for climate data (literature review)*. KNMI,De Bilt.

[10] Sobol I.M. (1973). *Numerical monte Carlo methods*. Nauka, Moscow. (in Russian)

The Monte Carlo Method for Determining the Vision System Characteristics

Irina Gendrina and Maria Alekseenko
National Research Tomsk State University, Tomsk, Russia
e-mail: igendrina@bk.ru

Abstract

Some of the Monte-Carlo algorithms for determining the point spread function of vision system through the atmosphere were considered. Calculations using the method of local estimate for the spherical and plane-parallel models of the atmosphere were conducted. The results of statistical simulation were processed using the regression analysis.

Keywords: Vision system, point spread function, local estimate, regression.

Introduction

Such method of studies as the system approach has been intensively developed in different fields of science, industry, and social life in recent decades. System approach is defined to be that which considers any system (object) to be a set of interrelated elements (components), and to have output (purpose), input (resources), connection with environment, and feedback. System approach represents a form of applications in the theory of knowledge and dialectics to studying the processes that occur in nature, community, and thinking. It essentially consists of fulfillment of the requirements of the general system theory; in accordance to this theory, each object in the process of its study should be considered as a large and complex system and, simultaneously, as an element of a more general system. One of the applications of the system approach is the optics [3] and, in particular, the atmospheric optics [4]. The main system characteristic in these fields is the point spread function (PSF); it is defined as the response L of linear system to the input signal, representing a point mass $\delta(x - x_1; y - y_1)$, located at a certain point $(x_1; y_1)$:

$$L\left[\delta(x - x_1; y - y_1)\right] = h\left(x, x_1; y, y_1\right)$$

An arbitrary object (function) $f(x, y)$ can be considered as a set of point masses. For instance:

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, y_1)\, \delta(x - x_1; y - y_1)\, dx_1 dy_1$$

. Then, a result of the system impact (image) can be represented in the form:

$$g(x, y) = L\left[f(x, y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, y_1)\, h\left(x, y; x_1; y_1\right) dx_1 dy_1$$

Obviously, regularities of the image distortion due to impact of any system can be studied by analyzing the effect of this system on the point spread function. We will consider a system "underlying system – atmosphere – receiving device" which can be regarded as a linear system [4]. These systems are conventionally called the vision systems. Image distortion in the vision systems may be caused by the properties of scattering medium and underlying surface, and by characteristics of receiving optical device. The available images of any objects should be analyzed, and possible distortions of object images should be predicted, by studying the point spread function of this system. One of the methods for the PSF determination in this case is to calculate the angular distribution of brightness of surface-based point source, measured with receiving device at the top of the atmosphere (TOA). The purpose of this work is to study the dependence of the angular distribution of brightness of radiation at TOA on the geometrical and optical observation conditions.

1 Problem Statement

Change in the brightness of radiation of point source in the medium in stationary case is described by the following integro-differential transfer equation:

$$(\vec{\omega}, gradI\,(\vec{r},\vec{\omega})) = -\sigma\,(\lambda,\vec{r}) + \sigma_s\,(\lambda,\vec{r}) \int\limits_{\Omega} I\left(\vec{r},\vec{\omega'}\right) g\left(\vec{r},\vec{\omega},\vec{\omega'}\right) d\vec{\omega'} + \Phi_0\,(\vec{r},\vec{\omega}) \quad (1)$$

Here, $\vec{x}\,(\vec{r},\vec{\omega})$ is a point of phase space $X = R \times \Omega$ of coordinates $\vec{r} \in R$ and directions $\vec{\omega} \in \Omega$. $\Phi_0\,(\vec{r},\vec{\omega})$ is the distribution density of sources. $I\,(\vec{r},\vec{\omega})$ is the intensity (brightness) at point $(\vec{r},\vec{\omega})$.

In the work, we consider two models of the atmosphere: a plane-parallel model and a spherical model. For the plane-parallel model of the atmosphere all quantities in formula (1) depends on just one coordinate, namely, the depth z , while the intensity of the scattered radiation will be a function of the coordinate z and direction of radiation, characterized by the zenith angle Θ and azimuth φ in the horizontal plane. Values of z vary from 0 at the bottom of the atmosphere to H at TOA, where H is the thickness of the medium. The source is on the Earth's surface, and it will be assumed to be the origin of the coordinates. The receiver is at the upper boundary of the medium and has the coordinates $(0, 0, H)$. The system has the circular symmetry; therefore, all can be thought to proceed in the YOZ plane corresponding to $\varphi = 0$. Layered-homogeneous plane-parallel model was chosen as a model of the medium; it consists of n homogeneous layers, the geometrical thickness of which is characterized by the coordinate z (Figure 1).

The second model, considered here, represents the layered-homogeneous spherical atmosphere. In this case, all quantities depend on the distance from Earth's surface h , while the intensity of scattered radiation will be the function of h and direction of radiation $\vec{\omega} = (\theta, \varphi)$. The Earth's center is assumed to be the origin of coordinates. The source is on the Earth's surface and has the coordinates $(0, 0, R_0)$; while the receiver is at TOA and has the coordinates $(0, 0, R)$, where R_0 is the radius of

Earth, and R is the outer radius of the atmosphere. In this model of the medium, the atmosphere with the thickness $H = R - R_0$ is divided into n spherical layers with radii R_i, $i = \overline{1,n}$, $R_n = R$ (Figure 2) [2].
The optical model of the atmosphere implies specification of the following parameters.
1) The coefficients of aerosol scattering $\sigma_s(h,\lambda)$ and absorption $\sigma_a(h,\lambda)$. Here, h is the height above the Earth's surface, and λ is the wavelength. The coefficients σ_s and σ_a are assumed to be piecewise constant, since the atmosphere is divided into homogeneous layers.
2) The scattering phase function $g(h,\mu,\lambda)$. Here, $\mu = \left(\vec{\omega},\vec{\omega'}\right)$ is the cosine of the scattering angle. The scattering phase function is specified by dividing the atmosphere into layers, in each of which the scattering phase function is assumed to be constant with altitude h.

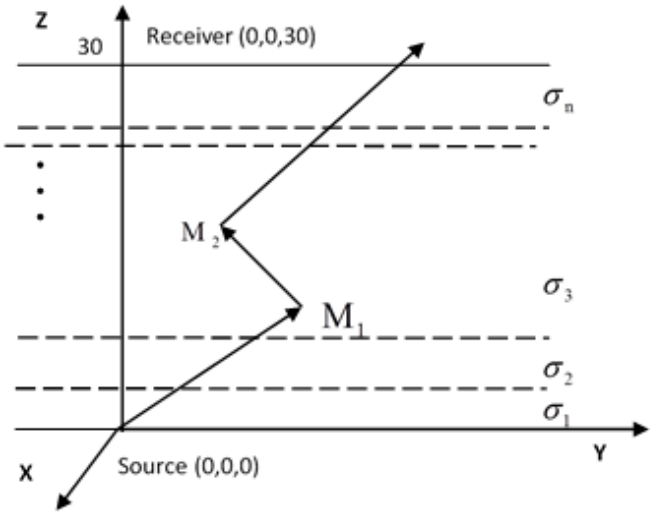


Figure 1: Shematic representation of the plane-parallel model of the atmosphere

One of the most universal methods for solving equation (1) is the method of statistical simulation, or the Monte Carlo method. This method is based on the integral transfer equation of the second kind with generalized kernel for the particle collision density:

$$f(\vec{x}) = \int_X k\left(\vec{x'},\vec{x}\right) d\vec{x'} + \psi(\vec{x}).$$

Monte Carlo method is usually used to estimate linear functionals of the form (2):

$$I_\varphi = (f,\varphi) = \int_X f(\vec{x})\,\varphi(\vec{x})\,d\vec{x}. \tag{2}$$

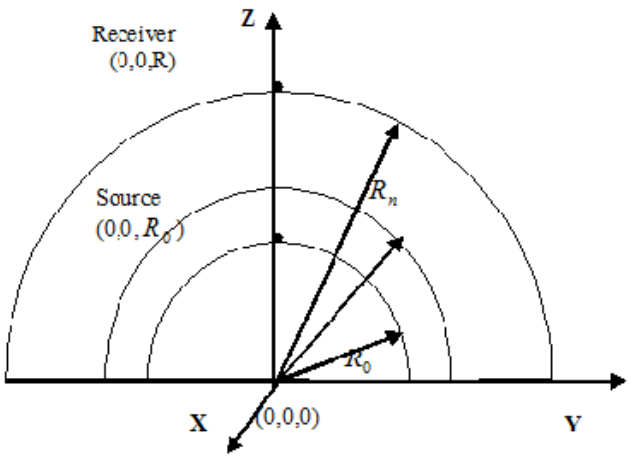


Figure 2: Shematic representation of the spherical model of the atmosphere

If $\{x_n\}$ is a 'physical' chain of collisions, then $I_\varphi = M\xi$, where $\xi = \sum_{n=0}^{N} Q_n \varphi\left(\vec{x_n}\right)$.
We single out two main algorithms of the Monte Carlo method.
1) The algorithm of direct simulation. This method is based on simulation of random trajectories of photon passage through the scattering medium. It is noteworthy that the characteristics of radiation process, necessary for analysis, are estimated in accordance with their physical meaning. A disadvantage of the direct simulation is that such characteristics as intensity, illumination, and others cannot be calculated with a sufficient accuracy. However, the direct simulation can be used to construct some other methods which make it possible to perform the required calculations. One of these methods is the algorithm of local estimate.
2) Algorithm of local estimate. The algorithm of local estimate consists of the calculation of the following functional:

$$J\left(\Theta_i\right) = \int\limits_{\Theta_i} \Phi\left(\vec{r^*}, \vec{\omega^*}\right) d\vec{\omega^*} = \int\limits_{X} l_i\left(\vec{x'}, \vec{x^*}\right) f\left(\vec{x'}\right) d\vec{x'} = M \sum_{n=0}^{N} Q_n \cdot l_i\left(\vec{x_n}, \vec{x^*}\right).$$

$$l_i\left(\vec{x'}, \vec{x^*}\right) = \frac{\exp\left[-\tau\left(\vec{r}, \vec{r^*}\right)\right] \cdot g(\mu^*)}{2\pi \cdot \left|\vec{r} - \vec{r^*}\right|^2} \cdot \Delta_i\left(\vec{s^*}\right).$$

Here, $\vec{s^*} = \frac{\vec{r} - \vec{r^*}}{\left|\vec{r} - \vec{r^*}\right|}$, $\mu^* = \left(\vec{\omega}, \vec{s^*}\right)$. $\Delta_i\left(\vec{s^*}\right)$ is the indicator of the region Θ_i, $\Phi\left(\vec{r^*}, \vec{\omega^*}\right)$ is the particle flux at a present point of the phase space $\vec{x^*} = \left(\vec{r^*}, \vec{\omega^*}\right)$. Q_n is the particle 'weight'.
We do not simulate the absorption, but multiply the 'weight' by the scattering probability and, namely, by the single scattering albedo. It is noteworthy that the variance decreases, and the average time of trajectory simulation on computer increases.
This algorithm is used to calculate the angular distribution of brightness, which repre-

sents the following quantity: $I(\Theta) = \frac{J(\Theta)}{\Theta}$, where Θ is the value of the corresponding to θ solid angle.

2 Initial data

We will consider the process of radiative transfer through aerosol-molecular atmosphere, which comprises a layer overcast, by neglecting the reflection from underlying surface.
We used the following data [1]:

1. Wavelength (mkm) in transparent windows: 0.347; 0.530; 0.694; 0.860; 1,060; 3,390; 10.60.

2. Lower boundary of atmosphere 0 km above Earth's surface, upper boundary H of the atmosphere 30 km above the Earth's surface.

3. Optical thickness for a cloudless atmosphere shown in Table 1 presented.

Table 1: Optical thickness of the cloudless atmosphere

Wavelength, mkm	Optical thickness
0,347	0,228
0,53	0,158
0,694	0,124
0,86	0,098
1,06	0,092
3,39	0,067
10,6	0,041

4. Lower boundary of the cloud layer - 1 km above the Earth's surface, Upper boundary - 2 km above the Earth's surface. The optical model of the cloud layer - haze H.

5. In this work, we considered two models of sources of radiation: Lambertian and isotropic sources. In case of isotropic source, density of the initial areas looks like: $p(\vec{\omega}) = \frac{1}{2\pi}$. For Lambertian this value is defined as: $p(\vec{\omega}) = \frac{\mu}{\pi}$.

For the statistical experiments was used a local estimate method in the scheme of constructing conjugate trajectories. Calculations were carried out based software package, which designed for a cloudless atmosphere. This complex was modified in order to be able to conduct simulations under the presence of clouds.
Numerical simulation results are shown in Table 2.

Table 2: The brightness of the scattered radiation for cloud atmosphere for various reception angles

Angle of reception in degrees	0,347 *mkm*	0,53 *mkm*	0,694 *mkm*	0,86 *mkm*	1,06 *mkm*	3,39 *mkm*	10,6 *mkm*
4,50	4,30E-07	5,01E-07	5,29E-07	5,46E-07	5,18E-07	4,91E-07	3,14E-07
13,5	1,31E-07	1,36E-07	1,01E-07	9,47E-08	8,07E-08	3,58E-08	3,65E-08
22,5	5,07E-08	4,30E-08	4,23E-08	3,50E-08	2,85E-08	8,37E-09	7,65E-09
31,5	2,29E-08	1,84E-08	1,20E-08	1,19E-08	7,86E-09	3,73E-09	2,17E-09
40,5	6,94E-09	8,43E-09	6,32E-09	5,25E-09	3,75E-09	6,66E-10	9,15E-10
49,5	3,30E-09	2,46E-09	3,19E-09	2,69E-09	1,97E-09	5,41E-10	4,77E-10
58,5	1,92E-09	1,71E-09	1,36E-09	1,37E-09	1,17E-09	3,49E-10	2,72E-10
67,5	1,18E-09	1,10E-09	7,91E-10	7,98E-10	6,56E-10	2,15E-10	1,25E-10
76,5	7,38E-10	1,07E-09	7,76E-10	5,14E-10	5,12E-10	9,37E-11	4,19E-11
85,5	1,76E-10	5,61E-10	9,50E-11	1,58E-10	1,05E-10	1,06E-11	5,66E-12

3 Statistical analysis of simulation result

Regression analysis was used to establish a functional link between the angular distribution of brightness and optical parameters. This analysis is widely used to restore the characteristics of aerosol and clouds, as well as to assess their impact on the climate.

The regression equation was built for the dependence of brightness on wavelength in the cloud spherical atmosphere at fixed angles reception. Regression coefficients and the coefficient of determination R^2 are shown in Table 3.

Table 3: Regression coefficients and the coefficient of determination for the cloudy atmosphere for fixed wavelengths

Wavelength, *mkm*	Coefficient b_0	Coefficient b_1	Coefficient R^2
0,347	-9,68	-2,56	0,904
0,53	-10,00	-2,44	0,948
0,694	-9,41	-2,70	0,913
0,86	-9,54	-2,68	0,944
1,06	-9,58	-2,73	0,948
3,39	-8,84	-3,28	0,940
10,6	-8,80	-3,38	0,921

For various optical models of the atmosphere was built the regression equation for the dependence of brightness on reception angle for considered wavelengths. This regression equation was obtained in the form of $y = b_0 + \frac{b_1}{x}$ for all wavelengths. The regression coefficients are shown in Table 4-5, coefficients of determination were also

given for each equation.

Table 4: Regression coefficients and the coefficient of determination for cloudless atmosphere for fixed wavelengths

Wavelength, mkm	Coefficient b_0	Coefficient b_1	Coefficient R^2
0,347	-9,69	-2,46	0,926
0,53	-9,73	-2,48	0,925
0,694	-9,62	-2,56	0,941
0,86	-9,55	-2,63	0,943
1,06	-9,58	-2,68	0,953
3,39	-9,10	-3,14	0,955
10,6	-8,86	-3,28	0,939

Statistical evaluation of the regression equations on the significance was conducted by F-test and Student's t-test. With confidence 0,9 it can be argued that the considered dependence are statistically significant. It should be noted that the obtained analytical data for the cloudless atmosphere are more accurate as compared with the data from the atmosphere cloud layer.

This once again confirms that the turbid layered homogeneous environment analysis and the preparation of some of the radiation brightness characteristics of a distorted unlike characteristics in the clear atmosphere.

The analysis shows that between the obtained angular distributions of intensity and wavelength for both cloudless and cloudy for the atmosphere, there is a link that can be with a good degree of accuracy to describe hyperbolic regression equation. This fact can be used for prediction, assessment and analysis of images of objects observed through the scattering medium.

Conclusions

In the course of the work was posed and solved the problem:

1. Construct a simulation model for the propagation of radiation, which based on the dual transport equation using a local assessment method.

2. Implemented algorithm of the Monte Carlo method for the calculation of the angular distribution of the radiation intensity of a point source in a cloudless sky.

3. Investigated dependence of the angular distribution of brightness on the wavelength, the thickness of the atmospheric model and radiation sources.

4. Applying regression analysis to establish a functional relationship between the angular distribution of brightness and the parameters of the model atmosphere.

References

[1] Krekov G.M., Rakhimov R.F. (1986).*Optical models of atmospheric aerosol*. Publishing House of Tomsk Affiliate of the Siberian Branch of the USSR Academy of Sciences, Tomsk.

[2] Marchuk G.I., Mikhailiov G.A., Nazaraliev M.A., Darbinjan R.A., Kargin B.A., Elepov B.S. (1976). *The Monte Carlo method in atmospheric optics*. "Nauka", Novosibirsk.

[3] Papoulis A. (1968).*Systems and transforms with applications in optics*. McGraw-Hill Book Company.

[4] Zuev V.E., Belov V.V., Veretennikov V.V. (1997).*Systems with applications in scattering media*. Publishing House "Spectrum" Institute of Atmospheric Optics of the Siberian Branch of the Russian Academy of Sciences, Tomsk.

The Uncertainty of a Control and the Control of an Uncertainty

Filimonov V.A.
Sobolev Institute of Mathematics, Omsk, Russia
e-mail: filimonov-v-a@yandex.ru

Abstract

The structure of the phenomenon of uncertainty is under consideration. Applied researches are some subjective cognitive processes that are immersed in a specific infrastructure. Statistical processing is considered in the paradigm of the simulation modeling as a simulation add-in model. The two-step test for goodness-of-fit and randomized survey are considered as the examples.

Keywords: research infrastructure, cognitive system, cross-technologies, subject, reflexive control, randomization.

Introduction

We consider that any statistical study is a system of subjective cognitive processes immersed in a specific infrastructure. In the capacity of such infrastructure we use cross-technologies of situational center [2, 3]. We will restrict ourselves here with a specification that such an approach requires methodological analysis and reflexive consideration both studied object and the researcher of this object as well. When a researcher writes the definition of his problem he must understand how and why all the components of the problem have been given to him. In this text we emphasize the importance of the concept "uncertainty". This concept requires the study of possible options of certainty and understanding of how to troubleshoot this certainty. G. W-F. Hegel wrote: "The Nothing from some Something, is a some specific Nothing (Das Nichts von irgend Etwas, ist ein bestimmtes Nichts)". We consider the uncertainty from the standpoints of its source and control. Under control we mean here the possibility of measuring of the uncertainty and the possibility of its reduction or increase. Objects and subjects are the sources of the uncertainty here. The size of this article does not allow to consider the fundamental publications devoted to the issues of uncertainty and ignorance, including such authors as N. Cusa, A.S. Narinyani and others. Table 1 presents some examples of the classification of uncertainty associated with statistical studies.

We limited to the references to the phenomena described in the literature on synergy [8], reflexive control [6, 7] and multi-dimensional optimization (the use of quasi-random numbers instead of pseudo-random) [10]. We describe two methods below: two-stage analysis of the goodness-of-fit and method of randomized survey. The choice of these methods is due to the fact that the basic idea lies not so much in the mathematical apparatus, but in the logic of the application.

Table 1: Classification and examples of uncertainty

The source of uncertainty	**Controlled uncertainty**	**Uncontrolled uncertainty**
Object	Parameters of a known distribution. Channels in the chaos. Quasi-random numbers.	Dynamical chaos. Jokers in chaos.
Subject	Reflexive control.	External Navigator of a Subject.

1 Statistical analysis in simulation systems

More than 30 years ago, the author on his own experience of simulation of digital communications came to the following conclusions [9]:

- it is advisable to replace two stages of the process "simulation and statistical processing of its results" to a single process "simulation that includes this statistical processing item";
- it is advisable to start project with the following prototype: "the simplest version of the system that contains the most complex component" item;
- the most complex components are the modules performing the measurements, including statistical. Such modules, in particular, may require a change in the concept of discretization in the model of the system under study.

In some cases for discrete modeling we can develop the recursive formulae that allow for each step to recalculate the statistical characteristics, both with accumulation, and with the use of sliding windows. It simplifies the application of the methods of sequential analysis of A. Wald as well. An important role for the component of the statistical processing is the creation of generators of the impacts. There are many problems associated with the "tails" of the distributions, correlation, etc. We confine ourselves here to the reference to quasi-random numbers, proposed in [10]. Deterministic quasi-random sequences of numbers in some cases may be a more appropriate and effective form of the imitation of random, rather than pseudorandom sequence obtained by physical and software generators. Additionally, we should mention the possibilities of the methods of designing of experiments. The establishment of the system of statistical processing in the simulation infrastructure allows to formulate and solve a critical challenge: to investigate the behavior of the component of the statistical analysis itself. Firstly, this allows to get confidence (or start to doubt) that the component is working correctly. Secondly, we can explore the behavior of a component in the case of violation of the prerequisites for its use. In particular, we can consider the impact of errors in the determination of the parameters to conclude the degree of agreement of the experimental and theoretical distributions. Sometimes

it helps to find the source of the error in the publications about the experiments that contain this error. Such virtual systems of statistical analysis will be a useful addition to an extensive arsenal of existing statistical methods.

2 Two-stage analysis of the goodness-of-fit

Two-stage processing of measurement results is metastatistics principle, consisting in the fact that the results of the first stage are the raw material for later. Its peculiarity consists in the fact that the results of the first stage are qualitatively different from the original information and can be independently interpreted. When you use goodness of fit tests of distributions of the phases can be described as follows [9]:

- computation of one value of criterion for each subset of the experimental data;
- analysis of the distribution of the obtained set of criterion values.

We explain this approach with the following example. We investigate a problem of determining the quality of operation of the analyzer of interference in the communication channel, where the distribution and the parameters of the interference depend on the time of the day. Let there be K realizations (data sets) of a certain observation. Each implementation contains in the general case, a different number N_k of data and is described by its individual theoretical distribution function F_k , $k = 1, ..., K$. The difference may lie in the values of the parameters and the type of distribution. The use of any standard criterion consent in case of correct operation of the analyzer will lead to the result that determined by the given confidence probability. If the threshold criterion was determined for a confidence level of 0.05, approximately 5% of the ideal implementations will be identified as wrong. And it is not the fault of the analyzer. In these circumstances, you can optionally use the following procedure. If for each of K realizations there are the computed values of the criterion (e.g., Kolmogorov), you should construct the empirical distribution of these values and then test the hypothesis that they have the Kolmogorov distribution. For other criteria you may need the exact marginal distribution. For the Kolmogorov criterion in the case of evaluation of distribution parameters by experimental data in [9] is proposed an approximation in the form of a generalized distribution of extreme values, and a table of the parameters of the approximation for several popular distributions as well.

3 Method of randomized survey

Randomization is a standard procedure for obtaining a sample for statistical studies and is used in various social processes. The example of the using of randomization for psychological protection of subjects is the penalty in the electric chair in the United States. There are several performers of the sentence who do not know, whose switch is the fatal one. There is also a method of randomization that is implemented only

in the conditions of the research subjects (respondents), capable to logical reflection. The uniqueness of this method is that it allows us to estimate the relative number of respondents with the attribute, which they usually deny. This capability is provided by a high potential anonymity of the respondents. The original method was proposed in 1965 [11], and later it was several times modified [1, 5].

The scheme of the method can be described as follows. We know total number N of the respondents in the group consisting of subgroups A and B. These respondents are not interested in revealing their membership to particular subgroup. The problem is to estimate the number of respondents in each subgroup, N_A and N_B respectively; $N_A + N_B = N$. A probabilistic experiment with two classes of possible outcomes C and D is proposed to the respondents. The probabilities of which is $P(C)$ and $P(D)$ are known and not equal 0.5; $P(C) + P(D) = 1$. Event C is associated with the assignment of the subject to subgroup A. Then the experiment is carried out. The results matched "YES" if the random event C (or D) fits to belonging the subject to the subgroup A (or B). Knowing the number of answers N_{YES} and N_{NO}, where $N_{YES} + N_{NO} = N$, we can obtain an estimate of the number of members of the group with characteristic A in the form $N_A = (NP(D) - N_{YES})/(1 - 2P(C))$. The mentioned experiments were as follows: the rotation of the roulette which are asymmetrically divided into sectors C and D; the throwing of dice or coins. In one modification, a respondent was throwing a coin either once or twice, and must gave only one response, so the researcher did not know if this is the answer to the main question, or the answer to the intermediate question.

Anonymity in all of these cases potentially was provided by promise of the experimenters about the absence of observation of a physical experiment. However, the capabilities of modern technical means allow to fix the outcome of any similar experiment without direct observation. Nevertheless, when you save a schematic of the method and calculation formulas absolute anonymity can be achieved through changes in the organization of the experiment. The choice events associated with a specific subgroup, passes to the reflexive system of the Respondent. This gives grounds to call this modification of the method of "reflexive randomized survey" [3]. The subject pre-selects for himself (and not informs anyone), what the outcomes of the experiment he connects with event C. For example, he (or she) can select one of the four suits of playing cards, one of the combinations simultaneously tossing two different coins, etc. Option 1:3 (one of four) is the easiest one both for calculation, and for the organization of the experiment. We emphasize that the method is workable only on the understanding position by all the participants of the whole process. To the author's experience, there is simple and effective experiment in which the researcher needs to ask four questions. The answers to the three demonstration questions are obvious for the participants, the fourth question is the main one. Here is the list of demo questions:

- Are you now in this audience? (The answer is definitely "Yes", the group B consists of 100% of respondents).

- Are you now outside this audience? (The answer is clearly "No"; 0%).

- Have you any neighbours at your right side? (The answer depends on the location of respondents in the audience and is easy calculated).

Respondents are convinced that estimations are obviously inaccurate, however, allow to assess the situation. The anonymity is obvious enough. It is also clearly the meaningless for the researcher to organize a surveillance and/or to distort the mechanism of generation of random events. The prospect of using this method in the organization of secret ballots in computer networks is rather important.The systems of distributed artificial intelligence, virtual agents and intelligent centaurs are the respondents.

4 Conclusions

Traditional descriptions of the statistical methods and their applications are scientific in the sense that the subjects are excluded from consideration and description. Sometimes this approach stops working, for example, in the field of medicine. In this case, one possible way of action is the consideration of the whole system, in which the formulation and solution of problems is performed. Understanding begins with a logical analysis or definition of the situation. The formation of such definition allows to clarify the nature of the uncertainty of this situation, and find ways of working with them. In the presence of inherent uncertainties it is possible to change the formulation of the problem, and, for example, begin to look for invariants or of the order parameters. Sometimes it can be possible to take a component of a specific situation with additional constraints and opportunities (as an example, the ultra-efficient estimations of L. Le Cam are). When using the measurement associated with the subjects, there is also the possibility of correction of the uncertainty. We will mention the phenomenon of perceptual bistability (hysteresis), which is described using catastrophe theory, "grey" and "black and white" scales, considered by D. A. Pospelov, and "the formula of human being" by V. A. Lefebvre, who describes the effects of categorization and origin of the "Golden section". The ability to manage uncertainty is well represented in [6]. In particular, it proved two theorems about the variety, according to which for any subject you can pick up some groups and sets of actions that ensure the desired behavior of the subject. It is noted that in some cases (war against terrorism) it is useful to provide the opponent with additional information that will make its behaviour more predictable. The variety of statistical methods is very large. The interaction of the Researcher with the Client is very similar to the interaction of the doctor with the patient. In both cases it is required to choose from the many sets only one set, that is adequate to the situation. In modern medicine there is a direction oriented to the production of medicines by the body of the patient itself. This approach is promising not only for medicine and education [4], but also to create effective technologies for the application of statistical methods in applied research.

References

[1] Adebola F.B., Johnson J.J. (2015). An Improved Warner's Randomized Response Model *International Journal of Statistics and Applications*. V.5(6): 263-267 DOI: 10.5923/j.statistics.20150506.01

[2] Filimonov V.A. (2014). Cartosemiotics on the "globe" of the cross-technologies of a situational center (In Russian) *Discussion posts to Cartosemiotics and to the theory of Cartography (Theoretical problems of Cartography and its neighbouring disciplines) / International correspondence seminar / Dresden*. V. 17, pp. 5-13.

[3] Filimonov V. A. (2014). Cross-technologies of situational center as a testing ground of Cybernetics (In Russian) *Mathematical structures and modeling*. No. 3 (31), pp. 99-108.

[4] Filimonov V.A. (2016). Proteins in the cage as the metaphor of cognitive pharmacy for training of mathematics (in Russian) *Actual problems of teaching mathematics in a technical university*. № 4, pp. 155-160. http://conf.nsc.ru/files/conferences/MathEducation-2016/fulltext/338766/338776/filimonov-v-a-omgtu-2016-2.pdf//

[5] Hussain Z., Shabbir J. (2008). Logit Estimation Using Warner's Randomized Response Model *Journal of Modern Applied Statistical Methods*. Vol. 7: Iss. 1, Article 11. Available at: http://digitalcommons.wayne.edu/jmasm/vol7/iss1/11

[6] Lefebvre V. A. (2010). *Lectures on the Reflexive Game Theory*. Leaf and Oaks Publishers.

[7] Lefebvre V. A. (2013). *What is animacy (in Russian)*. Kogito-centre, Moscow.

[8] Malinetskii G. G. (2016). *Mathematical foundations of synergetics: Chaos, structures, computational experiment. (in Russian)*.URSS, Moscow.

[9] Pollak Yu., Filimonov V. A. (1988). *Statistical computer simulation of means of communication (in Russian)*. Radio and sv'yaz, Moscow.

[10] Sobol I. M. (1969). *Multidimensional quadrature formulas and Haar functions (in Russian)*. Nauka, Moscow.

[11] Warner S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*. V. 60, N. 309, pp. 63-69.

Ontology Relations Completeness Evaluating Method Based on Formal Concepts Analysis Theory

Bato Merdygeev and Sesegma Dambaev
East Siberia State University of Technology and Management, Ulan-Ude, Russia
e-mail: mainisjusticeone@gmail.com, damseg@gmail.com

Abstract

Currently, many intelligent systems use ontology as a knowledge base. The effectiveness of such a system directly depends on the effectiveness of the knowledge presented in the ontology. Methods are needed to help the domain expert effectively evaluate the quality of the ontology. The paper presents developed approach and methods to the analysis of the domain ontology based on formal concepts analysis. The approach allows evaluating the completeness of ontology relations. We used a concept lattice and formal contexts to analyze the relations of ontology.

Keywords: ontology; relation; ontology analysis; ontology evaluation; completeness of ontology relations; completeness of ontology; formal concept analysis; concept lattice.

Introduction

The development and research of intelligent data mining techniques is a rapidly developing area. Many intelligent systems use ontology as a knowledge base. In computer science, the term "ontology" means the formal representation of knowledge. It is used as a form of knowledge representation of the real world, or part of it. [1]

Regardless of the type of ontology its creation is a laborious and expensive task. At the same time, there is a possibility to get inefficient or incorrect knowledge of ontology. To avoid it is necessary to evaluate the quality of ontology at every stage of its production.

There are many methods for analyzing and evaluating the quality of ontology [2–6]. Often, such methods have their own specific use, requirements for the structure of ontology, low level of automation, as well as a high level of participation of the domain expert, which leads to large expenditures of resources and time. The development and research of methods of automated ontology quality evaluation with minimal use of the domain expert is an actual scientific and technical task.

In this paper to analyze the ontology we propose an approach and methods based on formal concept analysis theory. The theory of formal concept analysis (FCA) [7–10] is an applied lattice theory. It can be used in the field of knowledge processing [11]. This approach allows analyzing the ontology structure based on relations of concept lattice.

1 Ontology analysis approach

1.1 Purpose of analysis

The use of the formal concept analysis theory makes it possible to evaluate the quality of the ontology knowledge structure. The structure of knowledge is represented by the structure of relations between terms of ontology. The purpose of the analysis of the developed approach is the completeness of the ontology relations.

The completeness of the ontology relations is indicator of integrity and consistency of the ontology relations. Ideally, the ontology is complete with respect to this indicator if all possible relations that are logically consistent with each other are presented in it, and all relationships that violate the integrity of the ontology's logical structure are missing. The completeness of the ontology relations shows how fully the knowledge of the relations between the terms of the domain is reflected in the ontology.

Ontology relations are classified into qualitative and quantitative relations [12]. In this approach only qualitative relations are considered.

1.2 Description of the approach

The basis of the analysis is to search for possible missing relations on the lattice using arrow relations.

The approach consists of the following steps:

1. Select the type of term relation that you want to analyze. Every relation type has its semantics, so the result of the analysis is interpreted according to the selected type.

2. Construct concept lattices for contexts where terms are taken as objects and attributes. [13-16] Depending on the relation type it is possible to use different methods of constructing a formal context to maximize the effectiveness of the analysis.

3. Search for discrepancies of the concept lattice and the structure of ontology relations. At this step, different methods can be used to analyze the concept lattice in order to evaluate the quality of the ontology. We use the developed methods:

    - ontology evaluation method based on a comparison of internal and actual structures of relations;
    - ontology evaluation method based on relations on the formal context arcs.

4. Result analysis. This analysis is performed by an expert. It determines the result of the analysis of the completeness of the ontology relations over the set of found inconsistencies and revealed hidden dependencies and result evaluation criteria for analysis methods.

This approach allows us to evaluate the completeness of ontology relations based only on data contained in the ontology, without using external sources. A search is made for internal hidden dependencies, which make it possible to determine the logicality of the available ontology relations. Also, hidden dependencies allow us to identify relationships that may need to be included in the ontology.

A domain expert is provided with a variety of situations found. The expert on the basis of the received data makes the final decision on each situation (to include or delete the relation from ontology). Thus, the expert does not waste time searching for suitable situations verifiable relations.

To accelerate the analysis, automatic decision-making is possible on the basis of special criteria for analyzing situations with relations. It is also possible to obtain an total value of the completeness of ontology relations, according to which the structural effectiveness of ontology knowledge is assessed.

2 Ontology evaluation method based on a comparison of internal and actual structures of relations

2.1 Categories of ontology terms

The main purpose of this method is to search for inconsistencies in the actual structure of the ontology relations with the concept lattice for this type of relations, i.e. relations that indicate these inconsistencies. A detailed description of the method is contained in [16].

The object of analysis of this method is an ontology constructed on the basis of the structure of the ontology described in [18]. Ontology in this case is presented in the form of related groups of terms, divided into categories: concept, action, state, event, property, quantity.

Relations between the terms of these categories can be divided into two types:

- relations of one category (Concept-Concept);
- relations between categories (Concept-Action).

The method makes it possible to compare ontological relations with the revealed structure of relations by means of a concept lattice. This allows us to identify the presence of some relations that may need to be included in the ontology.

2.2 Relations of one category

These relations include such types of relations as “Concept-Concept”, “Action-Action” and etc.

Let the set of relations of a particular type of source ontology is T_O, and the set of relations derived from a concept lattice of relations of the same type is T_R. Then the set of inconsistencies relations of one category is defined as

$$N_T : T_O \backslash T_R \cup T_R \backslash T_O. \quad (1)$$

It should be separated by a set of lattice inconsistencies ($N_{TR} : T_R \backslash T_O$) and ontology inconsistencies ($N_{TO} : T_O \backslash T_R$) as they may have a different weight in determining the completeness of the ontology relations.

As a result, we get a lot of incredible inconsistencies. Such inconsistencies can be a great multitude, which may confuse the expert.

2.3 Relations between categories

These relations include such types of relations as "Concept-Action", "Action-Concept", "Concept-State" and etc.

Let the union of qualitative relations of one category of original ontology is $\cup T_{Oi}$, and the set of relations derived from the concept lattice of relations of the same type is A_R.

Then the set of inconsistencies of the relation is defined as

$$N_A : (\cup T_{Oi} \backslash A_R) \cup (A_R \backslash \cup T_{Oi}). \quad (2)$$

It should be separated by a sets of lattice inconsistencies ($N_{AR} : \cup T_{Oi} \backslash A_R$) and ontology inconsistencies ($N_{TO} : A_R \backslash \cup T_{Oi}$) as they may have a different weight in determining the completeness of the ontology relations.

Thus, inconsistencies, which are available to the expert for consideration, determined by

$$N : N_T \cap N_A. \quad (3)$$

As a result, the expert receives the set is not appropriate for the two parameters of relations. This allows it to draw a conclusion about the completeness of the ontology relations.

The criterion of evaluation in this case can be considered the number of found inconsistencies of different types.

3 Ontology evaluation method based on relations on the formal context arcs

3.1 Relations on arcs of formal context

On the basis of [9] we define the concepts of relations on arcs of a formal context.

The formal context K is a triple $< G, M, I >$, where G, M are sets and $I \subseteq G \times M$ is the binary relation between G and M. The elements of G are objects, the elements of M are attributes, and I is the incidence of the context $< G, M, I >$.

$$A' := \{m \in M | \ (g, m) \in I, \forall g \in A\}, \ where \ A \subseteq G \quad (4)$$

$$B' := \{g \in G | \ (g,m) \in I, \forall m \in B\}, \ where \ B \subseteq M \tag{5}$$

A pair (A, B) is a formal concept of $< G, M, I >$ if and only if

$$A \subseteq G, \ B \subseteq M, \ A' = B \ and \ A = B' \tag{6}$$

A and B are called the extent and intent of the formal concept (A, B), respectively.

The arrow relations [9] of the formal context $< G, M, I >$ are defined as follows: for $g, \ h \in G$ and $m, \ n \in M$, let's say:

$$g \swarrow m :\Leftrightarrow (g,m) \notin I \ and \ if \ g' \subseteq h' \ and \ g' \neq h' \Rightarrow hIm \tag{7}$$

$$g \nearrow m :\Leftrightarrow (g,m) \notin I \ and \ if \ m' \subseteq n' \ and \ m' \neq n' \Rightarrow gIn \tag{8}$$

$$g \updownarrow m :\Leftrightarrow g \swarrow m \ and \ g \nearrow m \tag{9}$$

For a given $g \in G$, there is an attribute $m \in M$, marked by $g \swarrow m$ if and only if γg is $\vee$-irreducible. Dual $g \nearrow m$ for the same $g \in G$ and only if μm is $\wedge$-irreducible.

On the basis of these types of relations a conclusion is made about possible missing of the necessary relations.

3.2 Search for possible missing relations

The criterion for determining the possible missing relations between an object and an attribute of the concept lattice is the number of arrow relations.

To determine the set of possible missing relations, it is necessary to search for all arrow relations between unrelated objects and attributes of the lattice.

On the basis of formula 7, the set $S_{\swarrow}(g,m)$, consisting of terms by which the relation $g \swarrow m$ is defined, is determined.

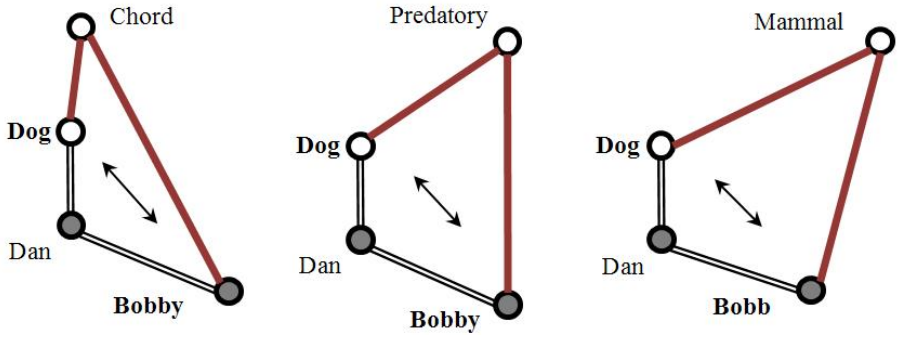


Figure 1: Relation $Bobby \updownarrow Dog$

$$S_{\swarrow}(g,m) = \{h : \ (g,m) \notin I \ and \ if \ g' \subseteq h' \ and \ g' \neq h' \Rightarrow hIm\} \tag{10}$$

On the basis of formula 8, the set $S_{\nearrow}(g,m)$, consisted of terms through which the relation $g \nearrow m$ is determined, is determined.

$$S_{\nearrow}(g,m) = \{n : \ (g,m) \notin I \ and \ if \ m' \subseteq n' \ and \ m' \neq n' \Rightarrow gIn\} \tag{11}$$

On the basis of formula 9, the pair $(g,m)_{\updownarrow}$ satisfies

$$S_{\updownarrow}(g,m) = (S_{\swarrow}(g,m), S_{\nearrow}(g,m)) \tag{12}$$

and represents all the terms that support the possibility that between g and m there must be the relation.

The found sets allow assuming that in the construction of ontology some relation elements were omitted. Also it can allow deducing new knowledge from concept lattice on certain type of relations. These data are provided by domain experts.

A set of all found terms can be used to evaluate the completeness of ontology relations.

$$S_K = \{S_{\updownarrow}(g,m) : \forall g \in G, \forall m \in M\},\ where\ S_{\swarrow}(g,m) \neq \emptyset\ and\ S_{\nearrow}(g,m) \neq \emptyset \tag{13}$$

S_K is the set of pairs that support possible missing relations. For each type of relations, a separate set of S_K is generated.

To reduce the number of unimportant relations, we should use threshold values for $S_{\swarrow}(g,m)$ and $S_{\nearrow}(g,m)$ before $S_{\updownarrow}(g,m)$ is included in the set S_K.

The criterion of evaluation in this case can be considered the number of pairs in sets S_K for each type of relations.

4 Combination of methods

Based on the methods presented, we can derive a general method of analysis. Steps of general method:

1. Separation of terms into categories. If there is a categorical apparatus in the ontology, then it is used in the method. Otherwise, the separation can be made based on the types of relations associated with the terms. In this case, you need a domain expert or an automatic classification program based on relation types.

2. Using first method "Ontology evaluation method based on a comparison of internal and actual structures of relations":

   (a) Construction of concept lattices by types of relations taking into account categories of terms.

   (b) Search for inconsistencies by the first method of analysis. Definitions of sets of inconsistencies for each type of relations and category of terms.

3. Using second method "Ontology evaluation method based on relations on the formal context arcs":

   (a) Analysis of the constructed lattices constructed during the analysis by the first method, using arrow relations to determine possible missed relations.

(b) Definition of sets of the found situations by types of relations.

4. Determination of the correspondences of the found situations by the first and second methods. The received data can be analyzed by the domain expert. Also, based on these data, you can determine the overall indicator of the completeness of the ontology relations.

Conclusions

Formal concept analysis provides additional opportunities for analysis of the ontology. Formal contexts and concept lattices allow us to divide the concepts of this ontology into separate relations, which provides a more detailed analysis of ontological knowledge.

The proposed approach to the analysis of the completeness of ontology-based lattice relations allows us to identify hidden relations between terms and provide them to the expert for further evaluation. Thus, it is possible to automate the search for possible inconsistencies in the ontology and subject area, which speeds up the work of the expert, and also allows us to identify hidden knowledge that can be derived from the original ontology.

These methods of evaluating the structure of ontology are based on one approach to evaluation and it is possible to combine them for more efficient analysis. These methods help the expert make decisions while analyzing the knowledge and structure of the ontology.

At the moment, the approach is being actively developed. It passed the tests on test ontologies with a small amount of data, which showed sufficient efficiency.

References

[1] Klesheva A.S., Shalfeeva E.A. (2005). Classification of ontology properties. Ontologies and their classification. *In: IACP*, 20 p., Preprint, Vladivostok.

[2] Dambaeva S., Busovikov P. Overview of ontology analysis methods. *Innovative Information Technologies: Matetials of the International scientific – practical conference*

[3] Lozano-Tello A. , Gomez-Perez A. (2004). Ontometric: A method to choose the appropriate ontology. *J.Datab.* Mgmt., 15(2): pp. 1–18.

[4] Spyns P. (2005). EvaLexon: Assessing triples mined from texts. *Technical Report 09*, STAR Lab, Brussels, Belgium.

[5] Gangemi A., Catenacci C., Ciaramita M., Lehmann J. (2005). Ontology evaluation and validation. An integrated formal model for the quality diagnostic task. *Laboratory of Applied Ontologies*, CNR, Rome, Italy.

[6] Babkin E., Nabiullin O. (2007). Ontology analysis based on the selection of semantically related entity groups. *HSE Academic Fund Programme*.

[7] Priss U. (2007). Formal Concept Analysis in Information Science. *Annual Review of Information Science and Technology*. Vol. **40**, pp. 521-543.

[8] Wille R. (2006). Formal Concept Analysis as Applied Lattice Theory. *Concept Lattices and Their Applications*, Springer, pp. 42-67.

[9] Ganter B., Wille R. (1999). Formal Concept Analysis: Mathematical Foundations. *Springer*.

[10] Wolff K. (1993). A first course in formal concept analysis. *In: F. Faulbaum editor, StatSoft '93*, Gustav Fischer Verlag, pp. 429-438.

[11] Wille R., Stumme G., Wille U. (1998). Conceptual Knowledge Discovery in Databases using Formal Concept Analysis. *Springer*. Verlag Berlin–Heidelberg.

[12] Naihanova L. (2008). The main types of semantic relationships between domain terms. *Proceedings of higher educational institutions. The Volga region №1*. Technical science. Informatics, Computer Science and Management, pp. 62-71.

[13] Stumme G., Maedche A. (2001). FCA-MERGE: Bottom-Up Merging of Ontologies. *IJCAI'01 Proceedings of the 17th international joint conference on Artificial intelligence*.

[14] Stumme G., Cimiano P., Hotho A., Tane J. (2004). Conceptual Knowledge Processing with Formal Concept Analysis and Ontologies. *ICFCA 2004: Concept Lattices*, pp. 189-207.

[15] Wille R. (1982). Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. *In: I. Rival (ed.): Ordered sets*. Reidel, Dordrecht-Boston, pp. 445-470.

[16] Priss U. (2000). Lattice-based Information Retrieval Knowledge Organization. Vol. **27**, 3, pp. 132-142.

[17] Merdygeev B., Dambaeva S. (2016). Evaluation of Ontology Quality based on Analysis of Relations in Concept Lattices. *CDUD 2016*, pp. 74-86.

[18] Naihanova L. (2008). The technology of creating methods for automatic construction of ontologies with the use of genetic and automatic programming. *Monograph*, Ulan-Ude.

Map as a Basis for Decision-Making in the Automated Learning Process

Victor Uglev, Sergey Cholodilov and Valeriya Cholodilova
Siberian Federal University, Zheleznogorsk, Russia
e-mail: uglev-v@yandex.ru

Abstract

Decision making in intelligent learning systems is based on the processing of a multitude of factors about the training course, models of the learner, learning situation and specific methodological knowledge (model teacher, methodologist and subject tutor). Their ordering and a compact representation in a single space will simplify the process of finding compromise solutions that take into account the subjective goals of each model. This paper proposes to use the method of mapping, on the basis of which to build a multidimensional Cognitive Maps of Knowledge Diagnosis, represents Atlas individual learning activities. To build and analyze maps in each case tested a number of hypotheses based on small statistical series of data and based on the use of nonparametric methods. One of the key methods of testing hypotheses in this approach is the method Shortliffe-Buchanan. To illustrate its application in testing hypotheses examples related to the level of formation of cognitive maps of diagnostic knowledge and its processing (analysis) in making decisions the solver learning system.

Keywords: Automation Educational System, Individualization, Decision-making, Cognitive Map of Knowledge Diagnosis, Shortliffe-Buchanan method.

Introduction

Intelligent automated educational systems (IAES) are the most advanced automation tools in distance and e-learning process. Such systems are forced to form and use the image of the pupil in order to take the most appropriate decisions in support of the individual learning process. Factors included in the parametric model of the student and the e-learning course, quite a lot, and their contribution to the process of making various decisions is nonlinear. A special method of "convolution" of the factor space, pre-concentrating the desired settings, would greatly simplify the work scheduler training system. Therefore, we consider the solution of this problem through the use of mapping technology, involving the inspection of a whole class of hypotheses, separately focusing on the use of the individual criteria nonparametric statistics.

The small volume of statistical data and the rather weak formalization of the object of influence (the student) do not allow in most cases IAES to rely on analytical laws of distributions of certain characteristics of the trainee (level of knowledge, competence level, personal importance of educational material, etc.). For this reason, the processing of student interaction protocols with IAES is mainly based on nonparametric methods for testing hypotheses [7] or heuristics [6]. It is obvious that the use of methods for nonparametric statistics for the scientific substantiation

of certain initial hypotheses is more preferable than to rely solely on a system of heuristic rules (for example, product rules) or the method of fuzzy logic. Thus, at the stage of processing the initial data extracted from the IAES protocols by an intelligent scheduler (decision making subsystem) during the formation and processing of the map, it becomes necessary to organize a check of a set of hypotheses so that decision making can be quickly implemented and explained

1 S1 Mapping in learning space of material

The development of individualized solutions by the IAES planner is based on a model of the learning process, in which the following stages of training can be distinguished: the formation of an individual course, the work with theoretical didactic material, the implementation of exercises, intermediate and final control (the first and last items occur once, others - iteratively). At the same time, individualization affects the composition of the educational material, the basic trajectory of its study, the composition and norms of evaluating the control material, as well as the content of the dialogue with the student [9]. Based on the methodology of the system approach [8], the process of automated learning should be considered as a joint “activity” of the student (including his model in IAES) and a system that includes the following models: teacher, methodologist and subject tutor [14]. Obviously, all these models have their own specific decision-making and can enter into a “conflict” when making decisions about a particular hypothesis. Conflict resolution requires that all of them rely on a single space of initial factors, minimized in composition to simplify and accelerate the decision-making process.

The space of knowledge about the elements of the learning process, like any system of knowledge, has metric characteristics [2]. At present, the development of a whole trend in semiotics - map semiotics is observed, which can also be successfully applied to objects of a non-geographic nature [13]. In this case, the cards perform a number of functions: folding large data sets, visualization, transfer of knowledge, - based on which you can solve the task of navigation, i.e. decision making. It turns out that from the existing space of factors it is necessary to allocate such a basis that will be invariant for all mentioned models from the IAES, and can be the basis for the concentration of the interesting subsets of factors used in decision making.

Allocating from the totality of the knowledge that is involved by the IAES planner in the situational solution of the "conflict" between the models makes the map the base on which, as in real situational centers, the possibility of finding compromise solutions is being discussed. Proceeding from this, we take as the basis (invariant) for the map the totality of the structural and semantic parameters of that didactic material, which is the basis of the electronic course. Obviously, for any student, regardless of his preferences, the formation of an individual educational trajectory begins with a typical (basic) working plan of the discipline, gradually acquiring the form of an individual one. We note a number of properties of the map describing the space of didactic material: metricity, scalability, vectorality [18].

As a visualization method, we use the "small world" model in which the can-

vas of the circle defines the basic sequence of studying didactic material, and the links between its elements are semantic dependencies within the electronic course. In addition, we will outline the main aspects that should be reflected on the map to ensure its tasks related to supporting the decision-making process: normative, knowledgeable, competent (in terms of components and developmental levels), subjective (model roles), Trajectory (including the dynamics of changes).

All of them can be reflected in different scales (from individual indicators to group scores on a set of students). Next, let's move from a general idea to a solution that could be implemented from the practical point of view as part of the IAES and used to make decisions based on various methods for testing hypotheses (including nonparametric ones).

2 Cognitive Map of Knowledge Diagnosist

Cognitive map of knowledge diagnosis (CMKD) is a map of the structure of didactic material, reflecting the result of the stage of summarizing information in the logic of the IAES work, automatically preparing (translating) information about the learning process into knowledge, in order to simplify the complex expert analysis of the learning situation and to develop an adequate response to the actions of the student. CMKD is focused on visual representation, drawing those aspects that are needed at the moment for decision making [10]. At the initial stage (before the questionnaire), the course map will be based on the basic composition of the discipline (Figure 1, left side), but then a transition to an individualized form should take place (Figure 1, right side). The elements of the charted space are the didactic units themselves, the basic trajectory of learning and maps (color marks the elements of the core of the discipline), as well as the semantic links within the educational material. The process of working with the map can be described in three stages: the synthesis of the base map, the formation and coordination of an individual map, decision making and visualization on the basis of an individual map that updates its content relative to the current learning situation (in dynamics). Thus, the first and second stages are necessary to ensure the first stage of the IAES work, and the third - to improve the quality of the implementation of subsequent stages [12]. Relying on the individualized composition of the electronic course, which is the metric basis for the map, other parameters from the learner model that are necessary for making decisions in the search for compromises between the scheduler models are superimposed on it. Thus, a set of maps or an atlas, characterizing various aspects of the process of personal learning at a given time (hereafter an atlas of individual learning activity) is obtained. Obviously, the map is oriented to accompanying individualized training, but in some cases it can also be fashioned to analyze group indicators. As an example, consider the process of testing a number of hypotheses that are verified in the process of CMKD formation using nonparametric statistics methods.

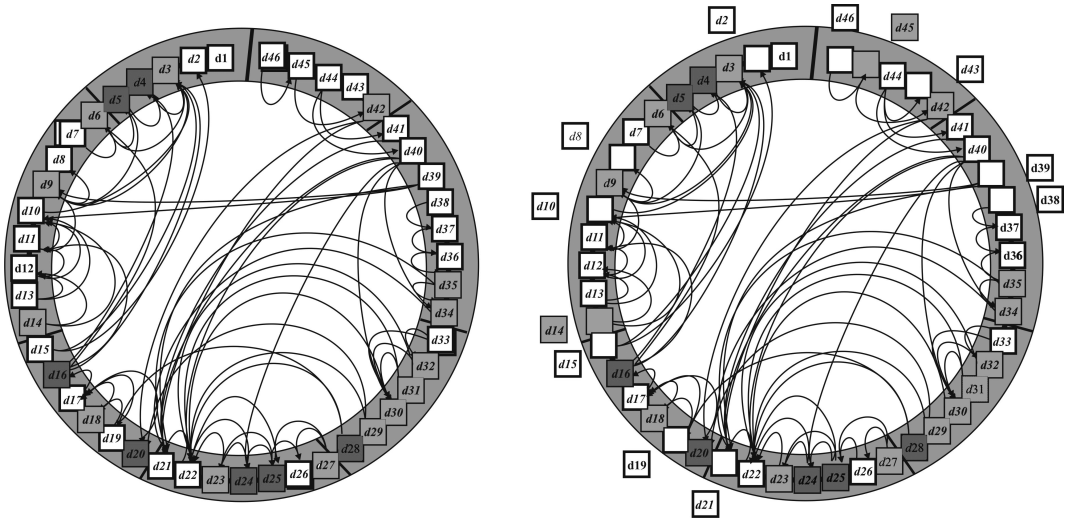


Figure 1: An example of CMKD rendering for the basic (left side) and individualized (right side) composition of the electronic course (normative aspect) of the discipline "Simulation"

3 Decision-making by Shortliffe-Buchanan nonparametric method

Nonparametric methods of statistics are widely used in pedagogical experiments and are usually aimed at checking the effectiveness of the application of certain innovative teaching methods [7], [4]. We will be interested in testing the hypotheses used in the construction of the CMKD as well as in making decisions with it. Conditionally, all processes for working with the card can be divided into two classes: those participating in the formation of the map, i.e. The processes that take place in order to generalize the initial data and reveal regularities, and participate in the use of the map data for decision-making, i.e. Conducting for the purpose of testing any hypotheses on the basis of an already existing map. We illustrate them using the method of Shortliffe and Buchanan [3]. Verification of hypotheses based on the result of analysis of a set of statistical data (evidence), as a rule, is of a probabilistic nature and requires formalized criteria. In the method of Shortliffe and Buchanan, the CF (certainty factor) confidence coefficient is calculated, which belongs to the interval from -1 to 1, where 0 is the uncertainty value. The quantitative estimation of CF, after a number of transformations according to [1], is carried out according to the formula (1).

$$CF\left[H/x\right] = \begin{cases} \frac{P(H/x)-P(H)}{1-P(H)} & if \quad P(H/x) \geq P(H) \\ \frac{P(H/x)-P(H)}{P(H)}, & if \quad P(H/x) < P(H) \end{cases}, \tag{1}$$

where $P(H/x)$ is the conditional probability that the hypothesis H is true for a series of evidence x, and $P(H)$ is an unconditional probability. For situations where

the number of certificates is small (less than 30), and a priori assessments of evidence allow estimating the CF to calculate the empirical value of the confidence and mistrust measure, it becomes possible to test several hypotheses in parallel on a common set of statistics. Next, consider two examples of the application of the Shortliffe and Buchanan method, illustrating both classes of processes when accompanying the IAES electronic course "Simulation Modeling" for students. The student records on the course and passes the survey, according to which the planner should formulate a proposal on the individual structure of the work program (the first stage of training). Let there be a number of answers from the X student to the question "On which didactic units of the course do you prefer to make an emphasis in teaching?" (Answers "Make an accent", "The meaning of the depth of study by default", "Do not make an accent" and "Do not know" 49 items [19]), supplemented with a matrix of expert judgments on four hypotheses reflecting the layout of the educational material: an ordered set H, including the options "Introductory course", "Basic course" (this is the default), "Practical course "," specialization courses "(depth). The parameters of each didactic unit are set in the e-course model (importance in the course, complexity, belonging to the core of the discipline, the basic depth of study and control), and there is also the result of choosing one of these hypotheses when answering the corresponding question during the questionnaire. Then, calculating the values ??of CF for each of the hypotheses, you can check a number of statements: "To what extent do the student's accents coincide with the preferred form of the e-course layout?", "Should the student's wishes be taken into account in didactic units?", "Is it necessary Initialize the dialogue to resolve the contradictions? ", etc. The result of testing the hypothesis of the subjective importance of an individual didactic unit for the student according to the method of Shortliffe and Buchanan will be its quantitative evaluation (the coefficient is sure spine). For example, suppose an ordered set of didactic units $D = d1, d2, d3, d4$ and a vector with order numbers of answers $X = 2, 3, 1, 1$ are given. Then, using the matrix of a priori expert assessments (see the table) and using formula (1), we can calculate the confidence factor of the scheduler in the preferred e-course layout for the student. It is also known that the student in the questionnaire explicitly noted the preference for the hypothesis h_2, i.e. I chose the "Basic Course." If we take into account that the questionnaire is not 5 and 49 points, then the final estimates of each of the hypotheses will take the following values: h_1 = 0.98, h_2 = 0.71, h_3 = -0.1, h_4 = 0.2. It can be seen that there is some discrepancy in the assessment of the depth of the course by the student and the planner. This can be both a basis for initializing the dialogue, and for drawing an additional series of parameters from the student model to the analysis [11].

Table 1: Fragment of the table of expert opinions on the manifestation of hypotheses from H

Displaying the results of the individual significance of the didactic course material on the map will look like it is shown in Fig. 2, left side (the structural aspect is

	h_1				h_2				h_3				h_4			
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
d1	1	1	0	1	0	0	0	0	1	1	0,5	0,5	1	0,5	0,1	0,5
d2	1	1	0	1	0	0	0	0	1	1	0	0,5	1	0,5	0,1	0,5
d3	1	1	1	1	1	0	0	0,5	1	1	0,3	0,5	1	0,5	0,1	0,5
d4	1	0,6	0	0,5	1	0,5	0	0,6	1	1	0,1	0,5	1	0,1	0,1	0,5

represented by the student model). Here you can see how, based on the basic version of the map, a new page of the atlas of individual learning activity is formed by imposing student preferences, which can be involved in decision-making [15].

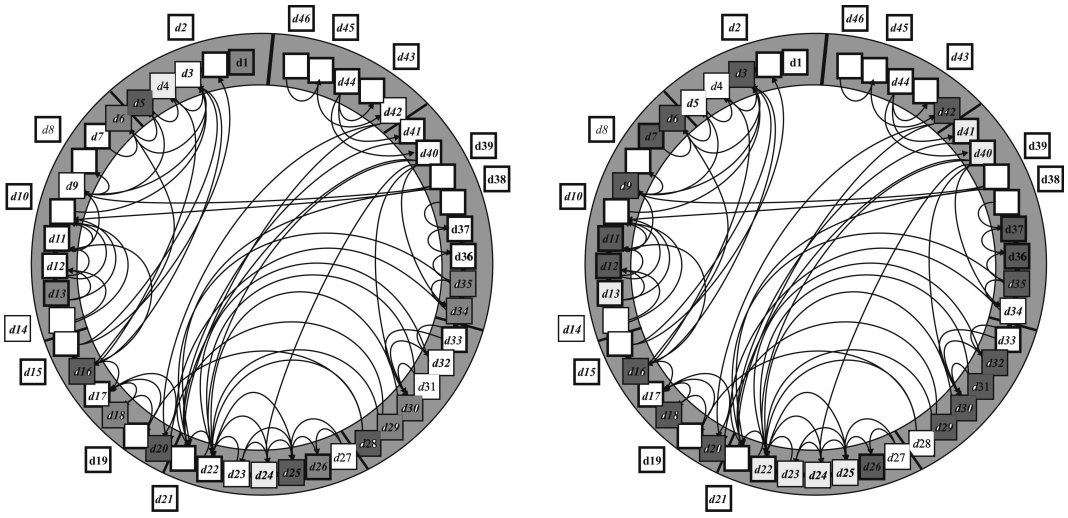


Figure 2: Example of rendering CMKD the emphasis in the course structure from the model of the pupil (left side) and the results of the evaluation of the level of development of competences on the input control (right side) for the discipline "simulation" of one of the students

Now consider an example of the application of the method of Shortliffe and Buchanan to the processing of statistics of responses to a series of tasks from the control and measurement material of the course. Since it is not difficult to extract information on correct answers and generalize it, we consider the competence component of the statistical series. To do this, we will assess the level of development of each of the competencies tested. For example, let the set of competencies ($H = 7$) of competences and a set of tasks from control and measurement materials be given, the answers to which can be reduced to an ordered vector of values [16]. Then the hypothesis testing about the level of development of one or another competence in a particular student can be reduced to application (1) if an a priori set of expert estimates is formed (by analogy with the table). Individual results of the application of the methodology from [20] for assessing the level of competence development are shown in Fig. 3 in the form of an individual competence profile (all axes are normal-

ized). But they have more importance for the further management of the learning process when they are reflected in the CMKD: in Fig. 2 (right side) the corresponding map from the atlas of individual educational activity is reflected. In this image, not only the metrical, semantic and competence characteristics of the learning process are seen, but also the dynamic characteristics of the process. This allows us to approach the verification of a number of concomitant hypotheses: "What didactical material should be studied first in order to increase the effectiveness of the development of h_i competence?", "How to respond to the student's request for a level of success with an emphasis on estimating the hypothesis h_i?", "On What kind of competence to emphasize when motivating a student in the dialogue process, if it is necessary to increase the importance (pay more attention) of the didactic material dn? ", Etc. Verification of such hypotheses is the basis for the process of managing the individual trajectory of training used in the IAES of the fourth generation [14].

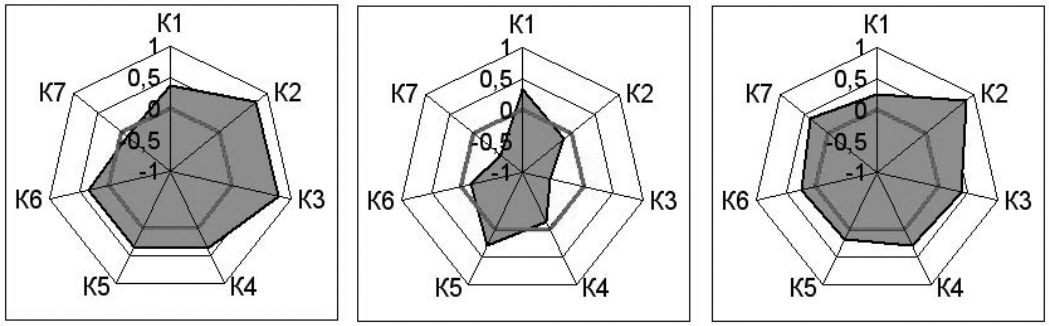


Figure 3: Individual competency development indicators of three students

In addition to revealing certain regularities, directly following the statistical (protocol) data, it is possible to test rather complicated hypotheses using the mapping mechanism. Let's give some formulations:

- Is the student's behavioral strategy consistent with the stated (personal) learning objectives?
- Is there a reaction of subconscious rejection of educational material (the so-called "cognitive immunity" [5])?
- Does the learner seek to "deceive" the teaching systems purposefully?
- Is the current tactic of interacting with the learner (reflective effect) effectively chosen by the solver?
- Should the model of the tutor defend the accents of the model of the student in the event of a "conflict of interests" when planning the composition of the work program of the discipline?

Many of these hypotheses involve methods of parametric and nonparametric statistics and allow me to synthesize a rather complex scheme of dialogue, deeply

individualizing the impact on the student. In this way, an effective feedback is obtained between the object of control - the student and the subject of control, i.e. Intellectual learning system [17].

Conclusions

The experimental approbation of the ideas presented in the article showed a sufficiently high effectiveness of the application of cognitive knowledge diagnostic maps in the development of individualized effects on the student in the process of training on the basis of intelligent automated learning systems. Despite the difficulties in typing the volume of the initial statistics (especially at the initial stage of training) and its "non-parametric" character, the combination of the corresponding statistical and heuristic methods of hypothesis testing made it possible to implement a "compromise" management of the educational process. A significant contribution to the implementation of the procedures was made by the method of Shortliffe and Buchanan, used to test hypotheses at all stages of the learning process (including the process of forming a cognitive knowledge diagnostic map). Synthesis of a set of maps in the form of an atlas of individual learning activities contributes to the formation of a compact and, at the same time, expressive presentation of the learning situation suitable for the automated version of the management of the individualized learning process.

References

[1] Artemova S.V., Pavlov V.I., Artemov A.A., Muromtsev D.Y. (2013). Decision support by ergatic element in uncertainty conditions by Shortliffe-Buchanan method. *Vestnik TSU*. Vol. **18**, part. 4, pp. 1435-1439.

[2] *Artificial intelligence* (1990). In 3 books. Reference. Under.ed. Popova E.V., Pospelov D.A., Zakharov V.N., Khoroshevsky V.F. Radio and communication, Moscow (in Russian).

[3] Buchanan B., Shortliffe E. (1984). *Rule-based Expert Systems*. Rdio and communication, Moscow (in Russian).

[4] Eremina S. (2015). The influence of education environment on the student's knowledge level based on the PISA survey data. *Applied Methods of Statistical Analysis. Nonparametric Approach - AMSA'2015*. NSTU, Novosibirsk, pp. 302-306.

[5] Filimonov V.A. (2015). Cognitive immunity as a problem and resource of information technology. *Mathematical modeling and computer simulation: proceedings of the III International scientific conference*. OmGTU, Omsk, pp. 149-150 (in Russian).

[6] Gavrilova T.A., Horoshevskiy V.F. (2001). *Knowledge Base of intellectual systems*. Piter, St. Petersburg (in Russian).

[7] Grabar M.I., Krasnyanskaya K.A. (1977). *The application of mathematical statistics in pedagogical research. Nonparametric methods*. Pedagogica, Moscow (in Russian).

[8] Tarasenko F.P. 2010. *Applied system analysis*. KnoRUS, Moscow (in Russian).

[9] Uglev V.A. (2010). On the specificity of individualization of training in Automated Training Systems. *Philosophy of Education*. Vol. **2**, pp. 68-74. Available at: www.phil-ed.ru/Text/NamberJourn.html (in Russian).

[10] Uglev V.A. (2012). The Cognitive Maps of Knowledge Diagnosis. *Open and Distance Learning*. Vol. **4**, pp. 17-23. Available at: ido.tsu.ru/other res/pdf/48-4.pdf (in Russian).

[11] Uglev V.A. (2013). Detection and intensification regularities while protocols to work of automated educational systems processing with the synthesis of Cognitive Maps of Knowledge Diagnosis. *Problems of Governance*. Vol. **6**, pp. 108-121 (in Russian).

[12] Uglev V.A. (2014). Implementation of Decision-making Methods in Intelligent Automated Educational System Focused on Complete Individualization in Learning. *AASRI Procedia*. Vol. **6**, pp. 66-72. DOI 10.1016/ j.aasri.2014.05.010.

[13] Uglev V.A. (2014). Variety of maps in the scientific cognition: between semiotics and cartosemiotics. *Geocontext: science almanac*. Vol. **2**, pp. 30-38 (in Russian).

[14] Uglev V.A. (2015). New generation of Intelligent Automated Educational Systems: main attributes and principles of organization. *Advanced methods and tools of intelligent systems: proceedings of all-Russian scientific-practical seminar*. NSTU, Novosibirsk, pp. 37-40 (in Russian).

[15] Uglev V.A. (2015). The possibilities of Individualization of Learning in the Automated Educational Systems with Cognitive Map of Knowledge Diagnosis. *Advanced methods and tools of intelligent systems: proceedings of all-Russian scientific-practical seminar*. NSTU, Novosibirsk, pp. 41-45 (in Russian).

[16] Uglev V.A., Dobronets B.S. (2017). Methodics of automatic measurement and estimation of the level of competences development. *Informatics and Education*. Vol. **2**, pp. 61-65 (in Russian).

[17] Uglev V.A., Filimonov V.A., Mishkina N.Yu. (2015). Hybrid approach to the management by feedback when working with Automated Educational Systems. *International Siberian Conference on Control and Communications (SIBCON)*. OmGTU, Omsk, pp. 1-4.

[18] Uglev V.A., Kovaleva T.M. (2014). An application of cognitive visual representation as a tool to support individual education. *Science&Education*. Vol. **3**, pp. 420-449. Available at: http://technomag.bmstu.ru/ doc/700661.html (in Russian).

[19] Uglev V.A., Ustinov V.A. (2011). *Simulation*. SFU publisher, Abakan (in Russian).

[20] Uglev V.A., Ustinov V.A. (2014). The new competencies development level expertise method within Intelligent Automated Educational Systems. *Advances in Intelligent Systems and Computing*. Vol. **293**, pp. 157-164. DOI 10.1007/978-3-319-07476-4_19.

Boosted Ensemble of the Nadaraya-Watson Estimators in Regression Task on the Boundary of the Feature Space

Ekaterina Mangalova and Olesya Chubarova
Siberian State Aerospace University, Krasnoyarsk, Russia
e-mail: e.s.mangalova@hotmail.com, kuznetcovao@mail.ru

Abstract

Nowadays, there is considerable interest in using ensembles learning technique to improve the performance of a single learner. Different ensemble learning methods are used in combination with one of the various single learners to avoid some negative effects of this learner implementation. This paper deals with the boundary problem of the Nadaraya-Watson estimator. The boosted ensemble of the Nadaraya-Watson estimators is proposed to solve this problem. Experimental results showed the high efficiency of ensemble approach in regression task on the boundary as well as in the middle of the feature space.

Keywords: Boosting, ensemble, nonparametric estimator, cross-validation, bandwidth, regression.

Introduction

Nowadays, there is considerable interest in using ensembles learning technique to improve the performance of a single learner. Most ensemble methods can be divided into two separate methods: bootstrap aggregating (bagging) and boosting.

Bagging is class of ensemble learning algorithms that fits base regressors each on random subsets of the original dataset and then aggregates their individual predictions (for example, by weighted averaging) to form a final prediction. Bagging is designed to improve stability and accuracy of a single learner. It allows to reduce variance and avoid overfitting [1].

Boosting is class of ensemble learning algorithms which convert weak learners to strong ones [2]. A weak learner is a predictor which predictions can be only slightly correlated with the actual responses (for example, it can be slightly better than mean value). In contrast, a strong learner is a predictor which produces predictions well-correlated with the actual responses. Boosting methods construct a composite learner by sequentially training weak learners. New weak learner is fit on dataset that has been transformed using the previously fitted ensemble. This has the advantage that only data that is likely to result in improved generalization performance is used for training [3].

In [4], [5] significant improvements of ensemble accuracy in comparison to a single base learner accuracy were shown on several machine learning benchmark problems.

Different ensemble methods are used in combination with one of the various single learners (nonparametric estimators, artificial neural networks, regression trees, linear

and polynomial regression, etc) to avoid some negative effects of this learner implementation. Some combinations "ensemble learning method - base learner" can be ineffective. For example, a critical factor in whether bagging will improve accuracy is the stability of the procedure for constructing. Improvement will occur for unstable procedures where a small change in training subsets can result in large changes in predictions. In [6] it is pointed out that artificial neural networks, regression trees, subset selection in linear regression are unstable, while Nadaraya-Watson estimator and k-nearest neighbors are stable. It means that only boosting technique can improve the performance of the Nadaraya-Watson estimator. In this paper we propose the boosted ensemble method which allows avoiding bias problems at or near the boundaries of the feature space.

The paper is organized as follows. In the next section we discuss accuracy problem near the boundary of the feature space. In the second section we introduce algorithm of this problem solving using ensemble technique, discuss some validation aspects and algorithm's computational properties. In the third section we provide experimental results of boosted ensemble of the Nadaraya-Watson estimators implementation. The paper finishes with a conclusion and perspectives for future work.

1 Problem description

The Nadaraya-Watson estimator is widely used technique for solving regression task:

$$\hat{y}\left(\bar{x} \mid G\right) = \frac{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c^j}\right) y_i}{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c^j}\right)} \tag{1}$$

where $G = \{\bar{x}_i, y_i, i = 1, 2, ..., n\}$ is dataset contained observations of input features $x^j, j = 1, ..., m$ and output feature y, K is kernel function with finite support (for example, the Epanechnikov kernel or the triangular kernel), $\bar{c}$ is vector of bandwidths. The estimator $\hat{y}\left(\bar{x} \mid G\right)$ is fit using training set G.

The Nadaraya-Watson estimator is becoming less accurate near the boundary of the feature space. Consider, for instance, the Nadaraya-Watson estimation in Figure 1a. On the one hand, fewer observations (Figure 1b) are averaged at the boundary and thus variance and bias can be enlarged. On the other hand, kernel weights are becoming more asymmetric as x approaches the boundary points (Figure 1c shows three kernel weights for $x = 0$, $x = 0.5$ and $x = 1$). This boundary effect is not present for x in the middle of the feature space, but for datasets with small sample size, a significant part of the feature space can be affected by the negative boundary behavior.

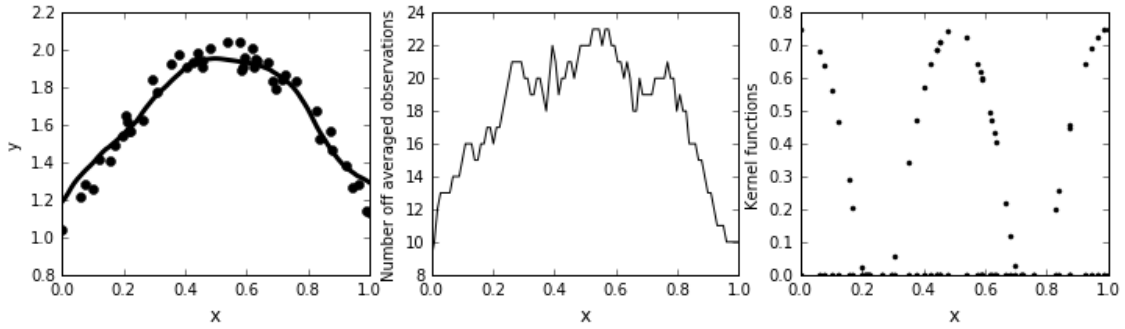


Figure 1: Boundary effect illustration: left plot illustrates the example of boundary effect for one-dimensional task, x is input feature, y is output feature; center plot shows how many observations are averaged in different points of the feature space, right plot shows 3 kernel weights for $x = 0$, $x = 0.5$ and $x = 1$

2 Boosted ensemble of nonparametric estimators

2.1 Algorithm description

The main idea of boosted ensemble of the Nadaraya-Watson estimator is sequentially fitting of weak learners (Nadaraya-Watson estimators) such the next estimator should be more accurate or not exist. This is achieved by bandwidth reducing and allows to focus on the boundary points.

The ensemble learning algorithm is given below:

1. Initialize model with a constant value

$$\hat{H}_0\left(\bar{x} \mid G\right) = n^{-1} \sum_{i=1}^{n} y_i \tag{2}$$

The calculated mean value is equivalent of nonparametric estimator with infinite bandwidths:

$$\hat{H}_0\left(\bar{x} \mid G\right) = \frac{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c_0^j}\right) y_i}{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c_0^j}\right)} \tag{3}$$

where $\bar{c}_0 = \left(c_0^j = \propto : j = 1, 2, ..., m\right)$

2. Fit a nonparametric estimator to pseudo-residuals, i.e. train it using the training set where output feature is replaced by the difference between actual and predicted values using current ensemble $\hat{H}_{q-1}\left(\bar{x} \mid G\right)$:

$$\hat{H}_q\left(\bar{x} \mid G\right) = \hat{H}_{q-1}\left(\bar{x} \mid G\right) + \begin{cases} \hat{h}_q\left(\bar{x} \mid G\right), & \sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c_q^j}\right) > 0 \\ 0, & \sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c_q^j}\right) = 0 \end{cases}, \tag{4}$$

$$\hat{h}_q(\bar{x} \mid G) = \frac{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c_q^j}\right)\left(y_i - \hat{H}_{q-1}(\bar{x} \mid G)\right)}{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c_q^j}\right)}. \quad (5)$$

This process is repeated and new weak learner is added in ensemble until the following inequality is true:

$$E\left(\hat{H}_q(\bar{x} \mid G)\right) < E\left(\hat{H}_{q-1}(\bar{x} \mid G)\right) \quad (6)$$

where $E\left(\hat{H}_q(\bar{x} \mid G)\right)$ is ensemble $\hat{H}_q(\bar{x} \mid G)$ error estimation.

Accuracy estimation is thoroughly discussed in section 2.3. Bandwidth vector $\hat{c}_q$ determination is discussed in section 2.4.

2.2 Cross-validation

Cross-validation should be used to estimate the accuracy of ensembles $\hat{H}_q(\bar{x} \mid G)$, $q = 1, 2, ..., Q$, and optimize sequence of bandwidth $\bar{c}_0, \bar{c}_1, ..., \bar{c}_Q$. This technique allows to avoid overfitting by estimating the number of weak learners Q aggregated in the ensemble.

K-fold cross-validation. In this method, the original dataset is randomly split into K equal sized subsets. Of the K subsets, a single subset containing n_{valid} observations is used as the validation data set to evaluate the model accuracy

$$E\left(\hat{H}_q\left(\bar{x} \mid G^k_{train}\right)\right) = \sum_{i=0}^{n_{valid}} L\left(\hat{H}_q\left(\bar{x}^k_{valid,i} \mid G^k_{train}\right), y^k_{valid,i}\right), \quad (7)$$

the remaining $K - 1$ subsets are used as training data set to fit the model $\hat{H}_q\left(\bar{x} \mid G^k_{train}\right)$, where L is error metric (for example, Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Scaled Error (MASE), Mean Absolute Percentage Error (MAPE)). This process is repeated K times such that each of the K subsets used exactly once as the validation data set. Then prediction accuracies are averaged over all K validation sets to estimate ensemble accuracy:

$$E\left(\hat{H}_q(\bar{x} \mid G)\right) = K^{-1} \sum_{k=1}^{K} E\left(\hat{H}_q\left(\bar{x} \mid G^k_{train}\right)\right). \quad (8)$$

Monte Carlo cross-validation. In this method, the dataset is randomly split K times into training and validation data sets. The proportion of the training/validation split (n_{train}/n_{valid}) is not dependent on the number of folds K and can be selected taking into account the dataset's size and the available computational power. For each split, the model is fit to the training data set and accuracy is estimated using the validation data set. Then prediction accuracies are averaged as in K-fold cross-validation (8).

It is important to note that estimation $\hat{H}_q\left(\bar{x} \mid G^k_{train}\right)$ is impossible without predictions for each sample in the original dataset: $\hat{H}_{q-1}\left(\bar{x}^k_{train,i} \mid G^k_{train}\right), i = 1, ..., n_{train}$,

and $\hat{H}_{q-1}\left(\bar{x}^k_{valid,i} \mid G^k_{train}\right), i = 1, ..., n_{valid}$. At each iteration, it needs to compute and store nK predictions. Moreover, kernel-based nonparametric estimator either takes $O\left(n_{train}\right)$ time for single prediction or takes $O\left(2^D \log n_{train}\right)$ when complex data structures such as kd-trees or cover trees are used [7]. This fact means that leave-one-out cross-validation as a particular case of K-fold cross-validation with $K = n$ is inapplicable in the case of a large number of observations. Choose of parameter K is very important. The larger parameter K, the higher compute complexity and the smaller estimated error variance. So, the value of K should be considered more carefully.

2.3 Bandwidth decreasing strategies

The main task of proposed ensemble learning method is to determine the sequence of bandwidths $\bar{c}_0, \bar{c}_1, ..., \bar{c}_Q$.

Simultaneous optimization of the sequence of bandwidths requires a huge amount of computing resources. To estimate the ensemble accuracy with new sequence of bandwidths, all weak learners included in the ensemble should be rebuild.

Instead of it, bandwidth decreasing strategy is proposed. On each iteration one of the bandwidths is decreased according the following rule:

$$\left(c^1_q, c^2_q, ..., c^m_q\right) = \left(b^{a^1_q} c^1_{q-1}, b^{a^2_q} c^2_{q-1}, ..., b^{a^m_q} c^m_{q-1}\right) \tag{9}$$

where values of vector $\hat{a}_q$ are shown the decreasing bandwidth, $a^j_q \in \{0, 1\}, j = 1, 2, ..., m$ and $\sum_{j=1}^m a^j_q = 1$, $b \in (0, 1)$ is decreasing coefficient. Decreasing coefficient b plays important role because it determines convergence rate of the algorithm (in other words, number of weak learners included in the ensemble).

Different methods can be used for the selection of the coordinate j, on which the next descent iteration would be performed. We consider two standard methods, namely, Cyclic Coordinate Descent and Greedy Coordinate Descent.

Cyclic Coordinate Descent. The different coordinate directions are used cyclically in course of the algorithm. If j-th bandwidth decreasing does not improve the ensemble accuracy, weak learner with such vector of bandwidth $\bar{c}_q$ is not included in ensemble and the next bandwidth is decreased. The process stops when there is not accuracy improvement m iterations in a row.

Greedy Coordinate Descent. In this method, at each iteration, the coordinate direction $j, j = 1, 2, ..., m$ is chosen in a greedy manner by the ensemble accuracy maximization. Greedy Coordinate Descent is intensive computationally because at each iteration, finding the coordinate direction that maximizes the ensemble accuracy with fixed stepsize is done in linear time.

2.4 Computational complexity

Proposed decreasing bandwidth strategy has properties that reduce the computational complexity of the ensemble method. Moreover, this computational complexity is comparable to the computational complexity of a single nonparametric estimator.

In the previous section, we impose the following restrictions on the sequence of bandwidths:

$$\bar{c}_1 \succ \bar{c}_2 \succ ... \succ \bar{c}_Q, \tag{10}$$

where $\bar{c}_{q-1} \succ \bar{c}_q$ means that $\forall j : c^j_{q-1} - c^j_q \geqslant 0, j = 1, 2, .., m$.

In context of Nadaraya-Watson estimators

$$\hat{h}_{q-1}(\bar{x} \mid G) = \frac{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c^j_{q-1}}\right)\left(y_i - \hat{H}_{q-2}(\bar{x}_i \mid G)\right)}{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c^j_{q-1}}\right)} \tag{11}$$

and

$$\hat{h}_{q}(\bar{x} \mid G) = \frac{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c^j_{q}}\right)\left(y_i - \hat{H}_{q-1}(\bar{x}_i \mid G)\right)}{\sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{x_i^j - x^j}{c^j_{q}}\right)} \tag{12}$$

it means that if on the $q-1$-th iteration some points are found in neighborhood of point $\bar{x}$ defined by kernel functions K with bandwidths $\bar{c}_{q-1}$, then on the q-th iteration only these points can be points in new neighborhood of $\bar{x}$ defined by kernel functions K with bandwidths $\bar{c}_q$. Step by step, less and less calculation of kernel function is executed. Figure 2 illustrates the effect of possible computational complexity reduction. Figure 2a shows two neighborhoods of $\bar{x} = (0, 0)$ defined by kernel functions with bandwidths $\bar{c}_{q-1} = (0.75, 0, 75)$ and $\bar{c}_q = (0.75, 0, 5)$, points located in these neighborhoods are depicted as double circles and triple circles correspondingly. Starting from some bandwidths $\bar{c}_{q+r}$, the situation of the lack of points in the neighborhood of $\bar{x}$ becomes possible. For example on Figure 2b there are no points in neighborhoods of $\bar{x} = (0, 0)$ defined by kernel functions with bandwidths $\bar{c}_{q+r} = (0.25, 0, 25)$ (bold solid rectangle), but the same neighborhood of $\bar{x} = (-0.5, 0.5)$ (bold dashed rectangle) contains 4 points. This can be interpreted as the continuation of ensemble learning procedure in some feature subspaces and the completion in other ones. The bandwidths decreasing sooner or later will reduces the number of calculations. But in the worst case, the computational complexity is constant. This occurs when the decreasing of the bandwidth does not improve accuracy, in other words, when the features are insignificant.

It is also important to note that computation complexity of ensemble and computation complexity of bandwidth optimization for single Nadaraya-Watson estimator are comparable. Instead of building the models with different vectors of bandwidths and calculating their accuracy to choose the best one, we aggregate all fitted estimators in ensemble.

3 Experimental results

The verification of the proposed ensemble of the Nadaraya-Watson estimators was conducted using the statistical modeling. Different types of boundary situations was

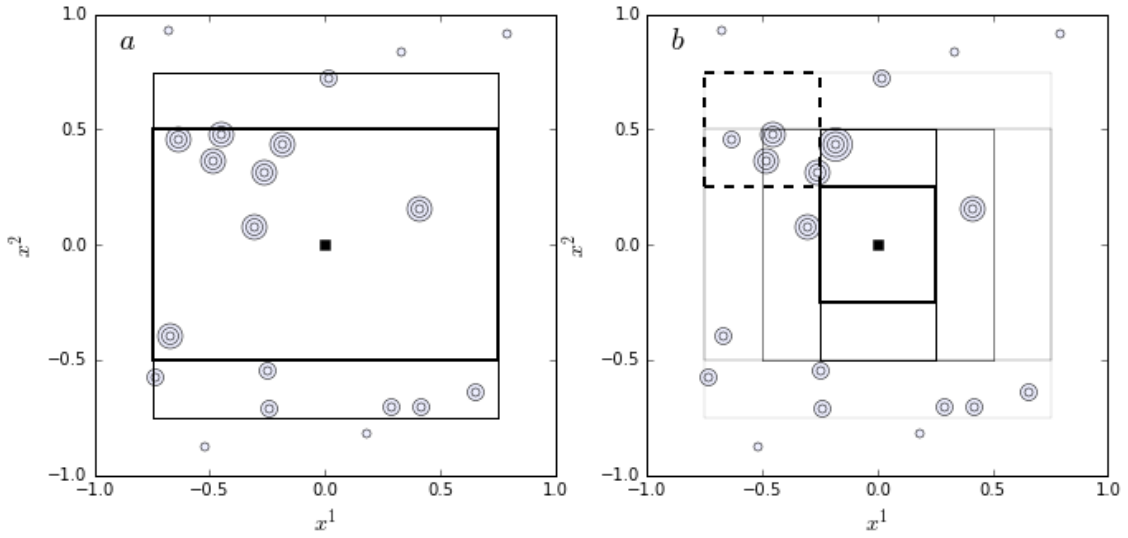


Figure 2: A simulation example, illustrating influence of bandwidths decreasing procedure on computations in 2-dimensional feature space. The left plot shows how computational complexity can be reduced for the next learner. Solid rectangle is the neighborhood of $\bar{x} = (0, 0)$ defined by kernels with bandwidths $\bar{c}_{q-1} = (0.75, 0, 75)$ which contains points depicted as double circles, bold solid rectangle is the neighborhood defined by kernels with bandwidths $\bar{c}_q = (0.75, 0, 5)$ and contained points depicted as triple circles. The right plot shows that starting from bandwidths $\bar{c}_{q+r} = (0.25, 0, 25)$, there are no points in neighborhoods of $\bar{x} = (0, 0)$ (bold solid rectangle) and ensemble learning process stops in such $\bar{x}$, but the same neighborhood of $\bar{x} = (-0.5, 0.5)$ (bold dashed rectangle) contains 4 points and ensemble learning process can be continued in such $\bar{x}$.

generating for one-dimensional feature space. Figure 3 illustrates one of the series of experiments. The boosted ensemble is usually more accurate on the boundary of the feature space and on the data gap (near $x = 0.2$) than a single Nadaraya-Watson estimator. The greater number of input features the more preferable ensemble implementation.

The proposed ensemble learning algorithm was tested on several machine learning benchmark problems and showed accurate predictions in comparison with such popular machine learning algorithms as Gradient Boosting and Random Forest.

Conclusions

The boosted ensemble of the Nadaraya-Watson estimators was proposed. Bandwidth decreasing strategy allows to reduce computational complexity. Also we are planning to adopt decreasing coefficient step by step to make learning process more computationally efficient. Conducted experimental results shows the efficiency of ensemble

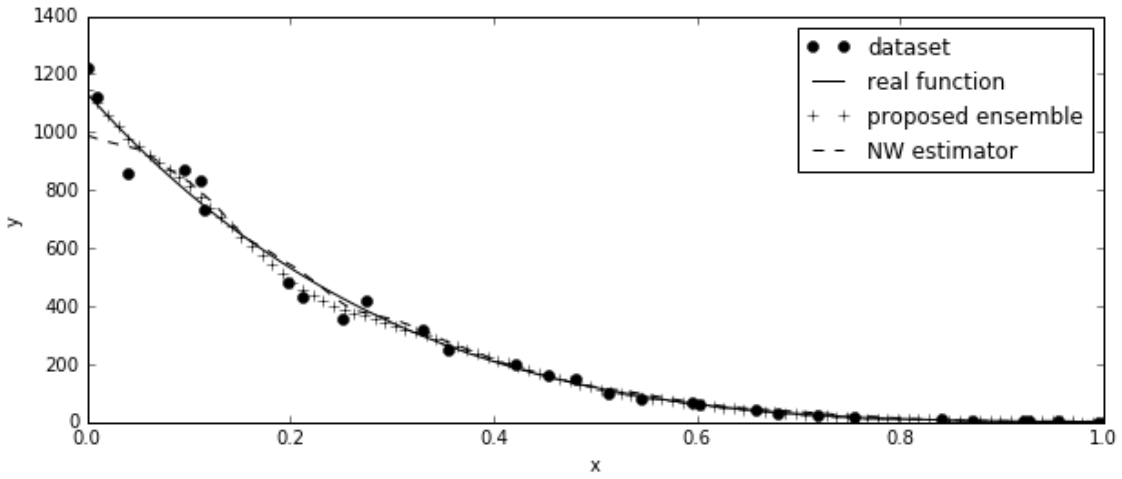


Figure 3: A simulation example, showing the efficiency of the boosted ensemble of the Nadaraya-Watson estimators (line of plusses) with comparison a single one (dotted line)

approach in regression task on the boundary as well as in the middle of the feature space.

References

[1] Breiman L. (1996). Bagging predictors. *Machine Learning*. Vol. **24**, pp. 123-140.

[2] Schapire R. E. (1990). The Strength of Weak Learnability. *Machine Learning*. Vol. **5**, pp. 197-227.

[3] Waterhouse S, Cook G (1997).Ensemble methods for phoneme classification. *Adv Neural Inf Processing Syst*. Vol. **9**, pp. 800-806.

[4] Dietterich T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*. Vol. **40**, pp. 139-157.

[5] Freund Y., Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proc. 13th International Conference on Machine Learning*. pp. 148–156.

[6] Breiman L. (1996).Heuristics of instability and stabilization in model selection. *The Annals of Statistics*. Vol. **24**, pp. 2350-2383.

[7] Ram P., Lee D., March W., Gray A.G. (2009). Linear-time algorithms for pairwise statistical problems. *Advances in Neural Information Processing Systems*. Vol. **22**, pp. 1527–1535.

Modeling Multidimensional Incomplete Sequences using Hidden Markov Models

Vadim E. Uvarov, Alexander A. Popov and
Tatyana A. Gultyaeva
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: uvarov.vadim42@gmail.com, a.popov@corp.nstu.ru,
t.gultyaeva@corp.nstu.ru

Abstract

This paper addresses the problem of multidimensional incomplete sequence modeling using hidden Markov models (HMMs). We propose modified Baum-Welch algorithm that can be used for training HMM on sequences that contain missing observations and modified forward-backward algorithm for classification of incomplete sequences. Additionally, we propose a modified Viterbi algorithm whcih can be used to decode and impute incomplete sequences using HMM. It was shown that both algorithms outperform other methods of missing observation handling, namely: elimination of missing data and imputation of missing observations using the mean of the neighboring observations.

Keywords: hidden Markov models, machine learning, sequences, Baum-Welch algorithm, missing observations, incomplete data, Viterbi algorithm, classification, decoding, imputation.

Introduction

Hidden Markov model (HMM) concept is a popular and powerful tool for sequence modeling. It was presented in late 1960-s and initially applied to speech recognition problems [1]. Despite the fact that this concept is relatively well studied, its usage in missing data scenarios is not properly investigated yet. Traditional algorithms that are used with HMMs cannot process sequences that contain missing observations i.e. incomplete sequences.

Some attempts to solve this problem were already made by several authors. For example, in [2] authors used marginalization and imputation approaches to classify incomplete noisy speech samples using HMMs. Such an approach proved to be more effective than the standard filtering methods. However in this work HMMs were trained on clean samples and sequence decoding problem was not addressed. In [3] missing observations in a sequence of movements were used to represent a situation when humanoid body part movement was occluded by some obstacle. The authors proposed a decoding algorithm that was used to infer the actual movements of a humanoid model from a sequence of data extracted from video-feed. However, once again HMMs were trained on a clean data.

In this work we address the problem of HMM training on incomplete sequences, decoding of incomplete sequences using HMM and classification of incomplete sequences using HMM. This paper continues the studies of HMM concept which are

carried out at the chair of theoretical and applied informatics at Novosibirsk State Technical University [4, 5].

1 Hidden Markov Model

The thorough description of hidden Markov model concept is out of scope of this paper. A good overview can be found in [1]. However we will introduce some notation to refer to elements of HMM and we will mention the basic algorithms.

We denote a sequence of Z-dimensional observations as $O = \{o_1, ..., o_T\}$, where T is the total number of observations in the sequence and o_t is the observation at time t. Hidden state of HMM is denoted as $s_i, i = \overline{1,N}$ where N is the total number of hidden states in a model and hidden state of model at time t is denoted as $q_t, t = \overline{1,T}$. We denote an HMM as $\lambda = \{\Pi, A, B\}$, where $\Pi = \left\{\pi_i = p\left(q_1 = s_i\right), \ \ i = \overline{1,N}\right\}$ - initial state distribution, $A = \left\{a_{ij} = p\left(q_{t+1} = s_j | q_t = s_i\right), \ \ i,j = \overline{1,N}\right\}$ - transition probabilities matrix and $B = \left\{b_i\left(o\right) = f\left(o|q = s_i\right), \ i = \overline{1,N}, \ o \in \mathbb{R}^Z\right\}$ - conditional probability density functions of multidimensional observations. In this paper we use mixtures of Gaussian distributions to model the observation densities, hence $b_i(o) = \sum\limits_{m=1}^{M} \tau_{im} g(o; \mu_{im}, \Sigma_{im}), \quad i = \overline{1,N}, \ o \in \mathbb{R}^Z$, where M is the number of mixtures, τ_{im} is the weight of m-th mixture in i-th hidden state, μ_{im} is Z-dimensional mean vector of m-th mixture in i-th hidden state, Σ_{im} is Z by Z covariance matrix of m-th mixture in i-th hidden state and $g(o; \mu_{im}, \Sigma_{im}), \ o \in \mathbb{R}^Z$ is Gaussian density function, i.e. $g(o; \mu_{im}, \Sigma_{im}) = \frac{1}{\sqrt{(2\pi)^Z |\Sigma_{im}|}} e^{-0.5(o-\mu_{im})^T {\Sigma_{im}}^{-1} (o-\mu_{im})}, \ o \in \mathbb{R}^Z$.

To perform a sequence classification given a sequence $O = \{o_1, ..., o_T\}$ and a set of HMMs $\lambda_1, \ ..., \ \lambda_D$ one would usually use a maximum likelihood criterion. A sequence O is referred to a model for which likelihood function is maximum, i.e. to a model $\lambda^* = \arg \max\limits_{\lambda \in \lambda_1, ... \lambda_D} \left(p\left(O|\lambda\right)\right)$. An efficient forward-backward algorithm is usually used to calculate a likelihood function $p\left(O|\lambda\right)$ [1]. A forward variable is defined as $\alpha_t(i) = p(\{o_1, o_2, ..., o_t\}, q_t = s_i | \lambda)$ and backward variable is defined as $\beta_t(i) = p(\{o_{t+1}, o_{t+2}, ..., o_T\} | q_t = s_i, \lambda), \ t = \overline{1,T}, \ \ i = \overline{1,N}$.

By decoding a sequence of observations $O = \{o_1, ..., o_T\}$ using a hidden Markov model λ one usually means finding the most probable sequence of hidden states $\{\widehat{q}_1, ..., \widehat{q}_T\}$ which correspond to the observations. Viterbi algorithm is usually used for this purpose [1]. This algorithm relies on calculation of probabilities $\delta_t\left(j\right) = p\left(q_t = s_j | o_t, \lambda\right), \ j = \overline{1,N}, \ t = \overline{1,T}$.

Training of hidden Markov models is usually performed with Baum-Welch algorithm, which is essentially a modification of EM-algorithm [1]. This algorithm iteratively maximizes the likelihood function and requires an initial approximation. Since it converges to a local maximum, it is usually run with a several different initial approximations. We denote additional probabilities used in Baum-Welch algorithm here: $\gamma_t(i) = p(q_t = s_i | O, \lambda), \ \xi_t(i,j) = p(q_t = s_i, q_{t+1} = s_j | O, \lambda)$ and $\gamma_t(i,m) = p(q_t = i, \omega_{it} = m | O, \lambda), \ \ t = \overline{1,T-1}, \ i,j = \overline{1,N}, \ m = \overline{1,M}$.

2 Incomplete Sequence Analysis

We define an incomplete sequence as a sequence where the value of some observations is undefined. We call such observations 'missing' and denote them with a sign $\emptyset$. It is clear that none of the standard algorithms cannot handle missing values, since calculation of probabilities $\alpha_t(i), \beta_t(i), \delta_t(j), \gamma_t(i), \xi_t(i,j)$ and $\gamma_t(i,m)$ requires calculation of $b_i(o_t)$ which in its turn requires the knowledge of the actual value of o_t.

A standard way of dealing with missing observations is to delete them from sequence completely and then glue the remaining parts together. We will call such method 'gluing'. The other standard way to eliminate the gaps is to fill them with the mean of the neighbor observations.

In the next to subsections we propose a more accurate and complex algorithms for dealing with incomplete sequences. The main idea of these algorithms is to suppose that $b_i(o_t) = 1$ when $o_t = \emptyset$ since any observation could have been in place of o_t and the probability of seeing any observation at time t is one.

2.1 Modified Viterbi Algorithm

The proposed modified Viterbi algorithm extends original Viterbi algorithm to the case of missing observations. Given an observation sequence $O = \{o_1, ..., o_T\}$ and HMM λ the steps of modified Viterbi algorithm are as follows:

1) initialization: $\delta_1(i) = \begin{cases} \pi_i, & o_1 = \emptyset \\ \pi_i b_i(o_1), & otherwise \end{cases}, \quad i = \overline{1,N}, \quad \psi_1(i) = 0;$

2) induction: $\delta_t(j) = \begin{cases} \max\limits_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], & o_t = \emptyset \\ \max\limits_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(o_t), & otherwise \end{cases},$

$j = \overline{1,N}, t = \overline{2,T}, \psi_t(j) = \arg\max\limits_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \quad j = \overline{1,N}, \ \ t = \overline{2,T};$

3) termination: $\widehat{q}_T = \arg\max\limits_{1 \le i \le N} [\delta_T(i)];$

4) backward inference of the most probable sequence of hidden states:

$$\widehat{q}_t = \psi_{t+1}\left(\widehat{q}_{t+1}\right), \ \ t = \overline{T-1, \ \ 1}.$$

In the result we will obtain the most probable sequence of hidden states $\widehat{Q} = \{\widehat{q}_1, \cdots, \widehat{q}_T\}$.

Besides decoding this algorithm can be also applied to impute the missing values. After decoding the missing value can be replaced with the most probable observation $\widehat{o}_t$ in the decoded hidden state $\widehat{q}_t$ (mean of corresponding distribution) or sampled from an distribution which corresponds to decoded hidden state $\widehat{q}_t$. It should be clear that after imputation one can also perform classification of imputed sequence or train model using that imputed sequence.

2.2 Modified Forward-Backward Algorithm

Given an observation sequence $O = \{o_1, ..., o_T\}$ and HMM λ the steps of modified forward variable computation algorithm are as follows:

1) initialization: $\alpha_1(i) = \begin{cases} \pi_i, & o_1 = \emptyset \\ \pi_i b_i(o_1), & otherwise \end{cases}, i = \overline{1,N};$

2) induction: $\alpha_{t+1}(i) = \begin{cases} \sum\limits_{j=1}^{N} \alpha_t(j) a_{ji}, & o_{t+1} = \emptyset \\ b_i(o_{t+1}) \left[\sum\limits_{j=1}^{N} \alpha_t(j) a_{ji} \right], & otherwise \end{cases}, \ i = \overline{1,N}$

$t = \overline{1,T-1}$.

Since $p(O|\lambda) = \sum\limits_{i=1}^{N} \alpha_T(i)$, this algorithm can be used for incomplete sequence classification using the maximum likelihood criterion.

The steps of modified backward variable computation algorithm are as follows:
1) initialization: $\beta_T(i) = 1, \quad i = \overline{1,N}$

2) induction: $\beta_t(i) = \sum\limits_{j=1}^{N} \beta_{t+1}(j) b_j(o_{t+1}) a_{ij}, \ i = \overline{1,N}, \ t = \overline{1,T-1}$.

The backward variables are used in modified Baum-Welch algorithm.

2.3 Modified Baum-Welch Algorithm

Given a set of training incomplete sequences $O^* = \left\{O^1, \ O^2, ..., \ O^K\right\}$ and some HMM approximation $\widehat{\lambda}$, one iteration of Baum-Welch algorithm consists of the following steps:
1) $\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{p(O|\widehat{\lambda})}, \quad i = \overline{1,N}, \ \ t = \overline{1,T-1};$

2) $\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{p(O|\widehat{\lambda})}, \ \ i,j = \overline{1,N}, \quad t = \overline{1,T-1};$

3) $\gamma_t(i,m) = \begin{cases} \gamma_t(i) \ \tau_{im}, & o_t = \emptyset \\ \gamma_t(i) \left[\frac{\tau_{im} g(o_t, \mu_{im}, \Sigma_{im})}{b_i(o_t)} \right], & otherwise \end{cases} \ i = \overline{1,N}, m = \overline{1,M}, t = \overline{1,T}.$

4) New approximation of HMM parameters $\widehat{\lambda}'$ is calculated as follows:

$$\hat{\pi}'_i = \frac{1}{K} \sum_{k=1}^{K} \gamma_1^{(k)}(i), \ \widehat{a}'_{ij} = \frac{\sum\limits_{k=1}^{K} \sum\limits_{t=1}^{T^k-1} \xi_t^{(k)}(i,j)}{\sum\limits_{k=1}^{K} \sum\limits_{t=1}^{T^k-1} \gamma_t^{(k)}(i)}, \widehat{\tau}'_{im} = \frac{\sum\limits_{k=1}^{K} \sum\limits_{t=1}^{T^k-1} \gamma_t^{(k)}(i,m)}{\sum\limits_{k=1}^{K} \sum\limits_{t=1}^{T^k-1} \gamma_t^{(k)}(i)},$$

$$\widehat{\mu}'_{im} = \frac{\sum\limits_{k=1}^{K} \sum\limits_{t=1}^{T^k-1} \gamma_t^{(k)}(i,m) \ o_t^k}{\sum\limits_{k=1}^{K} \sum\limits_{t=1}^{T^k-1} \gamma_t^{(k)}(i,m)}, \widehat{\Sigma}'_{im} = \frac{\sum\limits_{k=1}^{K} \sum\limits_{\substack{t=1 \\ o_t \neq \emptyset}}^{T^k-1} \gamma_t^{(k)}(i,m) (o_t^{(k)} - \widehat{\mu}'_{im})(o_t^{(k)} - \widehat{\mu}'_{im})^T}{\sum\limits_{k=1}^{K} \sum\limits_{\substack{t=1 \\ o_t \neq \emptyset}}^{T^k-1} \gamma_t^{(k)}(i,m)}.$$

After iteration steps $\widehat{\lambda}'$ replaces $\widehat{\lambda}$ and new iteration begins. This iterative process continues until likelihood function converges to local maximum.

3 Evaluation

3.1 Training on Incomplete Sequences

To evalute the training algorithms for incomplete sequences we used the original HMM with the following parameters: $N = 3, M = 3, Z = 2$. $A = \begin{bmatrix} 0.1 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.6 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$, $\{\tau_{im},\ i = \overline{1,N},\ m = \overline{1,M}\} = \begin{pmatrix} 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$, $\{\mu_{im},\ i = \overline{1,N},\ m = \overline{1,M}\} = \begin{pmatrix} \begin{pmatrix} 0 & 0 \end{pmatrix}^T & \begin{pmatrix} 1 & 1 \end{pmatrix}^T & \begin{pmatrix} 2 & 2 \end{pmatrix}^T \\ \begin{pmatrix} 3 & 3 \end{pmatrix}^T & \begin{pmatrix} 4 & 4 \end{pmatrix}^T & \begin{pmatrix} 5 & 5 \end{pmatrix}^T \\ \begin{pmatrix} 6 & 6 \end{pmatrix}^T & \begin{pmatrix} 7 & 7 \end{pmatrix}^T & \begin{pmatrix} 8 & 8 \end{pmatrix}^T \end{pmatrix}$, all covariance matrixes $\{\Sigma_{im},\ i = \overline{1,N},\ m = \overline{1,M}\}$ were diagonal with 1 on diagonal. Using this HMM we generated $K = 100$ sequences with the length $T = 100$. We were varying the percent of gaps in sequences from 0% to 90%. We were comparing 4 different training approaches for incomplete sequences: modified Baum-Welch algorithm (Baum-Welch), gluing of incomplete sequences (gluing), imputation using the modified Viterbi algorithm (Viterbi) and mean imputation (mean) - mean of 10 neighbours were taken. To measure training performance we used loglikelihood of original sequences and distance between models as was proposed in [1]: $D_s = \frac{D(\lambda,\widehat{\lambda})+D(\widehat{\lambda},\lambda)}{2}$, $D(\lambda_1,\lambda_2) = \frac{1}{T}\,|\ln p(O^2\,|\lambda_1) - \ln p(O^2\,|\lambda_2)\,|$, where O^2 is the sequences generated by model λ_2. Average results of 50 launches with different randoms seeds are presented in Fig. 1.

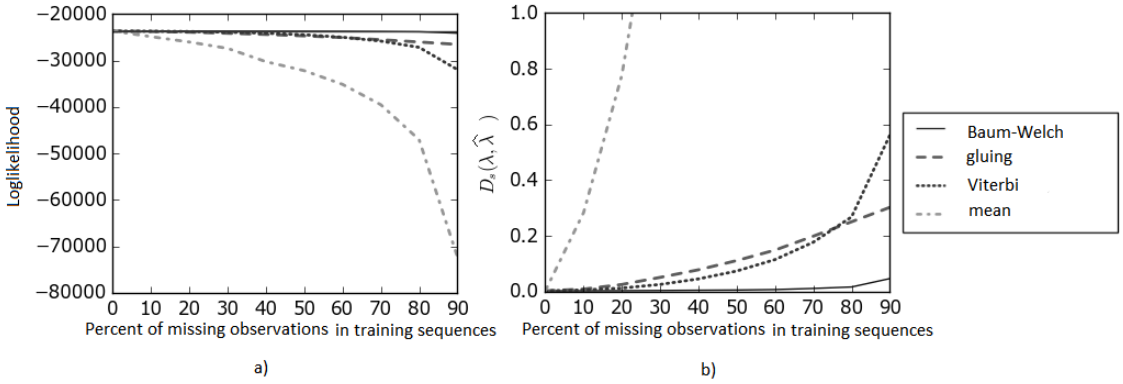


Figure 1: a) Average Loglikelihood of Complete Sequences for HMM Trained on Incomplete Sequences b) Average Distance Between True and Estimated HMM

3.2 Classifying Incomplete Sequences

To evalute the classification algorithms for incomplete sequences we used two HMMs with the same parameters as in the previous subsection except for the transition probabilities matrix: $A = \begin{bmatrix} 0.1 + \Delta A & 0.7 - \Delta A & 0.2 \\ 0.2 & 0.2 + \Delta A & 0.6 - \Delta A \\ 0.8 - \Delta A & 0.1 + \Delta A & 0.1 \end{bmatrix}$. Using these HMMs we generated $K = 100$ sequences with the length $T = 100$. We were varying the percent of gaps in sequences from 0% to 90%. We classified these sequences using the original HMMs. We were comparing 4 different classification approaches for incomplete sequences: modified forward-backward algorithm (forward-backward), gluing of incomplete sequences (gluing), imputation using the modified Viterbi algorithm (Viterbi) and mean imputation (mean) - mean of 10 neighbours were taken. To measure training performance we used accuracy metric. Average results of 50 launches with different randoms seeds are presented in Fig. 2.

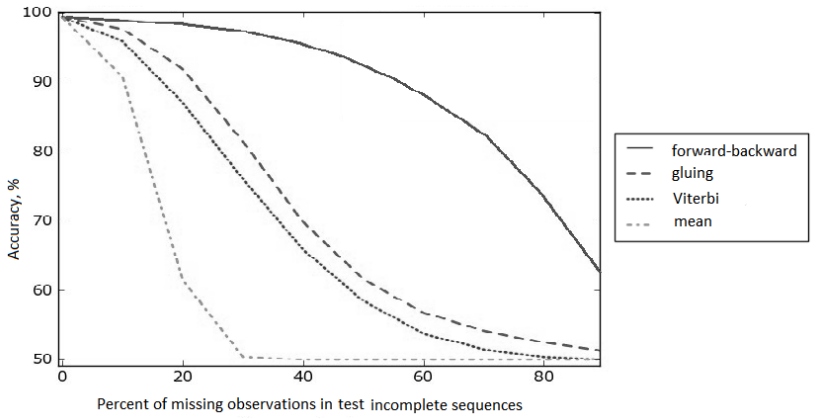


Figure 2: Accuracy of Incomplete Sequence Classification

3.3 Imputation and Decoding of Incomplete Sequences

To evalute the classification algorithms for incomplete sequences we used HMM with the same parameters as in the first subsection of this section. We generated $K = 100$ sequences with the length $T = 100$. We were varying the percent of gaps in sequences from 0% to 90%. We imputed and decoded these sequences using the original HMMs. We were comparing 2 different imputation and decoding approaches for incomplete sequences: the modified Viterbi algorithm (Viterbi) and mean imputation (mean) - mean of 10 neighbours were taken. To measure decoding performance we used percent of correctly decoded states and to measure imputation performance we used mean of norms of difference between actual and imputed observations. Average results of 50 launches with different randoms seeds are presented in Fig. 3.

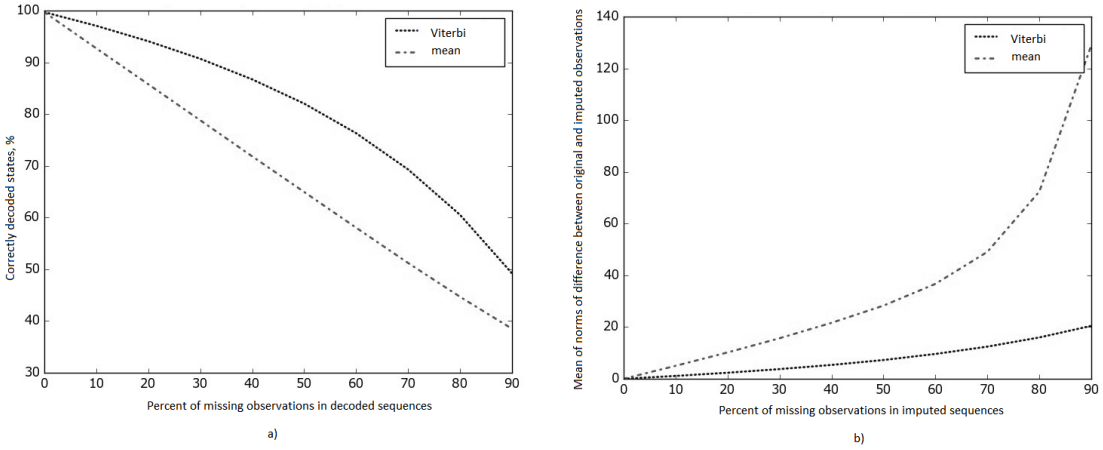


Figure 3: a) Percent of Correctly Decoded Observations b) Mean of Norms of Difference Between Actual and Imputed Observations

Conclusion

In this paper we proposed a method for training hidden Markov models on incomplete sequences, and methods for classification, decoding and imputation of incomplete sequences using hidden Markov models. Computer experiment results show that the proposed methods outperform the standard methods of dealing with missing observations.

References

[1] Rabiner L.R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition *Proceedings of the IEEE*. Vol. **77**, pp. 257-285.

[2] Cooke M., Green P., Josifovski L., Vizinh A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data *Speech Communication*. Vol. **34**, pp. 267-285.

[3] Lee D., Kulic D., Nakamura Y. (2008). Missing motion data recovery using factorial hidden Markov models *IEEE International Conference on Robotics and Automation*. pp. 1722-1728.

[4] Gultyaeva A., Popov A., Kokoreva V., Uvarov V. (2015) Classification of observation sequences described by Hidden Markov Models *Proceedings of the International Workshop Applied Methods of Statistical Analysis: Nonparametric approach*. pp. 136-143.

[5] Gultyaeva A., Popov A., Uvarov V. (2016) Training Hidden Markov Models on Incomplete Sequences *Proceedings of 13th International Conference on Actual Problems of Electronic Instrument Engineering*. Vol. **1**, pp. 317-320.

Informational and Educational Interaction for Multilingual Environment

Margarita V. Karaseva
Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
e-mail: karaseva-margarita@rambler.ru

Abstract

The structure of the learning process in the individual language training is presented. The implementation of the basic components of the multilingual environment of the informational and educational interaction based on the concept of the network training environment is considered. The formation root of the informational and terminological basis of such environment is frequency dictionaries; they determine the priority of these of those terms training. It is proved that the use of such dictionaries obtained due to the analysis of linguistic material improves the quality of foreign vocabulary training.

Keywords: multilingual environment, training process, informational and terminological basis, adaptation.

Introduction

Recently, due to the development of computer technologies new forms of education have become widespread. One of them is a system of individual education. The use of a computer in the training process helps to make the process of getting knowledge more interesting, exciting, highly operational. Improving the quality of education is achieved through the individual work of the student with educational material.

The introduction of automated training systems and active use of new information technologies in educational institutions have wide opportunities for the intensification and individualization of the learning process [9].

A training system is a set of interacting elements: training programs developed to meet the real needs for the improved user knowledge; departments that identify the training needs, manage and evaluate the effectiveness and quality of training; environment of institutions and departments directly engaged in training regardless of their organizational and legal forms and types that implement the adopted training programs.

1 Application of the multilingual environment

The implementation of automated training systems directly for the training purpose is related to the computer's performance of the following interrelated functions: management of educational activity; storage and delivery of the educational information; modeling of laboratory experiments, phenomena, situations, regularities, etc.; analysis of messages and responses of trainees; registration, storage and processing of learning results of students.

The multilingual adaptive training technology [2] is used to develop training systems that facilitate the intensive accumulation of the professionally foreign vocabulary for specialists and students who are faced with the foreign terminology while working with foreign literature or collaborating with foreign colleagues.

Two interrelated components are in the center of modern multilingual adaptive-teaching technology; they are an information-terminological basis (ITB) and training technology [3].

The multilingual adaptive training technology corresponds to a number of possible training processes [4]. For example, a structure that reflects the situation of simultaneous training of the group of students using one multilingual information and terminology basis is presented in Figure 1. A particular implementation of this structure is a system of one-dimensional language training (English-Russian, English-German informational and terminological basis, etc.).

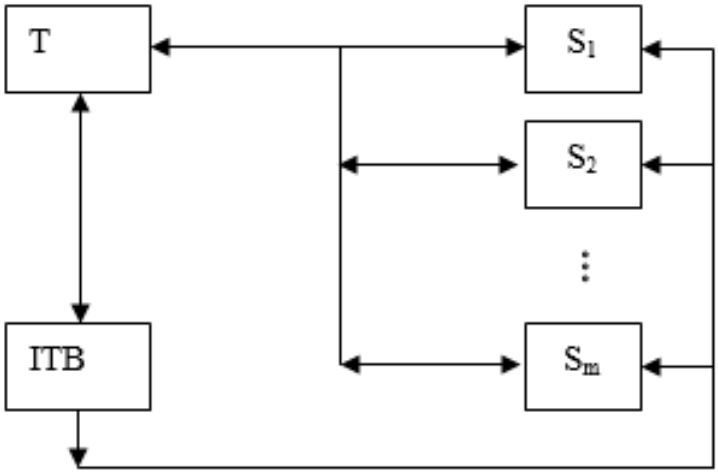


Figure 1: Simultaneous training of the group of students using one informational and terminological basis T is a teacher (in general case, a virtual one); S_1, S_2, ..., S_m are students

Letś note that this structure leads to an increase in the productivity of training only with a careful selection of a group of students. We will achieve only the productivity of the "slow learner" student having different students in the real system, and defining the goal as compulsory education of everyone [5], [6].

The simultaneous training of the group using some (or some different) ITB, i.e. each student independently (according to the recommendation of the teacher) chooses his basis (ITB_1, ITB_2, ..., ITB_k) corresponding to the structural diagram shown in Figure 2. The multiple relationships of components are added to this structure and it becomes more flexible. Not only the training time decreases but also its effectiveness increases as are elements of the individual approach to the student, and there is the possibility of adaptation of the ITB to the requirements of the student.

If in the previous structure the multiple links between the teacher and students are established then we will receive the training process which in [5] is called the

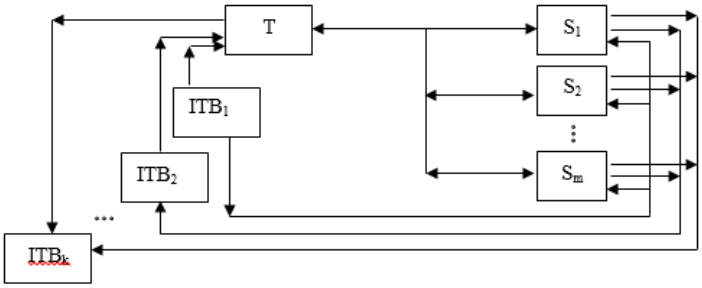


Figure 2: Simultaneous training of the group using some informational and terminological basis

individual training in the group when a teacher controls the training process, i.e., he decides the next portion of knowledge and skills the student needs to receive. The advantage of this scheme can be realized only if such a component as a teacher is duplicated. It will increase the capacity of this component and it will allow using the advantages of both individual and group training. In this case, it is not necessary to assume that the teachers (teaching algorithms) are the same. It is possible to build adaptive learning algorithms that are customizable in each specific case (at each training session) for a particular trainee. This approach corresponds to the structure with many teachers implementing different training algorithms or adapting to the student thereby reducing the training time of a weak student and increasing the efficiency of training a strong student.

We will point out a lot of informational connections of the structural scheme of the training process, a multilingual environment of the informational and educational interaction. It should be noted that unlike the traditional training process that becomes multidimensional with a lot of subject disciplines, the training process that is realized with the help of multilingual adaptive training technology remains to be a one-dimensional process in the presence of a multidimensional informational and terminological basis that can be realized as one-dimensional; the correspondence between the multilingual components of ITB_1, ITB_2, ..., ITB_k is to be established as explicit.

As a rule, some elements of the structure of the training process, related to each other through the medium of educational interaction [4] form an information and educational network (in our case, a multilingual environment). It is obvious that the implementation of the multilingual information and education environment in traditional education is practically an impossible problem. The automation of the training process for this structure requires the automation of the ITB formation (due to the query of students up to the dynamic formation of the ITB), as well as the computer support of the technology of adaptive training in an interactive mode.

The informational and terminological basis of the multilingual technology is based on the results of the analysis of the linguistic material. One should understand as the linguistic material here a certain number of texts of the interested subject field of

the language being studied. The volume of the language material can vary depending on the tools of analysis, the availability of original texts and a necessary number of terms.

The basis for the formation an informational and terminological basis is frequency dictionaries; according to these dictionaries the priority is determined for training certain terms. The implementation of such dictionaries obtained during the analysis of language material, improves the learning process of foreign vocabulary qualitatively [2], [5]. According to the relational model of the information and terminology basis developed by the authors all its elements are divided into two groups according to their frequency characteristics: lexemes with the largest frequency (the main lexemes) and lexemes whose frequency does not exceed a certain numerical threshold. Of these, later, relational series are formed, each of which uniquely corresponds to one of the main lexemes. In this case, the structure of the multilingual component (ML-component) reflecting the main lexeme is supplemented by the corresponding transition probabilities.

The information and terminological basis of the multilingual adaptive training technology, as mentioned above, is presented in the form of electronic frequency dictionaries, which further frequency division is used in training [5], [6].

Frequency dictionaries differ in a number of features, the most important of which are the language, content and volume of texts that served as material for the dictionary of a certain natural language, its word and word forms.

The authors developed an English-German-Russian frequency dictionary on system analysis in electronic engineering and aerospace [11]. This dictionary reflects some of the important qualitative and quantitative aspects of the use of lexicon in system analysis and informatics in English, German and Russian as a result of statistical analysis and description of texts that will contribute to the organization of the rational assimilation of vocabulary and the accumulation of the vocabulary.

The peculiarity of such training is in the formation of the associative links within the trained terminology taking into account that the most number of links occurring in the main lexemes, specially selected as the most significant ones for training.

The electronic dictionary being information and terminological support for the multilingual adaptive training technology allows automating the process of training the terminology of a foreign language taking into account the relations of its elements with the elements of the terminology of the previously studied languages and the generation of an associative field around the memorized terms.

Conclusions

Based on the concept of the environment educational structure and relying on the particular formalized representations of the educational process in language training, the analysis of the possible structures of the multilingual environment of the informational and educational interaction was conducted. It led to the conclusion that the multilingual adaptive training technology corresponds to a number of possible training processes (depending on the real implementation in the form of a computer

system). At the same time, the problem of taking into account associative characteristics of lexemes during the formation of informational and terminological support the generation of an associative field around memorized terms using associations between elements of various foreign terminologies is considered. That makes it possible to fill vocabulary out in practice intensively.

References

[1] Monakhov M. Yu. (2001). TInformational educational network. *Information technologies*. No. **7**, pp. 36-47.

[2] Kovalev I. V., Karaseva M. V., Leskov V. O. (2009). Components of information support of the multilinguistic adaptive traning technology. *Control systems and information technologies*. Vol. **35**,No. **1.3** pp. 360-363.

[3] Kovalev I. V., Karaseva M. V., Suzdaleva E. A.(2002). System aspects of the organization and application of the multilingual adaptive training technology. *Educational technologies and society*. Vol. **5**,No. **2** pp. 198-212.

[4] Kovalev I. V., Kovaleva T. A., Karasyova M. V., Ezhemanskay S. N.(2004). System aspects of multilingual adaptive training technology organization and usage. *Proceedings of Modeling and Simulation, MS'2004 AMSE International Conference on Modelling and Simulation, MS'2004. sponsors: University Lyon 1, France, Assoc. for the Adv. Model. AndSimul. Tech. Enterprises, AMSE, French Research Council, CNRS, Rhone-Alpes Region, Hospitals of Lyon. Lyon-Villeurbanne.*

[5] Kovalev I. V, Zelenkov P. V., Yarkova S. A., Shevchuk S. F. (2007). SOptimization of data processing in distributed educational environments. *Software products and systems*. No. **3** p. 28.

[6] Kovalev I. V., Karaseva M. V., Leskov V. O.(2009). Algorithmization of procedures for the unity of the related terms in the structure of the information and terminological basis. *Software products and systems*. No. **4** p. 28.

[7] Kovalev I. V., Karaseva M. V.(2013). *English-German-Russian frequency dictionary on system analysis in electronic engineering and aerospace industry.* Siberian State Aerospace University, Krasnoyarsk.

Logistic Regression as a Diagnostic Model for Stochastic Systems[1]

Alexander N. Tyrsin[1,2], Elena V. Chistova[2], Kirill K. Kostin[3]
[1] *Ural Federal University named after the first President of Russia B.N. Yeltsin, Yekaterinburg, Russia*
[2] *Institute of Economics, Ural branch of the Russian Academy of Sciences, Yekaterinburg, Russia*
[3] *South Ural State University (national research University), Chelyabinsk, Russia*
email: at2001@yandex.ru, elvitvas@ya.ru, lemwwwar@gmail.com

Abstract

It is proposed to apply logistic regression as a diagnostic model for multidimensional stochastic systems. Control problems for multidimensional systems have been stated as extremum problems. The idea is improving an object status through its discrimination as a member of the efficient objects class. The solution is found by means of optimal change of the object state vector. An example of developing and analyzing a model of relation of life quality to regional social and economic indicators is given.

Keywords: logistic regression, stochastic system, model, diagnostics, control, maximum likelihood method.

Introduction

Direct application of cause-and-effect models to complex stochastic systems is often difficult since an acceptable level of certainty cannot be reached. In such cases, categorization is usually used. A benefit of categorization is the possibility of generating statistically significant categorization rules without direct cause-and-effect relationship modeling.

Logistic regression is a common way of multivariate data categorization used in various fields [2,6,7,10]. The primary purpose of logistic regression is dividing a set of input values with a linear boundary into two areas corresponding to the two specified categories. Logistic regression provides probability of a certain event in the 0...1 range [7]. It should be noted that the number of categories L may exceed two; in this case it is called a multinominal regression. It can be obtained, e.g., with $L-1$ independent logistic regressions. For this reason, we consider categorization into two categories.

However, applying logistic regression only as a categorization rule limits practical applicability of the method.

The paper studies possible logistic regression application as a diagnostic model.

[1]The study has been supported by the Russian Foundation for Basic Research (Project No. 16-06-00048a).

1 Building logistic regression with the maximum-likelihood criterion

Currently logistic regression is usually built as an optimization problem with the maximum-likelihood criterion [8].

Let us have a set of precedents (a learning sample).

$$(\mathbf{x}_i, y_i), i = 1, 2, ..., n,$$

where $\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{im} \end{pmatrix}$ is a vector of ith object values;

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T \\ \mathbf{X}_2^T \\ \dots \\ \mathbf{X}_3^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix};$$

$y_i \in \{-1; 1\}$ is a binary variable that identifies the ith object membership in a corresponding category, e.g., the first category when it is $y_i = -1$ and the second category when it is y_i=1, m is the number of each object's properties; n is the number of observations.

The following function is used for categorization:

$$h(\mathbf{x}) = \frac{1}{1 - e^{-\mathbf{b}^T\mathbf{x}}} \tag{1}$$

Its values lie within the (0; 1) range. The threshold value is $h(\mathbf{x}) = 0.5$. The $\delta(z) = \frac{1}{1+e^{-z}}$ function is called "a logistic function".

The $\mathbf{b}^T = (b_0 \quad b_1 \quad \dots \quad b_m)$ coefficient vector in equation (1) defines the linear boundary. It is generally described by a hyperplane equation:

$$W(\mathbf{x}) = b_0 + \sum_{j=1}^{m} b_j x_j = 0 \tag{2}$$

We define the D_1 range of possible $\mathbf{x}$ values for the first category as $D_1 = \{\mathbf{x} : W(\mathbf{x}) < 0\}$, and $D_2 = \{\mathbf{x} : W(\mathbf{x}) > 0\}$ for the second category. Then $\forall \mathbf{x} \in D_1$ $h(\mathbf{x}) < 0.5$ and $\forall \mathbf{x} \in D_2$ $h(\mathbf{x}) > 0.5$. If $\mathbf{x}$ belongs to the hyperplane (2), then $h(\mathbf{x}) = 0.5$.

Then for an arbitrary observation $\mathbf{x}^*$ the probability of its membership in the first category is $P(\mathbf{x}^* \in D_1) = 1 - h(\mathbf{x}^*)$, and in the second category it is $P(\mathbf{x}^* \in D_2) = h(\mathbf{x}^*)$.

As shown in [11], the $\mathbf{b}$ vector definition by maximizing the likelihood logarithm is equivalent to minimizing the composed function:

$$Q(\mathbf{b}) = \sum_{j=1}^{n} \ln(1 + e^{-y_i b^T x_i}) \to \min_{b \in R^m} \tag{3}$$

The (3) minimization problem has no analytical solution. For this reason, iterative descent methods applied to extremum problems are used to evaluate the **b** coefficient vector. Paper [11] presents an iterative Newton Raphson method for solving the (3) problem.

2 Logistic regression model properties study

We evaluate whether equations (1) and (2) can be considered a mathematical model. The primary goals of a model are [1]: explaining the object design, its structure, and internal properties; ensuring the object's controllability, determining the most appropriate control methods for the specified objectives and criteria; forecasting direct and indirect consequences of the specified action applied to the object.

Each $\mathbf{x}_i$ vector is a set of the ith object values. The value of the y_i binary variable indicates the ith object membership of a category. For definiteness, we postulate that the first category ($y = -1$) is a set of objects having "poor" properties while the second category ($y = 1$) is a set of objects with "good" properties.

Using (2) we can evaluate $W(\mathbf{x}^*)$ with the $\mathbf{x}^*$ set of actual property values. If $W(\mathbf{x}^*) < 0$, the object belongs to the first category with a probability of $1 - h(\mathbf{x}^*)$. If $W(\mathbf{x}^*) > 0$, the region belongs to the second category with a probability of $h(\mathbf{x}^*)$. If $W(\mathbf{x}^*) = 0$, the situation is undetermined. The region can belong to both the first and the second category with a probability of 0.5.

The mathematical model's parameters are the **b** logistic regression vector coefficient, and the **x** property value vector. The **b** vector components are exogenous variables while that of the **x** vector are endogenous variables.

Let us consider an object that has a set of actual property values expressed as the $\mathbf{x}^*$ vector. According to (1) it is possible to offer a probabilistic estimation of the region resident's life quality. Indeed, it belongs to the high life quality regions with a probability of $h(\mathbf{x}^*)$. The probability can serve as a value of the goal function. The higher the $h(\mathbf{x}^*)$ value, the more is the probability of the object's membership in the second category.

The object properties depend on the $\mathbf{x}^*$ position with respect to the (2) hyperplane. Numerically they can be defined, e.g., as a distance from point $\mathbf{x}^*$ to hyperplane (2)

$$d = \frac{|\mathbf{b}^T \mathbf{x}^*|}{|\mathbf{b}|}$$

and the gradient $\mathrm{grad}h(\mathbf{x}^*)$ of the $h(\mathbf{x})$ function at the point $\mathbf{x}^*$

$$\mathrm{grad}h(\mathbf{x}^*) = \left(\frac{\partial h(\mathbf{x}^*)}{\partial x_1}, \ldots, \frac{\partial h(\mathbf{x}^*)}{\partial x_m}\right) = \frac{exp\{-\mathbf{b}^T \mathbf{x}^*\}}{(1+exp\{-\mathbf{b}^T \mathbf{x}^*\})^2}(b_1, \ldots, b_m).$$

So, grad $h(\mathbf{x}^*)$ is always orthogonal to the (2) hyperplane. The $\mathbf{b}$ vector defines its direction, while the $q(\mathbf{x}^*)$ value determines its length. Since the coefficient vector determines the equation of the (2) hyperplane accurate to a factor, for practical applications it would be convenient to convert it to standard measure, i.e., $\mathbf{b}^0 = \frac{\mathbf{b}}{|\mathbf{b}|}$.

Let us explain the dividing plane coefficients b_j, $j = 1, 2, \ldots, m$. There is a vector $\mathbf{x}$. We replace one of its variables (coordinates), e.g., kth, with $\triangle x_k$. All the other variables are kept unchanged (fixed). As a result we obtain the new $\mathbf{x}^{'}$ vector that has $\forall j \neq k \; x_j^{'} = x_j$ and $x_k^{'} = x_k + \triangle x_k$.

After simple rearrangement, we obtain

$$h(\mathbf{x}^{'}) = \frac{h(\mathbf{x})}{h(\mathbf{x})(1-e^{-b_k \triangle x_k})+e^{-b_k \triangle x_k}} = \frac{h(\mathbf{x})}{h(\mathbf{x})+(1-h(\mathbf{x}))e^{-b_k \triangle x_k}},$$

$$\ln \frac{h(\mathbf{x}^{'})}{1-h(\mathbf{x}^{'})} - \ln \frac{h(\mathbf{x})}{1-h(\mathbf{x})} = b_k \triangle x_k,$$

that is, changing x_k variable by one while the other variable values are fixed changes $\ln \frac{h(\mathbf{x})}{1-h(\mathbf{x})}$ by b_k ones.

The $h(\mathbf{x}^{'})$ function's derivative with respect to increment $\triangle x_k$ varies directly with the b_k coefficient. Therefore, to increase $h(\mathbf{x}^{'})$ for positive b_k values the $\triangle x_k$ increments shall also be positive, and shall be negative for the negative values.

An object can be controlled by solving extremum problems. The essence of such problems is improving an object status through increasing the $h(\mathbf{x})$ probability by optimal change of the $\mathbf{x}$ vector of properties. Let us consider some of such problems.

Problem 1. Maximizing the $h(\mathbf{x})$ probability under constraints to the property changes is

$$\begin{cases} h(\mathbf{x}) \to max, \\ x_j = x_j^* + \triangle_j, j = 1, \ldots, m, \\ \triangle_j \in G_j, j = 1, \ldots, m, \end{cases} \tag{4}$$

where $\triangle_j$ is the jth component change, G_j is the acceptable region of the jth component changes.

The (4) problem does not account for economic constants and costs incurred by changing the x_j components. If we take them into account, we obtain the following problem

$$\begin{cases} h(\mathbf{x}) \to max, \\ x_j = x_j^* + \triangle_j, j = 1, \ldots, m, \\ \triangle_j \in G_j, j = 1, \ldots, m. \\ \upsilon_j(\triangle_j) \leq V_j, j = 1, \ldots, m, \end{cases} \tag{5}$$

where $\upsilon_j(\triangle_j)$ is the jth component change cost function, V_j is the highest acceptable cost of the jth component change.

Problem 2. Function $h(\mathbf{x})$ reaching the specified p_0 probability with min costs incurred by changing the $\mathbf{x}$ vector of social and economic properties. In this case the problem can be rendered as follows:

$$\begin{cases} \sum_{j=1}^{m} \upsilon_j(\triangle_j) \to min, \\ x_j = x_j^* + \triangle_j, j = 1, \ldots, m, \\ \triangle_j \in G_j, j = 1, \ldots, m. \\ h(\mathbf{x}) = p_0 \end{cases} \tag{6}$$

Direct consequences of a specific action applied to the object are evaluated as the goal function value in the (4)-(6) problems. Indirect consequences can be evaluated ad hoc from the social and economic property vector values obtained through solving an optimization problem.

Thus, the model represented by equations (1) and (2) has the primary properties of a mathematical model.

3 An example of logistic regression application to modeling the relation of life quality to regional social and economic indicators

Let us consider the forecasting capability of the logistic regression model with a specific example. We will study the correlation between life quality and the region's social and economic indicators. The data have been borrowed from the Russian Statistics Agency reports for 2013 [9]. This very year is chosen to avoid any misinterpretation, since in 2014 the tense international situation led to the Russian economy's growth slowdown, and then to its decline. Such periods of recessions strongly affect the life quality. To evaluate the life quality, we use the Human Development Index (HDI) [4].

It has been impossible to directly obtain statistically significant HDI vs. region's social and economic indicators relation (such as multiple regression equations). Let us consider the problem of categorizing regions into two categories: with low HDI ($y = -1$) and high HDI ($y = 1$).

In order to be able to compare the absolute values we convert them to a single relative scale by dividing the values by the region's annual average population size. To improve the level of certainty, the statistical data for 24 Russian regions have been rejected for the following reasons: very low HDI; very high HDI; HDI close to the average HDI over the entire selection; lack of data.

As a result, two categories have been established. The first category embraces the regions with low HDI (below 0.84). The total number of such Russian regions is 28. The second category embraces the regions with HDI equal to 0.85 or higher. The total number of such Russian regions is 31. The data used are listed in Table 1. Table 2 contains the estimated dividing hyperplane equation (2) coefficients. The equation has been developed with a learning sample.

The correct recognition probability over the entire learning sample is 0.966.

Extra parameters have to be specified to use logistic regression-based control (see problems (4)-(6)). Such parameters are region-specific.

Table 1: Social and economic indicators considered

Indicator	Symbol
Aggregate birth rate, units	X_1
Life expectancy at birth, years/100	X_2
Number of mortgages extended to the local residents, per 1,000 people	X_3
Fixed asset investments per capita expressed in effective prices, RUB per 100,000 people	X_4
Gross regional product (GRP) per capita, RUB per 1,000,000 people	X_5
Fixed asset investment to GRP ratio, units*10	X_6
Oncology-related mortality rate: number of cases per 10 million people	X_7
Paramedic salary to the average Russian region salary ratio, units*10	X_8
Nurse (pharmacist) salary to the average Russian region salary ratio, units*10	X_9
Total unemployment rate, %/10	X_{10}
Number of physicians, per 100,000 people	X_{11}
Consolidated utility maintenance budget comprised of the region's allocations and the National Utilities Fund allocations, 1,000 RUB per 10 people	X_{12}
Consolidated public health budget comprised of the region's allocations and the National Public Health Fund allocations, 1,000 RUB per 10 people	X_{13}

The logistic regression coefficients for X_1, X_6, X_7, X_9, X_{10}, X_{12}, and X_{13} appear negative. It means that if all other variables are fixed to improve the region's life quality ($h(\mathbf{x})$ probability) these variables are to be decreased. For the other variables the $h(\mathbf{x})$ probability increase is achieved by increasing their values (if the other variables are fixed).

For such variables as X_7 and X_{10} the reasons for having negative coefficients are obvious while the rest have to be explained.

First, the negative X_1 coefficient represents the phenomenon known as "life quality to birth rate negative feedback" [3].

Second, the negative X_6 coefficient can be explained by the fact that this variable is relative. Also, regions with higher living standards tend to use investments more efficiently. In other words, a more developed region (with higher life quality) spends less fixed asset investments than a less developed region to obtain the same gross regional product increase. For this reason, lower X_6 coefficient with other factors being equal means more efficient consumption of fixed asset investments.

Table 2: Estimated dividing hyperplane coefficients (3)

b_0	b_1	b_2	b_3	b_4	b_5	b_6
0,072	−2.381	0.807	0.879	10.004	3.974	−1.911
b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
−1.149	0.165	−1.770	−2.739	0.688	−4.223	−0.870

Third, the negative X_9, X_{12}, and X_{13} coefficients can be explained by inefficiency of the public health and utility services. The available statistical data show that alternative spending patterns lead to a more significant life quality increase. Certainly, this does not mean that public health and utility sector should not be developed.

The highest coefficient is for variable X_4. It is consistent with Keynesian theory: as the total amount of investments rises, the revenue is also increased by k-fold of the investment increase [5].

Conclusions

It has been shown that logistic regression can be applied as a diagnostic model for complex stochastic systems. The logistic regression coefficients are explained within the diagnostic model framework. An example of logistic regression application to studying the relation of life quality to regional social and economic indicators is given.

References

[1] Ashihmin V.N., Gitman M.B., Keller I.E., Naymark O.B., Stolbov V.Yu., Trusov P.V., Frik P.G. *Introduction to Mathematical Modeling*. Moscow, Logos Publishing, 2005. 440 p. (in Russian)

[2] Azen R., Walker C.M. *Categorical Data Analysis for the Behavioral and Social Sciences*. Routledge, 2011. 283 p.

[3] Becker G.S. The Economic Way of Looking at Life // *Journal of Political Economy*. 1993. Vol. 101. No. 3. pp 385-409.

[4] Jahan S. The Future of Human Development Measures // *United Nations Development Programme. Human Development Report*. 01 June 2016. URL: http://hdr.undp.org/en/content/future-human-development-measures (access date: April 3, 2017).

[5] Keynes J. *The General Theory of Employment, Interest and Money*. London, Macmillan, 1936. 301 p.

[6] Lachin J.M. *Biostatistical Methods: the Assessment of Relative Risks. - 2nd edition.* Wiley, 2011. 644 p.

[7] Magnus Ya.R., Katyshev P.K., Peresecky A.A. *Econometrics 101. Introductory course - 6th edition, amended.* Moscow, Delo Publishing, 2004. 576 p. (in Russian)

[8] Myatlev V.D., Panchenko L.A., Riznichenko G.Yu., Terekhin A.T. *Probability Theory and Mathematical Statistics. Mathematical Models.* Moscow, Academy Publishing, 2009. 320 p. (in Russian)

[9] Russian Regions. Social and Economic Performance. 2014: Statistical Report. / Russian Statistics Agency. Moscow, 2014. 900 p. (in Russian)

[10] Shoukri M.M., Pause C.A. *Statistical Methods for Health Sciences. - 2nd edition.* CRC Press, 1999. 390 p.

[11] Vorontsov K.V. *Regression Estimation Algorithm Lectures.* 2007. 37 p. 15.03.2016. URL: http://www.ccas.ru/voron/download/Regression.pdf (access date: April 3, 2017). (in Russian)

Time-Series Forecasting for Big Data

Natalia Galanova and Victor Demin
2GIS, Novosibirsk, Russia
e-mail: natalia.galanova@gmail.com

Abstract

In this paper we discuss in general the problems of time-series forecasting: seasonality detection, trend elimination and different types of adaptive forecasting models.

Keywords: trend, seasonality, time-series forecasting, exponential smoothing, adaptive models, big data.

Introduction

The problem of the time series forecasting is actual and was investigated by many researchers. Various forecasting models exist in economics, industry, science and other fields of life. Our company provides local search service and we often have a requirement to predict the amount of users attention, which can be measured in clicks, calls or some other users activity. However, we have a number of features and requirements for prediction models. The main feature is a big number of different times series - we have to forecast more than a million time series per day. Consequently, we can't use only one forecasting model. Moreover, time series can change their characteristics and for the same data at different timepoints may be necessary to use different forecasting models. In this paper, we consider an algorithm that is a combination of standard methods for identifying and elimination trends, identifying and taking into consideration seasonality and also discuss various adaptive forecasting models. In summary, such a mixture of algorithms allows us to forecast hundreds of thousands of times series per day, which can have a different structure, with acceptable accuracy.

1 Seasonality

In time series data, seasonality is the presence of variations that occur at specific regular intervals, such as weekly, monthly, or yearly. Seasonality may be caused by various factors, such as weather, vacation or holidays and consists of periodic, repetitive, and generally regular and predictable patterns in the levels of a time series. In our data, seasonality is based on the product features, like increasing of the users interest for tires in october or march. There are two main reasons for studying seasonal component: to understand data features and its seasonal effect, and to use the past patterns of the seasonal variations in forecasting.

Usually the detection of the seasonal components is based on the autocorrelation function. The autocorrelation function of a random process is the Pearson correlation between values of the process at different times, as a function of the time lag k:

$$p_k = \frac{\sum_{i=k+1}^{n} (X_i - X^k)(X_{i-k} - X^{k+1})}{\sqrt{\sum_{i=k+1}^{n} (X_i - X^k)^2 \sum_{i=k+1}^{n} (X_{i-k} - X^{k+1})^2}}, \tag{1}$$

where

$$X^k = \frac{\sum_{i=k+1}^{n} X_i}{n-k}, \; X^{k+1} = \frac{\sum_{i=k+1}^{n} X_{i-k}}{n-k}. \tag{2}$$

The most common approach to the detection of the seasonal components is based on the graphical analysis of the autocorrelation function in dependence of the lag value.

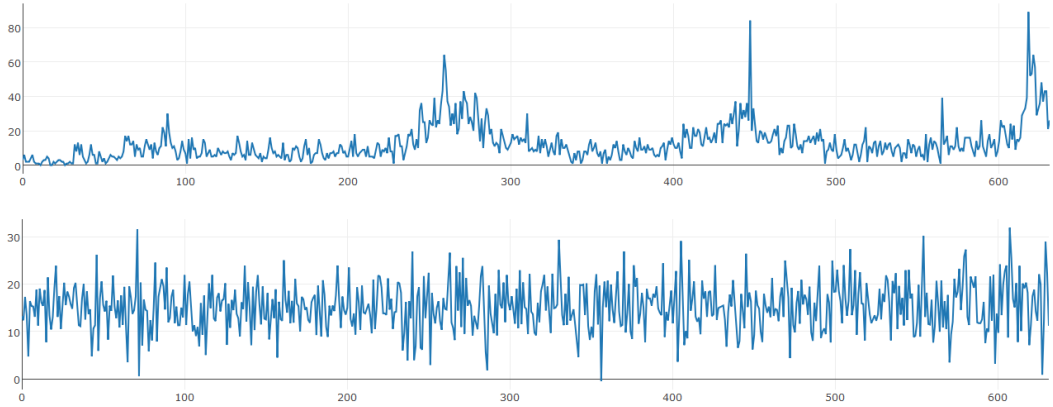


Figure 1: Time series with and without seasonality

Figure 1 presents two different time series: the first one is the real data which contains week and year seasonal component (this conclusion is based on our expert opinion). The second time series was generated as the normal random value as an example of the time series without any seasonal component. Figure 2 presents autocorrelation functions for lag values from 1 to 370. For the first time series we can see the periodicity in the values of the autocorrelation function, that means that we have year seasonal component. For the second time series we can not see any periodical variations in the autocorrelation function, which means that there is no any seasonal components in this data.

The main disadvantage of this approach is that it is applicable only if we have small amount of time series to analyze, and can use graphical analysis. But we work with thousands of time series and we can not detect seasonality by graphical procedures.

That is why we use approach based on testing the hypothesys of the autocorrelation significance for several fixed lag values. Based on the data features we fixed the following lags:

$$lags = \{7, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330, 365\}. \tag{3}$$

For testing autocorrelation significance for particular lag value k we need to check

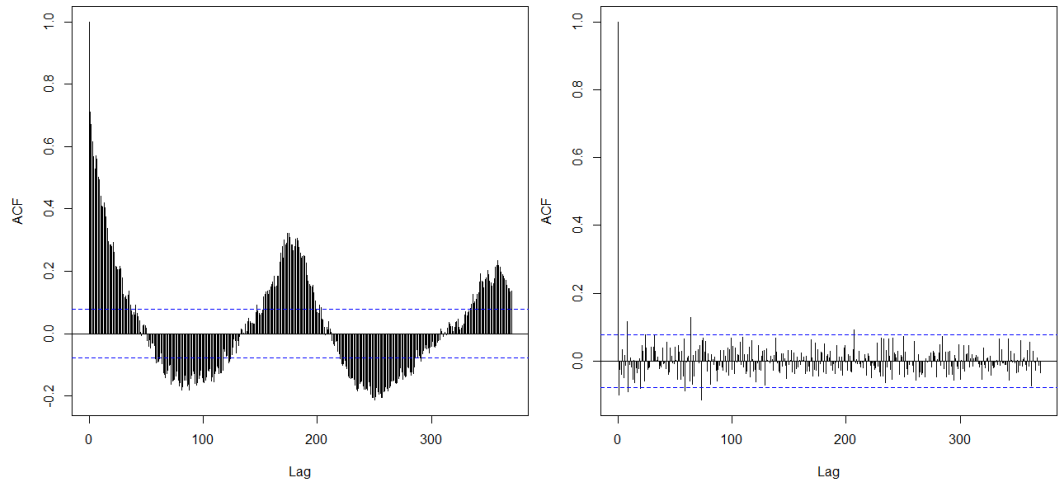


Figure 2: Autocorrelation functions for time series with and without seasonality

the null hypothesis of the following form:

$$H_0 : p_k = 0 \tag{4}$$

against the alternative hypothesis

$$H_1 : p_k > 0. \tag{5}$$

For testing this null hypothesis we can use statistic [3]:

$$t = \frac{p_k\sqrt{n-k-2}}{\sqrt{1-p_k^2}}. \tag{6}$$

Under the null hypothesys this statistic belongs to the Student's distribution with $n-k-2$ degrees of freedom.

Table 1: Significance testing of the autocorrelation values

lag	Data with seasonality			Data without seasonality		
	acf	S_n	$p-value$	acf	S_n	$p-value$
7	0.53	15.71	0.00	-0.01	-0.28	0.61
180	0.38	8.73	0.00	-0.04	-0.93	0.82
365	0.49	9.39	0.00	0.003	0.05	0.48

Table 1 presents results for significance testing of the autocorrelation function values for different lags. It can be seen from the table that lag values 7, 180 and 365 are significant for time series with seasonality. And for time series which are random normal data without seasonality there are no significant lags. Thus, testing significance of the autocorrelation function values helps us to automate seasonality detection procedure and apply it to the thousands of time series.

2 Trend

The main problem of the trend existence in the data is that it leads to the errors in the seasonality detection. As it was shown above, for seasonality detection we use an approach based on testing significance of the autocorrelation function for different lags. If we will apply this approach to the data with trend we will obtain incorrect results, because trend will increase the values of the autocorrelation function. That is why before the seasonality testing we need to check if there is a trend in the data. If we found out that trend is significant we need to extract it somehow.

To check if the data contain trend we need to test the following null hypothesis:

$$H_0 : m_i = m, i = 1, 2, ..., n. \tag{7}$$

Against the alternative hypothesis:

$$H_1 : |m_{i+1} - m_i| > 0, i = 1, 2, ..., n - 1. \tag{8}$$

For testing null hypothesis we use the inversion test [2]. Let say that there is an inversion in the data if we have the following situation:

$$x_i > x_j, \ i < j \leqslant n, \tag{9}$$

Test statistic of inversion test has the following form:

$$I^* = \frac{I - \frac{n(n-1)}{4}}{\sqrt{\frac{2n^3+3n^2-5n}{72}}}, \tag{10}$$

where I is the number of inversions in the data and under the null hypothesis test statistic belongs to the standard normal distribution. If we reject the null hypothesis we have to say that our data contain significant trend and we need to extract it.

The easiest way of trend extraction is to use a simple linear regression. And we can say that for our time series it works well for many cases. But also we have a problem of a piecewise linear trend, the example of such time series is presented at the Figure 3. Based on the data review we can say that in general we have one of the three situations: no trend in the data, simple linear trend or piecewise linear trend with one breakpoint.

The problem of the piecewise linear trend is that we need to find the optimal break-point. For finding the optimal breakpoint we use the approach suggested in [4].

So we can formulate the algorithm of the trend extraction:

1. First we test the null hypothesis of the form (7) by the inversion test (10);
2. If we reject the null hypothesis, which means that we have significant trend, we build two trend models based on the data - linear trend and piecewise linear trend;
3. Than we choose the trend model that has less sum of residuals.

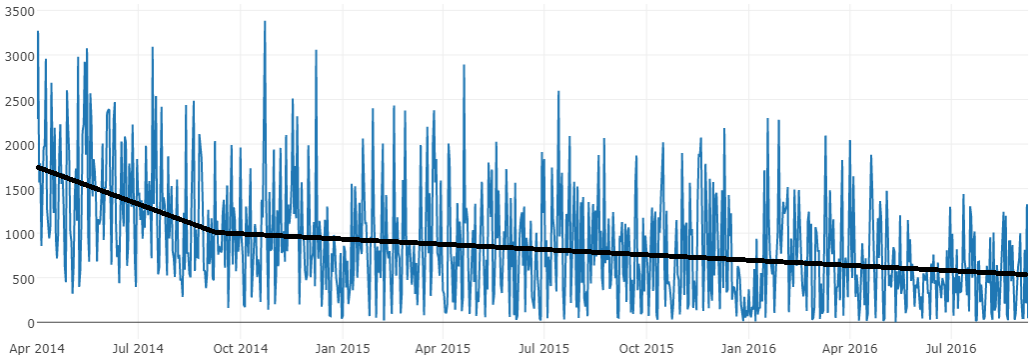


Figure 3: Time series with piecewise linear trend

3 Forecasting models

While choosing forecasting modes we faced the following main problems:

1. We have a great amount of data, at the same time we need to forecast more than a million of time series;

2. We need to calculate new forecast values each day and this procedure must be quick and effective;

3. We have a lot of time series for periods more than two years, so the models with parameters estimation (like autoregressive models) will work too long;

4. We have many different types of time series: with or without trend, having various seasonality components, time series with changepoints and so on.

Based on these problems we decided to use the class of adaptive forecasing models and also to use adaptive selection discussed below.

The simplest forecasting model we use is the exponential smoothing or Braun's model:

$$\widehat{a}_{1,t} = \alpha \cdot x_t + (1-\alpha) \cdot \widehat{a}_{1,t-1}, \tag{11}$$

where $\alpha \in [0,1]$ is the smoothing parameter. The forecast value at the timepoint t for the timepoint v can be defined as

$$\widehat{x}_v(t) = \widehat{a}_{1,t}. \tag{12}$$

The next model is the exponential smoothing model with week seasonal component:

$$\begin{aligned} \widehat{a}_{1,t} &= \alpha \cdot (x_t - g_{t-7}) + (1-\alpha) \cdot \widehat{a}_{1,t-1}, \\ \widehat{g}_t &= \alpha \cdot (x_t - \widehat{a}_{1,t}) + (1-\alpha) \cdot \widehat{g}_{t-7}, \end{aligned} \tag{13}$$

and the forecast value at the timepoint t for the timepoint v can be defined as

$$\widehat{x}_v(t) = \widehat{a}_{1,t} + \widehat{g}_{t-7+v\%7}. \tag{14}$$

But we often have more than one seasonal component in our data - like weekly, monthly and yearly seasonality at the same time series. For such cases we suggest the following model considering multiple seasonal components:

$$\begin{gathered} \widehat{a}_{1,t} = \alpha \cdot (x_t - g^*_{t-1}) + (1-\alpha) \cdot \widehat{a}_{1,t-1}, \\ g^*_t = \frac{1}{k} \sum_{i=1}^{k} \widehat{g}_{t-lags[i]}, \\ \widehat{g}_t = x_t - \widehat{a}_{1,t}, \end{gathered} \tag{15}$$

where *lags* is the list of the significant lag values, defined by the seasonality detection procedure discussed above.

The forecast value at the timepoint t for the timepoint v can be defined as

$$\widehat{x}_v(t) = \widehat{a}_{1,t} + g^*_{t+v}. \tag{16}$$

As we said before, our data can contain trend - linear or piesewise linear. That is why all these three models - simple exponential smoothing, exponential smoothing with week seasonality and multiseasonal model can be applied by two ways: for raw data or for data after trend elimination. If there is no significant trend in time series, we build forecasting models by the raw data $x_1, x_2, ..., x_n$ and then forecast value from timepoint t for the timepoint v can be defined as (12), (14) or (16) depending on the forecasting model. But if there is significant trend in the data, before building forecasting model we need to extract it. Let say that we can describe this trend by some function $f(t)$ which is linear or segmented linear function, so after trend elimination we work with data:

$$\begin{gathered} z_1, z_2, ..., z_n, \quad . \\ z_i = x_i - f(t_i) \end{gathered} \tag{17}$$

So the forecast value from timepoint t for timepoint v for time series with trend consists of two parts - forecast value of the trend function $f(t)$ and forecast value of the model built for data without trend $z_1, z_2, ..., z_n$:

$$\widehat{x}_v(t) = f(v) + \widehat{z}_v(t). \tag{18}$$

Often it is not obvious what model is the best for prediction. Even if there is a significant trend and seasonal components in the data - it is not necessary that model with trend and multiple seasonality is the best for prediction at the particular timepoint t. That is why we use approach based on the adaptive selection - for one time series we build several models at the same time and at the timepoint t when we need to calculate forecat value we choose the model, that at that moment has the smallest error:

$$\widetilde{e}_t = (1-\gamma) \cdot \widetilde{e}_{t-1} + \gamma \cdot |e_t|, \tag{19}$$

where e_t is the forecasting error of the model at the time t and γ is the smoothing parameter which is usually taken as 0.1.

4 Some results

Every month we recalculate trend significance and trend models for more than a million time series. Among our data we have 69% of time series with significant trend and 57% of them can be described by linear trends and 43% by segmented linear trends respectively. Also every month we restart seasonality detection procedure for our time series. At the moment only 30% of our time series were detected to have any seasonal components. Usually the problems of seasonality detection are based on the data insufficiency - time series may be too short to detect seasonality for some lag values or also there can be too many null values in the data. The "popularity" of seasonal components for our fixed lags (3) is presented in the Table 2.

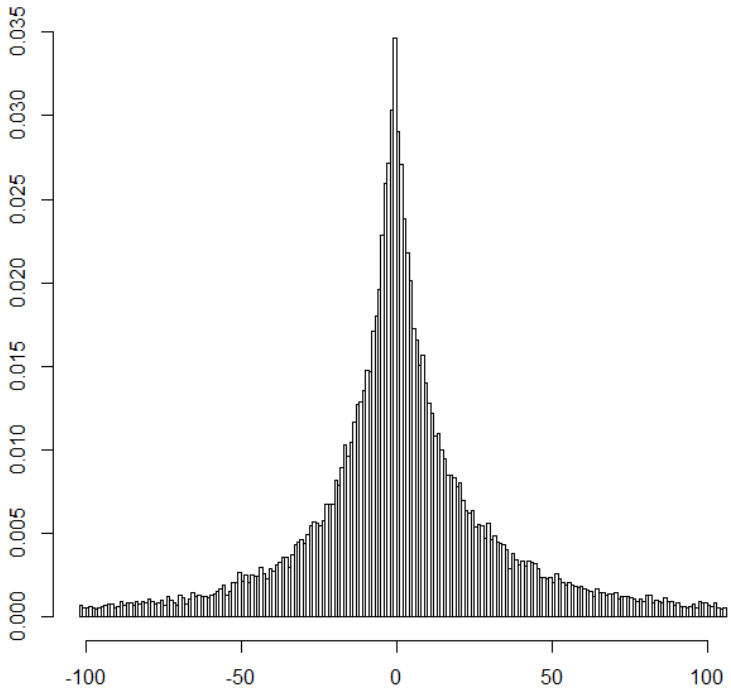


Figure 4: Forecast errors

Table 2: Frequency of seasonal components

Lag	7	30	60	90	120	150	180	210	240	270	300	330	365
time series %	45	39	32	39	35	23	23	31	22	20	28	27	28

Based on the business needs, we forecast our time series for thirty days period which is very long for adaptive models, because usually these models are used for short-term forecasting. But nevertheless we obtain acceptable accuracy. The Figure 4 presents histogramm for the forecast errors of users clicks on various firms. These errors are calculated by thirty days period and forecast values are based on the best

model, chosen by adaptive selection:

$$e = \sum_{i=1}^{30} \widehat{x}^{*}_{t+i}(t) - \sum_{i=1}^{30} fact_{t+i}. \tag{20}$$

As we can see from Figure 4 obtained errors are symmetric and have zero mean.

Conclusions

Thus, we have developed an algorithm that allows to make long-term forecasts for more than a million time series, which can have a different structure. Such an algorithm is a combination of methods for revealing and eliminating trend, identifying the seasonal components, various adaptive models and the model of adaptive selection. Suggested algorithm satisfies the speed and quality requirements of the received forecasts and is used to predict users attention for several thousand companies.

References

[1] Lukashin Ju.P. (2003). *Adaptivnye metody kratkosrochnogo prognozirovanija vremennyh rjadov*.Finance and Statistics, Moscow. (in Russian)

[2] Veretelnikova I.V., Lemeshko B.Yu. (2015). Tests for an Absence of Trend . *Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Nonparametric Approach"l.* pp. 80-91.

[3] Fjorster Je., Rjonc B. (1983). *Metody korreljacionnogo i regressionnogo analiza*.Finance and Statistics, Moscow. (in Russian)

[4] Muggeo V.M.R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine*. Vol. **22**, pp.3055-3071.

Anomalies Detection in Big Data Time Series

Victor Demin, Natalia Galanova, Anastasia Zamashchikova
2GIS, Novosibirsk, Russia
e-mail: vicdemin@gmail.com

Abstract

In this paper we discuss the problems of anomalies detection procedures in time series data. We consider various types of time series, having different structure: trend, seasonal components and changepoints. An algorithm for searching anomalies of different types, both outliers and changepoints in various time series is suggested.

Keywords: time series, anomalies, outliers, changepoints, trend, seasonality, big data.

Introduction

Our company presents local search service and provides information about more than 500000 companies in over 100 cities in Russia and other countries. Almost every week we release new features in our products, which present online, desktop and mobile applications, and of course there is always a chance of errors. We want to discover these potential errors as soon as it is possible.

We collect a lot of various data from users activity in our products and these data mostly present time series. Based on finding anomalies in these time series we want to solve the problem of errors detection after releases in our products.

We have several features that we need to take into account in our anomalies detection algorithm.

First of all, we work with a great amount of time series and, as a consequence, the time series may have different structures: time series mostly consisting of zeroes, time series with seasonal and trend components, time series with changepoints and so on. Besides, the series can change its structure with time.

Secondly, the types of anomalies in the data can be different: there can be outliers and changepoints in the data and also they can be mixed. We define as outlier an observation that is distant from other observations. As a changepoint we define the point of significant changes of the time series in the mean or variance values. In Figure 1 you can see the examples of time series with an outlier and a changepoint.

Thirdly, because our main goal is to find errors after releases, we are only interested in the anomalies that have occurred in the last few days and do not want to find changepoints or outliers in the "past". That means that we suppose that we have already found these anomalies after corresponding releases.

Forthly, we need an algorithm to be quick because we work with a lot of time series and need to analyse it every day.

And the last one - the number of false positive anomalies, found by this algorithm, must be reduced to a minimum. This requirement arises because we have to check

many time series every day and each found anomaly is going to be analysed by quality assessment engineers.

Generally, we can divide our time series into several types. First of all, there are sparse or/and short time series. We just skip such time series and do not look for anomalies in them because there is not enough data for analysis. Secondly, we have time series with high variability. The problem of such time series is that it is difficult even for a person to determine the existense of anomalies. That is why we also need to skip such time series because the attempts to find anomalies in them usually lead to a lot of false positive anomalies. And also we have a lot of time series with trend and seasonal components in various combinations: stationary time series without any seasonal components, nonstationary time series without any seasonal components, stationary time series with seasonality, nonstationary time series with seasonality and time series with one or more changepoints.

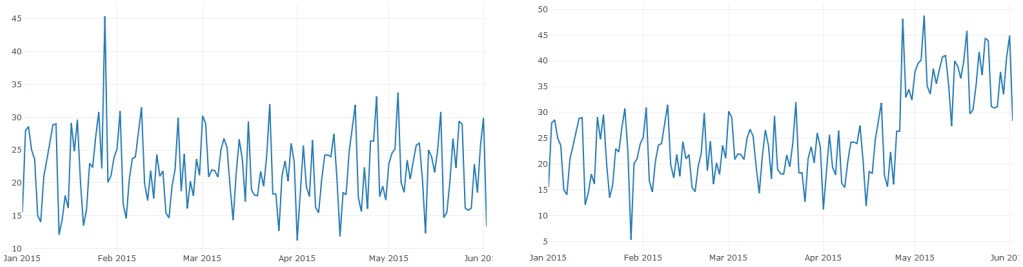


Figure 1: Time series with an outlier and a changepoint

1 ESD test

In our algorithm of anomalies detection we use ESD test. The generalized (extreme Studentized deviate) ESD test [1] is used to detect one or more outliers in the sample of values from approximately normal distribution. The generalized ESD test performs k separate tests: test for one outlier, test for two outliers and so on up to k outliers, where k is the upper bound for the suspected number of outliers in the data. The generalized ESD test is defined for the hypothesis:

$$\begin{aligned} H_0 &: \textit{There are no outliers in the data set} \\ H_1 &: \textit{There are up to k outliers in the data set} \end{aligned} \tag{1}$$

The algorithm of this test is the following:

1. We compute test statistic of the form:

$$R_i = \frac{max\,|x_i - \bar{x}|}{s}, \tag{2}$$

where $\bar{x}$ and s are sample mean and standard deviation respectively;

2. Remove the observation that maximizes $|x_i - \bar{x}|$;

3. Recompute the statistic (2) with $n-1$ observations;

4. Repeat the process 1-3 until k observations have been removed. This results in the k test statistics $R_1,\ R_2, ..., R_k$;

5. Corresponding to the k test statistics, compute the following k critical values:

$$\lambda_i = \frac{(n-i) \cdot t_{p,n-i-1}}{\sqrt{\left(n-i-1+t^2_{p,n-i-1}\right) \cdot (n-i+1)}},\ i = 1, 2, ..., k, \tag{3}$$

where $t_{p,v}$ is the $100p$ percentage point from the Student distribution with v degrees of freedom and

$$p = 1 - \frac{\alpha}{2 \cdot (n-i+1)}. \tag{4}$$

6. The number of outliers is determined by finding the largest i such that $R_i > \lambda_i$.

2 Algorithm

Let $X_n = \{X_1, ..., X_n\}$ to be a time series of size n, in which we want to find anomalies. The first step is to determine the type of the time series, because depending on it we should find anomalies in different ways. As we said before, our data can contain seasonality which can cause a lot of problems while finding anomalies. Since almost all of our data are statistics related to the users activity and it is always aggregated by days, we assume by default that there is always weekly seasonality in our data. As we know from our data on weekends and holidays the users behaviour is usually inadequate. That is why we exclude all weekends and holidays from the consideration. It helps us to avoid the problems related with the seasonality presence in prospect.

After "excluding" seasonality we need to make sure that we have enough data for the following analysis. If in the resulting time series more than 40% of zero observations or the number of observations is less than 15, we consider such data set as a sparse or/and short time series and do not try to find anomalies in it.

Then we need to check if our time series is stationary or not. To check if the data contain trend we need to test the following null hypothesis:

$$H_0 : m_i = m, i = 1, 2, ..., n. \tag{5}$$

Against the alternative hypothesis:

$$H_1 : |m_{i+1} - m_i| > 0, i = 1, 2, ..., n-1. \tag{6}$$

For testing null hypothesis we use the inversion test [2]. Let say that there is an inversion in the data if we have the following situation:

$$x_i > x_j,\ i < j \leqslant n, \tag{7}$$

Test statistic of inversion test has the following form:

$$I^* = \frac{I - \frac{n(n-1)}{4}}{\sqrt{\frac{2n^3+3n^2-5n}{72}}}, \tag{8}$$

where I is the number of inversions in the data and under the null hypothesis test statistic belongs to the standard normal distribution. If we reject the null hypothesis we have to say that our data contain significant trend and then from the original time series X_n we go to the time series of differences:

$$Y_{n-1} = \{Y_1, ..., Y_{n-1}\} = \{(X_2 - X_1), ..., (X_n - X_{n-1})\}. \tag{9}$$

After eliminating trend by taking differences (if it was nesessary) we go to finding changepoints in the resulting time series. Note that at this step we could find only changepoints in variance, not in mean value of the time series. For searching changepoints we use the algorithm from the *changepoint* R package [3]. As we said before - we are only interested in finding anomalies in the last few days, so if we found any changepoints - we just neglect all the data which are earlier than the latest of found changepoints.

After that we are sure that our resulting time series has no trend, no changepoints in means or variances and no seasonal components, so we can say that at this step we work with the sample of independent and identically distributed values which can contain outliers. For finding outliers in this sample we use ESD test which is described above.

If we found any outliers by the ESD test, we remove them from the data and if we worked with the series of differences, we return to the original time series but already without outlier values. And now we again try to find changepoints in our data without outliers. At this step we already could find changepoints in mean values. This algorithm can return a great number of changepoints, because it is extremely sensitive to outliers, trends and seasonality in the data. But at this step of the algorithm we have already discarded seasonality and outliers from our time series. However, the elimination of trend is a very difficult task if there are changepoints in the data. Therefore, after finding the changepoints, we do an additional check: we construct two linear regression models for the left and for the right area of the point that is supposed to be a changepoint and based on the difference of the corresponding slope values we make a decision - if this point is a significant changepoint or it is not. If we found any siginificant changepoints - we take into consideration just the latest one and "forget" about the previous.

So, as a result we can obtain the following types of anomalies: changepoint in variances, changepoint in means and outliers.

3 Example

Let's show how the algorithm works on the example of the time series shown in Figure 2. This series is extremely difficult for searching anomalies, since it has seasonality, trend, outliers and the changepoint.

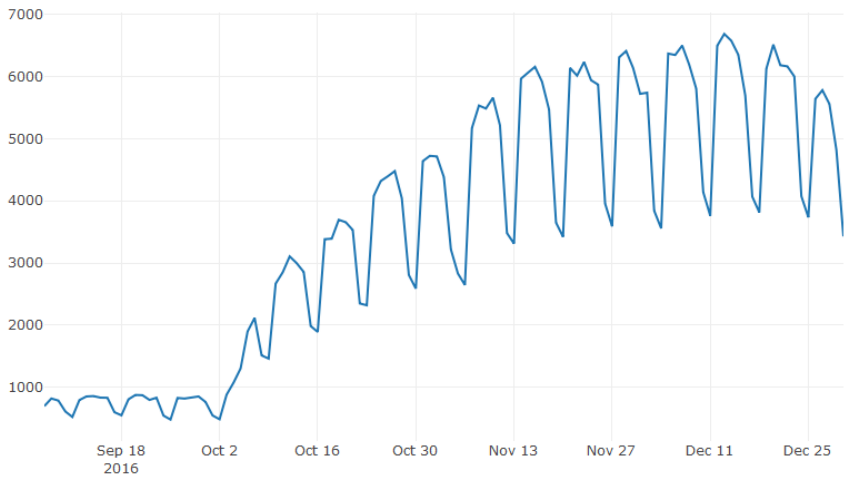


Figure 2: Initial time series for finding anomalies

The first step is to throw out the weekends and holidays. As a result, we obtain a series presented in Figure 3.

Figure 3: Time series without weekends and holidays

Then we need to check the hypothesis about trend existence in the data. We do it with the inversion test and obtained statistic the following: $I^* = 9.806911$ and $p-value = 0$. Since we reject the null hypothesis we can say that there is a significant trend in the data and we need to go to the series of differences. In the obtained time series we look for a changepoint. The time series of differences and found changepoint in the variances are shown in the Figure 4.

After that, we exclude the left side of the time series and in the right side we

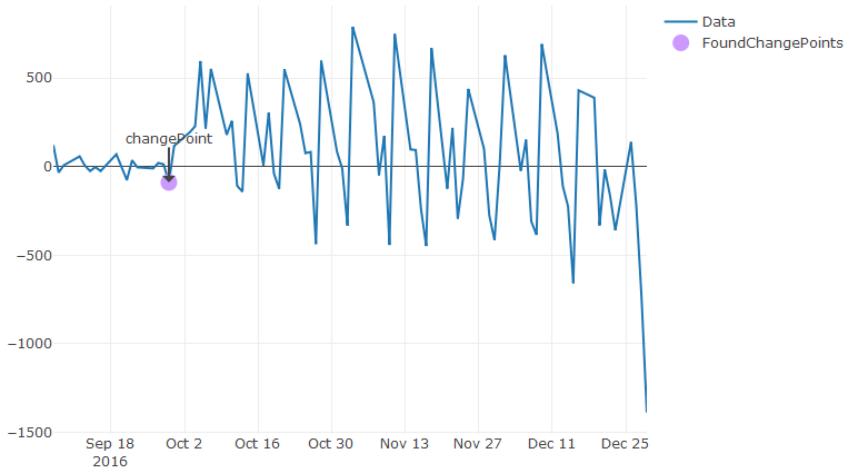


Figure 4: Changepoint in variances, found in the time series of differences

look for outliers using ESD test. Figure 5 shows that the ESD test found one outlier, which we need to exclude from the data.

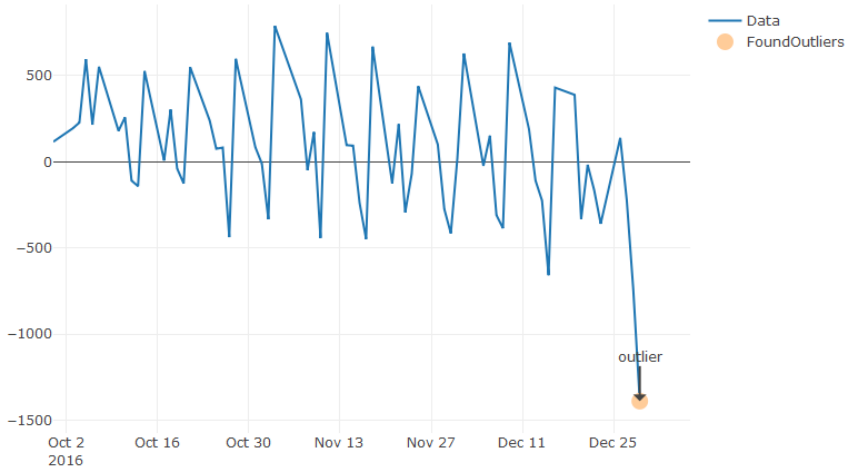


Figure 5: Outlier found by the ESD test

Then we get back from the differences to the original time series without outliers and in the resulting time series we again look for changepoints. Three found changepoints in means are shown in Figure 6. And after our additional check, based on the linear regressions, only two changepoint remains as siginificant: the first and the last one.

As a result, we received one outlier, one changepoint in variances and two changepoints in means, but we will only inform about the last changepoint in means and

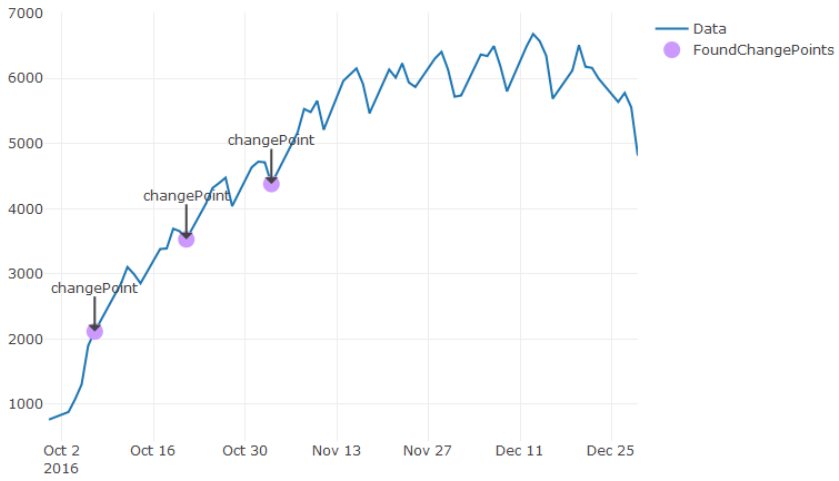


Figure 6: Changepoints in means, found in the original time series without outliers

the outlier, since the other two changepoints happened too long ago and they are no longer of our interest.

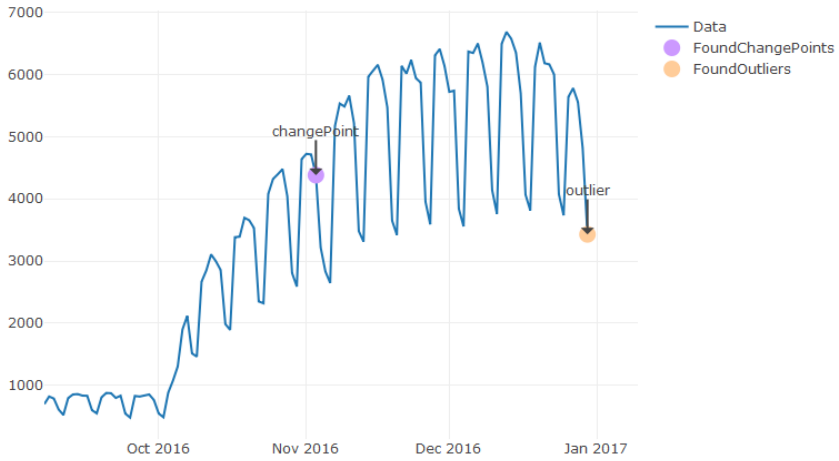


Figure 7: Initial time series and found anomalies

Conclusions

Thereby, this paper provides an algorithm for searching anomalies in the time series which present data grouped by days. And it is assumed that these time series possibly can contain weekly seasonality. The main feature of this algorithm is that it can work with time series of any type: sparse and/or short time series, noise series, series with

a trend, series with changepoints and outliers. This is achieved due to the fact that great attention is paid to the preliminary analysis of the time series - declining "problematic" time series, elimination of weekends and holidays and taking differences if there is significant trend in the data. And also the algorithm uses the different tools for searching both outliers and changepoints. In addition, this algorithm is oriented to the particular practical application and that is why we ignore some of the found anomalies. But this is a restriction of the algorithm's application and the algoritm by itself is able to find all anomalies in the data.

References

[1] Rosner B. (1983).Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*. Vol. **25 (2)**, pp.165-172.

[2] Veretelnikova I.V., Lemeshko B.Yu. (2015). Tests for an Absence of Trend . *Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Nonparametric Approach"l*. pp. 80-91.

[3] Killick R., Haynes K. and Eckley I.A. (2016). changepoint: An R package for changepoint analysis. *R package version 2.2.2, <URL: https://CRAN.R-project.org/package=changepoint>*.

APPLIED METHODS OF STATISTICAL ANALYSIS.
NONPARAMETRIC METHODS IN CYBERNETICS AND SYSTEM ANALYSIS

Proceedings

of the international workshop

Krasnoyarsk, 18-22 September 2017

Научное издание

ПРИКЛАДНЫЕ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА.
НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ В КИБЕРНЕТИКЕ И СИСТЕМНОМ АНАЛИЗЕ

ТРУДЫ

Международного семинара

Красноярск, 18-22 сентября 2017 г.

На английском языке

Подписано в печать 29.08.2017. Формат 60 × 84 1/8.
Усл. печ. л. 47,5. Уч.-изд. л. 88,35.
Тираж 50 экз. Заказ № 1069.

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20