BELARUSIAN STATE UNIVERSITY

VIENNA UNIVERSITY OF TECHNOLOGY

RESEARCH INSTITUTE FOR APPLIED PROBLEMS
OF MATHEMATICS AND INFORMATICS

BELARUSIAN REPUBLICAN FOUNDATION
FOR FUNDAMENTAL RESEARCH

BELARUSIAN STATISTICAL ASSOCIATION

COMPUTER DATA ANALYSIS AND MODELING:

THEORETICAL AND APPLIED STOCHASTICS

Proceedings of the Eleventh International Conference
Minsk, September 6-10

MINSK
Publishing center BSU
2016

UDC 519.22(06)

Editors: Prof. Dr. 5. *Aivazian,* Prof. Dr. *P. Filzmoser,* Prof. Dr. *Y. Kharin*

Computer Data Analysis and Modeling: Theoretical and Applied Stochastics : Proc. of the Eleventh Intern. Conf., Minsk, Sept. 6—10,2016. — Minsk : Publishing center of BSU, 2016. — 315 p.

ISBN 978-985-553-366-6.

This collection of papers includes proceedings of the Eleventh International Conference

"Computer Data Analysis and Modeling: Theoretical and Applied Stochastics" organized by the Belarusian State University and held in September 2016 in Minsk. Papers are reviewed by qualified researchers from Belarus, Russia, Austria, Germany, Lithuania, Poland. The papers are devoted to the topical problems: robust and nonparametric data analysis; statistical analysis of time series and forecasting; multivariate data analysis; design of experiments; probability and statistical analysis of discrete data; econometric analysis and modeling; survey analysis and official statistics; computer intensive methods, algorithms and software; computer data analysis in applications.

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

UDC 519.22 (06)

ISBN 978-985-553-366-6

PROGRAM COMMITTEE

Honorary Chair

S. Ablameyko — Rector of the Belarusian State University (Minsk, Belarus)

Co-Chairs

Prof. Dr. S. Aivazian (Moscow, Russia)
Prof. Dr. P. Filzmoser (Vienna, Austria)
Prof. Dr. Yu. Kharin (Minsk, Belarus)

Members

W. Charemza (Leicester, United Kingdom)
D. Chibisov (Moscow, Russia)
C. Croux (Leuven, Belgium)
K. Ducinskas (Klaipeda, Lithuania)
R. Dutter (Vienna, Austria)
G. Dzemyda (Vilnius, Lithuania)
A. Egorov (Minsk, Belarus)
H. Friedl (Graz, Austria)
I. Gaishun (Minsk, Belarus)
K. Hron (Olomouc, Czech Republic)
V. Krasnoproshin (Minsk, Belarus)
B. Lemeshko (Novosibirsk, Russia)
G. Medvedev (Minsk, Belarus)
V. Mkhitarian (Moscow, Russia)
Yu. Mishura (Kiev, Ukraine)
S. Morgenthaler (Lausanne, Switzerland)
V. Mukha (Minsk, Belarus)
Yu. Pavlov (Petrozavodsk, Russia)
D. Perrotta (Ispra, Italy)
G. Shevlyakov (St.-Petersburg, Russia)
A. Shiryaev (Moscow, Russia)
E. Stoimenova (Sofia, Bulgaria)
M. Templ (Vienna, Austria)
N. Troush (Minsk, Belarus)
A. Tuzikov (Minsk, Belarus)
V. Witkovsky (Bratislava, Slovak Republic)
A. Zaigrajew (Torun, Poland)
A. Zubkov (Moscow, Russia)

Local Organizing Committee

Co-chairs

P. Mandrik, Dean of the Faculty of Applied Mathematics and Informatics
Yu. Kharin, Director of the Research Institute for Applied Problems of Mathematics and Informatics

Members

V. Malugin — Conference Secretary
E. Ageeva, I. Bodiagin, O. Golos, A. Kharin, O. Kutsapalova, M. Mitskevich, E. Orlova, I. Pirshtuk, S. Staleuskaya, E. Vecherko, V. Voloshko

To the 95th Anniversary
of the Belarusian State University

To the 15th Anniversary
of the Research Institute for Applied Problems
of Mathematics and Informatics

PREFACE

The Eleventh International Conference "Computer Data Analysis and Modeling: Theoretical and Applied Stochastics" (CDAM'2016) organized by the Belarusian State University on September 6-10, 2016, is devoted to the topical problems in computer data analysis and modeling. Statistical methods of computer data analysis and modeling are widely used in variety of fields: computer support of scientific research; decision making in economics, business, engineering, medicine end ecology; statistical modeling of complex systems of different nature and purpose. In the Republic of Belarus computer data analysis and modeling have been developed successfully for the last 35 years. Scientific conferences CDAM were held in September 1988, December 1990, December 1992, September 1995, June 1998, September 2001, September 2004, September 2007, September 2010, and September 2013 in Minsk.

The Proceedings of the CDAM'2016 contain 81 papers. The topics of the papers correspond to the following scientific problems: robust and nonparametric statistical analysis of time series and forecasting, multivariate data analysis, statistical classification and pattern recognition, signal processing, statistical modeling, modeling of complex systems in different applications, statistics in economics, finance and other fields, software for data analysis and statistical modeling.

The Organizing Committee of the CDAM2016 makes its acknowledgements to Belarusian State University, Research Institute for Applied Problems of Mathematics and Informatics, Belarusian Republican Foundation for Fundamental Research, Vienna University of Technology, software company "ITransition", Belarusian Science and Technology Association "Infopark", and the CDAMCSS project within the OeAD's programme IMPULSE for financial support.

S. Aivazian
P. Filzmoser
Yu. Kharin

CONTENTS

PLENARY LECTURES

Datta S., Dutta S. A Rank-Sum Test for Clustered Data When the Number of Subjects in a Group within a Cluster is Informative 12

Dürre A., Fried R., Vogel D. The Spatial Sign Covariance Matrix and its Application for Robust Correlation Estimation 13

Fokianos K. Binary and Count Time Series Analysis 20

Jakimauskas G., Sakalauskas L. Implementation of the Poisson-Gaussian Regression Model in Empirical Bayes Estimation of Small Probabilities 21

Kharin Yu.S., Zhurak M. Statistical Analysis of Discrete Spatio-Temporal Data by Conditional Autoregressive Models 25

Matilainen M., Miettinen J., Nordhausen K., Oja H., Taskinen S. ICA and Stochastic Volatility Models 30

Mishura Yu.S. Mixed Power Variations with Statistical Applications 38

Monti G.S., Filzmoser P., Deutsch R. Robust Estimation Approach for Hazardous Concentration Levels Using Species Sensitivity Distribution 44

Oliveira M.R., Vilela M., Pacheco A., Valadas R., Salvador P. Extracting Information from Interval Data Using Symbolic Principal Component Analysis 45

Orsingher E. Some Fractional Extensions of the Poisson Process 53

Serov A.A., Zubkov A.M. Two-Sided Inequalities for the Average Number of Elements in the Union of Images of Finite Set Under Iterations of Random Equiprobable Mappings 55

Shevlyakov G.L., Vasilevskiy N.V. Performance Study of Linfoot's Informational Correlation Coefficient and its Modification 58

Stabingis G., Bernatavičienė J., Dzemyda G., Imbrasienė D., Paunksnis A. Automated Classification of Arteries and Veins in the Retinal Blood Vasculature 64

Stoimenova E. Comparison of Partially Ranked Lists 68

Vencalek O. Data Depth and its Applications in Classification 74

Zaigrajew A., Alama-Bućko M. Optimal Choice of Order Statistics Under Confidence Region Estimation in Case of Large Samples 79

SECTION 1
ROBUST AND MULTIVARIATE DATA ANALYSIS

Abdushukurov A.A., Muradov R.S. Estimation of Two-Dimensional Survival Function by Random Right Censored Data 81

Brodinova S., Zaharieva M., Filzmoser P., Ortner T., Breiteneder C. Group Detection in the Context of Imbalanced Data 87

Chernov S.Y. Error Probabilities in Sequential Testing of Simple Hypotheses for Dependent Observations 88

Hoang H.S., Baraille R. On Stochastic Perturbation Method for Estimation of High Dimensional Matrix 90

Hoffmann I., Filzmoser P., Croux C. Robust and Sparse Multiclass Classification by the Optimal Scoring Approach 94

Kharin A., Ton T.T. Evaluation of Sequential Test Characteristics for Time Series with a Trend 96

Nikolov N.I. Lee Distance in Two-Sample Rank Tests 100

Ortner T., Filzmoser P., Brodinova S., Zaharieva M., Breiteneder C. Forward Projection for High-Dimensional Data 104

Savelov M.P. On the Sequential Chi-Square Test 105

Smirnov P.O., Shirokov I.S., Shevlyakov G.L. On Approximation of the Q_n-estimate of Scale by Highly Robust and Efficient M-estimates 107

Walach J., Filzmoser P., Hron K., Walczak B. A Pairwise Log-Ratio Method for the Identification of Biomarkers 111

Zhuk E.E., Dus D.D. Assignment of Arbitrarily Distributed Random Samples to the Fixed Probability Distribution and its Risk 112

SECTION 2
STATISTICAL ANALYSIS OF TIME SERIES AND SPATIAL DATA

Ageeva H. Forecasting of Regression Model Under Classification of the Dependent Variable 117

Badziahin I.A. On Parameter Estimation of Stationary Gaussian Time Series Observed Under Right Censoring 121

Baranovskiy A.G., Troush N.N. Analysis of Self-Similarity Property of α-stable Processes 122

Chibisov D.M. Asymptotic Optimality of the Chi-Square Test in the Class of Permutation-Invariant Tests 126

Ducinskas K., Dreiziene L. Expected Error Rates in Classification of Gaussian CAR Observations 127

Egorov A.D. Evaluation of Expectation of a Class of Poisson Functionals 131

Nikitsionak V.I., Bachar A.M. Significance Level Analysis for Adaptive Algorithm of Stationary Poisson Stream Processing 135

Orlova E.N. Statistical Analysis of Markov Chains with the Periodically Changed Transition Probability Matrices 136

Ralchenko K.V. Consistent Estimators of Drift Parameter in Stochastic Differential Equations Driven by Fractional Brownian Motion 138

Sakhno L.M. Statistical Inference for Random Fields in the Spectral Domain Based on Tapered Data 142

Svidrytski A., Yatskou M., Apanasovich V. An Improved k-Nearest Neighbors Algorithm for the Analysis of Two-Color DNA Microarray Data with Spot Quality Factors 143

Vorobeychikov S.E., Burkatovskaya Y.B. Guaranteed Change Point Detection of Linear Autoregressive Processes with Unknown Noise Variance 147

SECTION 3
PROBABILISTIC AND STATISTICAL ANALYSIS OF DISCRETE DATA

Cheplyukova I.A. On the Limit Distribution of the Maximum Vertex Degree in a Conditional Configuration Graph 152

Filina M.V., Zubkov A.M. Some Remarks on the Noncentral Pearson Statistics Distributions 156

Gurevich G., Vexler A., Zhao Y. Modern Empirical Likelihood Concepts 160

Iskakova A.S. Modeling Unbiased Estimators with Good Asymptotic Properties for the Sum of Multivariate Discrete Independent Random Variables 165

Kharin Yu.S., Maltsew M.V. On One Generalization of Markov Chain with Partial Connections 168

Leri M.M. On Robustness of Configuration Graphs with Random Node Degree Distribution 172

Menshenin D.O. Asymptotic Properties of Binary Sequences Obtained by the Neumann Transform 176

Pavlov Yu.L. On Random Graphs in Random Environment 178

Radavičius M. Hoeffding Type Inequalities for Likelihood Ratio Test Statistic .. 182

Voloshko V. Steganographic Capacity of Locally Uniform Markov Covers 185

Zubkov A.M., Kruglov V.I. On Coincidences of Tuples in a Binary Tree with Randomly Labeled Vertices 190

SECTION 4
ECONOMETRIC MODELING AND FINANCIAL MATHEMATICS

Kirlitsa V.P. Exact D-optimal Designs Experiments for Linear Model with Heteroscedastic Observations 195

Lappo P.M., Yakushava T.A. Some Approaches to Classification of Subjects of Foreign Economic Activity by Risk Level 199

Lialikova V.I., Khatskevich G.A. Modeling the Regions of Belarus Competitiveness Based on Panel Data 201

Malugin V.I., Novopoltsev A.Yu. Statistical Estimation and Testing of Turning Points in Multivariate Regime-Switching Models 205

Medvedev G.A. On the Probability Distribution Processes Some Models of Interest Rates 210

Mukha V.S. Minimum Distance from Point to Linear Variety in Euclidean Space of the Two-Dimensional Matrices 215

Navitskaya K., Zhalezka B. Modeling of Regional Socio-Economic Development of Belarus 219

Novopoltsev A.Yu. Multivariate Linear Regression with Heterogeneous Structure and Asymmetric Distributions of Errors 223

Zmitrovich A.I., Krivko-Krasko A.V., Lysenko T.V. Automated Report on the Business Plan of the Investment Project 227

Zuev N.M. Calculation of European Options with Absolute Criteria 231

SECTION 5
SURVEY ANALYSIS AND OFFICIAL STATISTICS

Agabekova N. Technique of Calculating the Gender and Age Scales of Real Consumer Expenditure 234

Bokun N. Micro-Entities and Small Enterprises Surveys in Belarus 237

Kolesnikova I. The R&D Intensity Factors of GDP 243

Kulak A., Sharilova Y. Statistical Assessment of Gender Issues in Social and Labor Sphere 247

Matkovskaya O. Application of Method Lags Model for the Analysis of the Impact Ecology on Health 250

Novikov M.M. The Relevant Leading Indicator of Macroeconomic Dynamics ... 253

Sharilova Y., Kulak A. The Statistical Validity of the Increase in Retirement Age in the Republic of Belarus 256

Soshnikova L.A. Methodological Approaches to the Reflection of Environmental Assets in SEEA and NAMEA 259

Visotski S. Modeling Competitive Advantage of Territories 263

SECTION 6
COMPUTER DATA ANALYSIS AND MODELING IN APPLICATIONS

Baklanova O., Baklanov A., Shvets O. Multivariate Analysis for Image Recognition System to Assess the Quality of the Mineral Species 268

Hubin A.A., Storvik G.O. On Mode Jumping in MCMC for Bayesian Variable Selection within GLMM 274

Kharin A., Filzmoser P., Gabko P. Development of the Master Program on Applied Computer Data Analysis within the Tempus Project "Applied Computing in Engineering and Science" 278

Klimenok V.I. Unreliable Queueing System with Backup Server 280

Kolchin A.V., Ionkina H.G. On Some Aspects in Acquisition of Brain Electrical Activity 284

Matalytski M., Naumenko V., Kopats D. Analysis and Application of G-network with Incomes and Random Waiting Time of Negative Customers 287

Minkevičius S., Greičius E. On the Inequality in Open Multiserver Queueing Networks 291

Pavlova O.S., Malugin V.I., Ogurtsova S.E., Novopoltsev A.Yu., Byk I.S., Gorbat T.V., Liventseva M.M., Mrochek A.G. Computer Analysis of Essential Hypertension Risk on the Base of Genetic and Environmental Factors 296

Sergeev R.S., Kavaliou I.S., Tuzikov A.V., Sprindzuk M.V. Bioinformatics Analysis of M.TUBERCULOSIS Whole-Genome Sequences 300

Starodubtsev I.E. Fractal Dimension as a Characteristic of Biological Cell AFM Images 304

Varlamov O.O., Danilkin I.A. Knowledge Representation and Reasoning. Mivar Technologies 308

PLENARY LECTURES

A RANK-SUM TEST FOR CLUSTERED DATA WHEN THE NUMBER OF SUBJECTS IN A GROUP WITHIN A CLUSTER IS INFORMATIVE

Somnath Datta[1], Sandipan Dutta[2]
[1]*Department of Biostatistics, University of Florida*
[2]*Department of Bioinformatics and Biostatistics, University of Louisville*
[1]*Gainesville and* [2]*Louisville, USA*
e-mail: [1]somnath.datta@ufl.edu, [2]sandipan.dutta@louisville.edu

Abstract

The Wilcoxon rank-sum test is a popular nonparametric test for comparing two independent populations (groups). In recent years, there have been renewed attempts in extending the Wilcoxon rank sum test for clustered data, one of which [1] addresses the issue of informative cluster size, i.e., when the outcomes and the cluster size are correlated. We are faced with a situation where the group specific marginal distribution in a cluster depends on the number of observations in that group (i.e., the intra-cluster group size). We develop a novel extension of the rank-sum test for handling this situation. We compare the performance of our test with the Datta-Satten test, as well as the naive Wilcoxon rank sum test. Using a naturally occurring simulation model of informative intra-cluster group size, we show that only our test maintains the correct size. We also compare our test with a classical signed rank test based on averages of the outcome values in each group paired by the cluster membership. While this test maintains the size, it has lower power than our test. Extensions to multiple group comparisons and the case of clusters not having samples from all groups are also discussed. We apply our test to determine whether there are differences in the attachment loss between the upper and lower teeth and between mesial and buccal sites of periodontal patients.

Keywords: correlated data, dental data, nonparametric tests, Wilcoxon rank-sum test, within-cluster resampling

References

[1] Datta S., Satten G. A. (2005). Rank-Sum Tests for Clustered Data. *Journal of the American Statistical Association*. Vol. **471**, pp. 908–915.

THE SPATIAL SIGN COVARIANCE MATRIX AND ITS APPLICATION FOR ROBUST CORRELATION ESTIMATION

A. Dürre[1], R. Fried[2], D. Vogel[3]
[1,2]*Fakultät Statistik, Technische Universität Dortmund*
Dortmund, GERMANY
[3]*Institute for Complex Systems and Mathematical Biology, University of Aberdeen*
Aberdeen, UNITED KINGDOM
e-mail: [1]alexander.duerre@udo.edu

Abstract

We summarize properties of the spatial sign covariance matrix and especially look at the relationship between its eigenvalues and those of the shape matrix of an elliptical distribution. The explicit relationship known in the bivariate case was used to construct the spatial sign correlation coefficient, which is a non-parametric and robust estimator for the correlation coefficient within the elliptical model. We consider a multivariate generalization, which we call the multivariate spatial sign correlation matrix.

1 Introduction

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote a sample of independent p dimensional random variables from a distribution F and $s : \mathbb{R}^p \to \mathbb{R}^p$ with $s(\mathbf{x}) = \mathbf{x}/|\mathbf{x}|$ for $\mathbf{x} \neq 0$ and $s(0) = 0$ the spatial sign, then

$$S_n(\mathbf{t}_n, \mathbf{X}_1, \ldots, \mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^{n} s(\mathbf{X}_i - \mathbf{t}_n) s(\mathbf{X}_i - \mathbf{t}_n)'$$

denotes the empirical spatial sign covariance matrix (SSCM) with location $\mathbf{t}_n$. The canonical choice for the location estimator $\mathbf{t}_n$ is the spatial median

$$\boldsymbol{\mu}_n = \underset{\boldsymbol{\mu} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} ||\mathbf{X}_i - \boldsymbol{\mu}||.$$

Beside its nice robustness properties like an asymptotic breakdown-point of 1/2, it has (under regularity conditions, see [12]) the advantageous feature that it centres the spatial signs, i.e.,

$$\frac{1}{n} \sum_{i=1}^{n} s(\mathbf{X}_i - \boldsymbol{\mu}_n) = 0,$$

so that $S_n(\boldsymbol{\mu}_n, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ is indeed the empirical covariance matrix of the spatial signs of the data. If $\mathbf{t}_n$ is (strongly) consistent for a location $\mathbf{t} \in \mathbb{R}$, it was shown

in [5] that under mild conditions on F the empirical SSCM is a (strongly) consistent estimator for its population counterpart

$$S(\mathbf{X}) = \mathbb{E}(s(\mathbf{X} - \mathbf{t})s(\mathbf{X} - \mathbf{t})').$$

There are some nice results if F is within the class of continuous elliptical distributions, which means that F possesses a density of the form

$$f(\mathbf{x}) = \det(V)^{-\frac{1}{2}} g((\mathbf{x} - \boldsymbol{\mu})V^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

for a location $\boldsymbol{\mu} \in \mathbb{R}^p$, a symmetric and positive definite shape matrix $V \in \mathbb{R}^{p \times p}$ and a function $g : \mathbb{R} \to \mathbb{R}$, which is often called the elliptical generator. Prominent members of the elliptical family are the multivariate normal distribution and elliptical t-distributions (e.g. [2], p. 208). If second moments exists, then $\boldsymbol{\mu}$ is the expectation of $\mathbf{X} \sim F$, and V a multiple of the covariance matrix. The shape matrix V is unique only up to a multiplicative constant. In the following, we consider the trace-normalized shape matrix $V_0 = V/\mathrm{tr}(V)$, which is convenient since $S(\mathbf{X})$ also has trace 1. If F is elliptical, then $S(\mathbf{X})$ and V share the same eigenvectors and the respective eigenvalues have the same ordering. For this reason, the SSCM has been proposed for robust principal component analysis (e.g. [13, 15]). In the present article, we study the eigenvalues of the SSCM.

2 Eigenvalues of the SSCM

Let $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ denote the eigenvalues of V_0 and $\delta_1 \geq \ldots \geq \delta_p \geq 0$ those of $S(\mathbf{X})$. Explicit formulae that relate the δ_i to the λ_i are only known for $p = 2$ (see [19, 3]), namely

$$\delta_i = \frac{\sqrt{\lambda}_i}{\sqrt{\lambda}_1 + \sqrt{\lambda}_2}, \ i = 1, 2. \tag{1}$$

Assuming $\lambda_2 > 0$, we have $\delta_1/\delta_2 = \sqrt{\lambda_1/\lambda_2} \leq \lambda_1/\lambda_2$, thus the eigenvalues of the SSCM are closer together than those of the corresponding shape matrix. It is shown in [8] that this holds true for arbitrary $p > 2$, so

$$\lambda_i/\lambda_j \geq \delta_i/\delta_j \text{ for } 1 \leq i < j \leq p \tag{2}$$

as long as $\lambda_j > 0$. There is no explicit map between the eigenvalues known for $p > 2$. Dürre et al. [8] give a representation of δ_i as one-dimensional integral, which permits fast and accurate numerical evaluations for arbitrary p,

$$\delta_i = \frac{\lambda_i}{2} \int_0^\infty \frac{1}{(1 + \lambda_i x) \prod_{j=1}^p (1 + \lambda_j x)^{\frac{1}{2}}} dx, \ i = 1, \ldots, p. \tag{3}$$

We use this formula (implemented in R [17] in the package sscor [9]) to get an impression how the eigenvalues of $S(\mathbf{X})$ look like in comparison to those of V_0. We first look at of equidistantly spaced eigenvalues

$$\lambda_i = \frac{2i}{p(p+1)}, \ i = 1, \ldots, p,$$

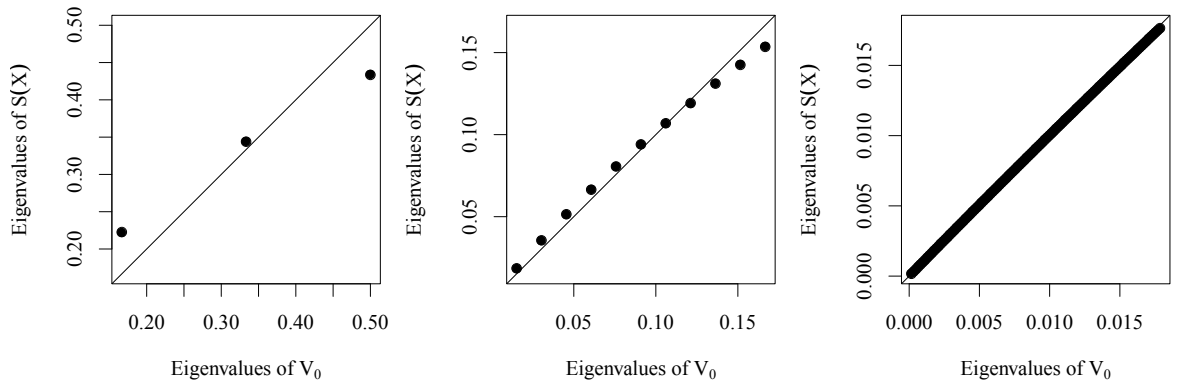


Figure 1: Eigenvalues of the SSCM wrt the corresponding eigenvalues of the shape matrix in the equidistant setting $p = 3$ (left), $p = 11$ (centre) and $p = 101$ (right).

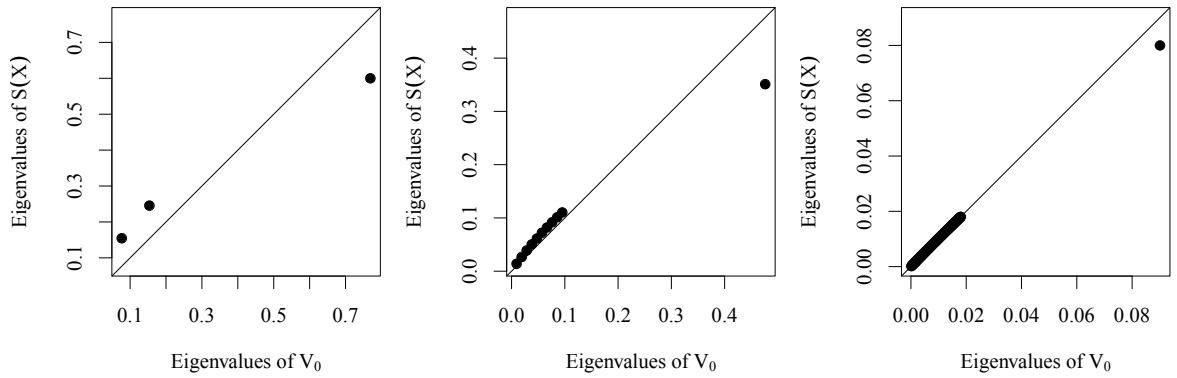


Figure 2: Eigenvalues of the SSCM wrt the corresponding eigenvalues of shape matrix in the setting of one large eigenvalue for $p = 3$ (left), $p = 11$ (centre) and $p = 101$ (right).

for different $p = 3,\ 11,\ 101$. The magnitude of the eigenvalues necessarily decreases as p increases, since $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \delta_i = 1$ per definition of V_0 and $S(\mathbf{X})$. As one can see in Figure 1, the eigenvalues of $S(\mathbf{X})$ and V_0 approach each other for increasing p. In fact the maximal absolute difference for $p = 101$ is roughly $2 \cdot 10^{-4}$. In the second scenario, we take $p - 1$ equidistantly spaced eigenvalues and one eigenvalue 5 times larger than the rest, i.e.,

$$\lambda_i = \begin{cases} \frac{i}{p((p+1)/2+5)-5} & i = 1, \ldots, p-1, \\ \frac{5(p-1)}{p((p+1)/2+5)-5} & i = p. \end{cases}$$

This models the case where the dependence is mainly driven by one principle component. As one can see in Figure 2, the distance between the two largest eigenvalues is smaller for $S(\mathbf{X})$ than for V_0. This is not surprising in light of (2). Thus in general, the eigenvalues of the SSCM are less separated than those of V_0, which is one reason why the use of the SSCM for robust principal component analysis has been questioned

(e.g. [1, 14]). However, the differences appear to be generally small in higher dimensions.

3 Estimation of the correlation matrix

Equation (1) can be used to derive an estimator for the correlation coefficient based on the empirical SSCM: the spatial sign correlation coefficient ρ_n ([6]). Under mild regularity assumptions this estimator is consistent under elliptical distributions and asymptotically normal with variance

$$\mathrm{ASV}(\rho_n) = (1-\rho^2)^2 + \frac{1}{2}(a+a^{-1})(1-\rho^2)^{3/2}, \tag{4}$$

where $a = \sqrt{v_{11}/v_{22}}$ is the ratio of the marginal scales and $\rho = v_{12}/\sqrt{v_{11}v_{22}}$ is the generalized correlation coefficient, which coincides with the usual moment correlation coefficient if second moments exists. Equation (4) indicates that the variance of ρ_n is minimal for $a = 1$, but can get arbitrarily large if a tends to infinity or 0.

Therefore a two-step procedure has been proposed, the *two-stage spatial sign correlation* $\rho_{\sigma,n}$, which first normalizes the data by a robust scale estimator, e.g., the median absolute deviation (mad), and then computes the spatial sign correlation of the transformed data. Under mild conditions (see [7]), this two-step procedure yields an asymptotic variance of

$$\mathrm{ASV}(\rho_{\sigma,n}) = (1-\rho^2)^2 + (1-\rho^2)^{3/2}, \tag{5}$$

which equals that of ρ_n for the favourable case of $a = 1$. Since (5) only depends on the parameter ρ, the two-stage spatial sign correlation coefficient is very suitable to construct robust and non-parametric confidence intervals for the correlation coefficient under ellipticity. It turns out that these intervals are quite accurate even for rather small sample sizes of $n = 10$ and in fact more accurate then those based on the sample moment correlation coefficient [7].

One can construct an estimator of the correlation matrix R by filling the off-diagonal positions of the matrix estimate with the bivariate spatial sign correlation coefficients of all pairs of variables. This was proposed in [6]. Equation (3) allows an alternative approach: First standardize the data by a robust scale estimator and compute the SSCM of the transformed data. Then apply a singular value decomposition

$$S_n(\mathbf{t}_n, \mathbf{X}_1, \ldots, \mathbf{X}_n) = \hat{U}\hat{\Delta}\hat{U}',$$

where $\hat{\Delta}$ contains the ordered eigenvalues $\hat{\delta}_1 \geq \ldots \geq \hat{\delta}_p$. One obtains estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ by inverting (3). Although theoretical results are yet to be established,

we found in our simulations that the following fix point algorithm

$$\begin{aligned}
\hat{\lambda}_i^{(0)} &= \delta_i, && i = 1, \ldots, p, \\
\tilde{\lambda}_i^{(k+1)} &= 2\hat{\delta}_i \left(\int_0^\infty \frac{1}{(1+\hat{\lambda}_i^{(k)} x) \prod_{j=1}^p (1+\hat{\lambda}_j^{(k)} x)^{\frac{1}{2}}} dx, \right)^{-1}, && i = 1, \ldots, p, \ k = 1, 2, \ldots \\
\hat{\lambda}_i^{(k+1)} &= \tilde{\lambda}_i^{(k+1)} \left(\sum_{j=1}^p \tilde{\lambda}_j^{(k+1)} \right)^{-1}, && i = 1, \ldots, p, \ k = 1, 2, \ldots
\end{aligned}$$

works reliably and converges fast. Let $\hat{\Lambda}$ denote the diagonal matrix containing $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, then $\hat{V} = \hat{U}\hat{\Lambda}\hat{U}'$ is a suitable estimator for for the shape of the standardized data and $\hat{R}$ with $\hat{r}_{ij} = \hat{v}_{ij}/\sqrt{\hat{v}_{ii}\hat{v}_{jj}}$ an estimator for the correlation matrix, which we call the *multivariate spatial sign correlation matrix*. Contrary to the pairwise approach, the multivariate spatial sign correlation matrix is positive semi-definite by construction.

Theoretical properties of the new estimator are not straightforward to establish. By a small simulation study we want to get an impression of its efficiency. We compare the variances of the moment correlation, the pairwise as well as the multivariate spatial sign correlation under several elliptical distributions: normal, Laplace and t distributions with 5 and 10 degrees of freedom. The latter three generate heavier tails than the normal distribution. The Laplace distribution is obtained by the elliptical generator $g(x) = c_p \exp(-\sqrt{|x|}/2)$, where c_p is the appropriate integration constant depending on p (e.g. [2], p. 209).

We take the identity matrix as shape matrix and compare the variances of an off-diagonal element of the matrix estimates for different dimensions $p = 2,\ 3,\ 5,\ 10,\ 50$ and sample sizes $n = 100,\ 1000$. We use the R packages mvtnorm [10] and MNM [16] for the data generation. The results based on 10000 runs are summarized in Table 1.

Except for the moment correlation at the t_5 distribution, the results for $n = 100$ and $n = 1000$ are very similar. Note that the variance of the moment correlation decreases at the Laplace distribution as the dimension p increases, but not so for the other distributions considered. The lower dimensional marginals of the Laplace distribution are, contrary to the normal and the t-distributions, not Laplace distributed (see [11]), and the kurtosis of the one-dimensional marginals of the Laplace distribution in fact decreases as p increases.

Equation (5) yields an asymptotic variance of 2 for the pairwise spatial sign correlation matrix elements regardless of the specific elliptical generator, which can also be observed in the simulation results. The moment correlation is twice as efficient under normality, but has a higher variance at heavy tailed distributions. For uncorrelated t_5 distributed random variables, the spatial sign correlation outperforms the moment correlation. Looking at the multivariate spatial sign correlation, we see a strong increase of efficiency for larger p. For $p = 50$ the variance is comparable to that of the moment correlation. Since the asymptotic variance of the SSCM does not depend on the elliptical generator, this is expected to also hold for the multivariate spatial sign correlation, and we find this confirmed by the simulations. The multivariate spatial sign correlation is more efficient than the moment correlation even under slightly heavier tails for moderately large p.

	n	100					1000				
	p	2	3	5	10	50	2	3	5	10	50
N	cor	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	sscor pairwise	1.9	1.9	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	1.9	1.6	1.4	1.2	1.0	2.0	1.7	1.4	1.2	1.0
t_{10}	cor	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.4	1.3
	sscor pairwise	2.0	1.9	1.9	2.0	1.9	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	2.0	1.7	1.3	1.2	1.0	2.0	1.7	1.4	1.2	1.0
t_5	cor	2.0	2.1	2.1	2.1	2.1	2.6	2.6	2.6	2.6	2.6
	sscor pairwise	2.0	2.0	1.9	2.0	1.9	2.1	2.0	2.0	2.0	2.0
	sscor multivariate	2.0	1.7	1.4	1.2	1.1	2.1	1.7	1.4	1.2	1.0
L	cor	1.6	1.5	1.3	1.2	1.1	1.6	1.5	1.3	1.2	1.1
	sscor pairwise	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	1.9	1.6	1.4	1.2	1.1	2.0	1.7	1.4	1.2	1.1

Table 1: Simulated variances (multiplied by $\sqrt{n}$) of one off-diagonal element of the correlation matrix estimate based on the moment correlation (cor), the pairwise spatial sign correlation (sscor pairwise) and the multivariate spatial sign correlation matrix (sscor multivariate) for spherical normal (N), t_5, t_{10}, and Laplace (L) distribution, several dimensions p and sample sizes $n = 100, 1000$.

An increase of efficiency for larger p is not uncommon for robust scatter estimators. It can be observed amongst others for M-estimators, the Tyler shape matrix, the MCD, and S-estimators (e.g. [4, 18]). All of these are affine equivariant estimators, requiring $n > p$. This is not necessary for the spatial sign correlation matrix. One may expect that the efficiency gain for large p is at the expense of robustness, in particular a larger maximum bias curve. Further research will be necessary to thoroughly explore the robustness properties and efficiency of the multivariate spatial sign correlation estimator.

References

[1] Bali J.L., Boente G., Tyler D.E., Wang J.L. (2011). Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics*. Vol. **39**, pp. 2852-2882.

[2] Bilodeau M., Brenner D. (1999). *Theory of Multivariate Statistics*. Springer, NY.

[3] Croux C., Dehon C., Yadine A. (2010). The k-step spatial sign covariance matrix. *Advances in data analysis and classification*. Vol. **4**, pp. 137-150.

[4] Croux C., Haesbroeck G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J. Multivariate Analysis*. Vol. **71**, pp. 161-190.

[5] Dürre A., Vogel D., Tyler D.E. (2014). The spatial sign covariance matrix with unknown location. *J. Multivariate Analysis*. Vol. **130**, pp. 107-117.

[6] Dürre A., Vogel D., Fried R. (2015). Spatial sign correlation. *J. Multivariate Analysis*. Vol. **135**, pp. 89-105.

[7] Dürre A., Vogel D. (2016). Asymptotics of the two-stage spatial sign correlation. *J. Multivariate Analysis*. Vol. **144**, pp. 54-67.

[8] Dürre A., Tyler D.E., Vogel D. (2016). On the eigenvalues of the spatial sign covariance matrix in more than two dimensions. *Statistics & Probability Letters*. Vol. **111**, pp. 80-85.

[9] Dürre A., Vogel D. (2016). *sscor: Robust Correlation Estimation and Testing Based on Spatial Signs*. R package version 0.2.

[10] Genz A et al. (2016), *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0.5.

[11] Kano Y. (1994). Consistency property of elliptic probability density functions. *J. Multivariate Analysis*. Vol. **51**, pp. 139-147.

[12] Kemperman J.H.B. (1987). The median of a finite measure on a Banach space. *Stat. Data Analysis Based on the L_1-Norm and Related Methods*. pp. 217-230.

[13] Locantore N. et al. (1999). Robust principal component analysis for functional data. *Test*. Vol. **8**(1), pp. 1-73.

[14] Magyar A.F., Tyler D.E. (2014). The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. *Biometrika*. Vol. **101**, pp. 673-688.

[15] Marden J.I. (1999). Some robust estimates of principal components. *Statistics & Probability Letters*. Vol. **43**, pp. 349-359.

[16] Nordhausen K., Oja H. (2011), Multivariate L_1 methods: the package MNM. *J. Statistical Software*. Vol. **43**, pp. 1-28.

[17] R Development Core Team (2016). R: A Language and Environment for Statistical Computing.

[18] Taskinen S. et al. (2006). Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. *Journal of Multivariate Analysis*. Vol. **97**, pp. 359-384.

[19] Vogel D., Köllmann C., Fried R. (2008). Partial correlation estimates based on signs. *Proc. 1st Workshop on Information Theoretic Methods in Science and Engineering*. Vol. **43**, pp. 1-6.

BINARY AND COUNT TIME SERIES ANALYSIS

K. Fokianos
University of Cyprus
Nicosia, CYPRUS
e-mail: fokianos@ucy.ac.cy

Abstract

We discuss some models for the statistical analysis of binary and count time series based on generalized linear models methodology. We outline the methods and tools needed for studying such models and we develop maximum likelihood estimation theory and diagnostics. The theory is extended to the general framework of time series following generalized linear models. Several real data examples complement the presentation.

IMPLEMENTATION OF THE POISSON-GAUSSIAN REGRESSION MODEL IN EMPIRICAL BAYES ESTIMATION OF SMALL PROBABILITIES

G. Jakimauskas[1], L. Sakalauskas[2]
Institute of Mathematics and Informatics, Vilnius University
Vilnius, LITHUANIA
e-mail: [1]gintautas.jakimauskas@mii.vu.lt, [2]leonidas.sakalauskas@mii.vu.lt

Abstract

The problem of implementation of the Poisson-Gaussian regression models in empirical Bayesian estimation of the small probabilities is considered. A bootstrap method using Monte-Carlo simulation is proposed. The method is applied to real-world USA cancer data combined with some possible regression variables, assuming they may have influence on the actual cancer data.

1 Introduction

Let us consider the problem of probability estimation of rare events in large populations (e.g., probabilities of some disease, homicides, suicides, etc.). The respective number of events depends on the population size and on the probability of a single event. Let us assume that probability of a single event depends only on population and these probabilities are the same in the same population. Moreover, assume that all events in all populations are independent. Under such assumptions number of events in each population will follow the Bernoulli distribution.

An event count refers to the number of times an event occurred in specific population. The benchmark model for count data is the Poisson distribution.

The Poisson distribution is the simplest distribution for modeling count data. However, it has one obvious limitation: its variance is equal to its mean. In case of real data we usually have so-called overdispersion: empirical variance is significantly bigger than empirical mean. In this case we can add some independent mixing distribution which increases variance of the combined distribution. By selecting parameters of the mixing distribution we can adjust the mean and the variance of the combined distribution to the empirical mean and the empirical variance of the real data. The simplest model adds gamma distribution to the Poisson distribution. The resulting distribution is known as negative binomial distribution or Poisson-gamma distribution. This distribution is more dispersed than the Poisson distribution. Obviously, negative binomial distribution can accommodate overdispersion but not underdispersion. There are many generalizations of the Poisson distributions (see, e.g., [2], [3], [6]).

Count data regression models have a widespread use (see, e.g., [2], [6]). The mean parameter of the Poisson-gamma model is usually parametrized using exponential link function of the regressors, in order to ensure that mean parameter is strictly greater than zero. As an alternative to the Poisson-gamma distribution, we will consider

Poisson-Gaussian distribution (see, e.g., [7], [8]). In this case the additional link function is not needed, and adding regression variables is very simple and clear. However, the calculations using Poisson-Gaussian model are much more complicated, and we need to use Hermite-Gauss numerical integration formulae (see, e.g. [1]).

2 Mathematical models

Let observed number of events $\{Y_j\} = Y_j,\ j = 1, \ldots, K$, be a sample of independent random variables $\{\mathbf{Y}_j\}$ with binomial distribution, respectively, with number of experiments $\{N_j\}$ and success probabilities $\{\lambda_j\}$. Clearly, $\{\mathbf{E}(\mathbf{Y}_j)\} = \{\lambda_j N_j\}$.

An assumption is often made (see, e.g., [5], [8]) that random variables $\{\mathbf{Y}_j\}$ have a Poisson distribution with parameters, respectively, $\{\lambda_j N_j\}$, i.e.

$$\mathbf{P}\{\mathbf{Y}_j = m\} = h(m,\ \lambda_j N_j),\ m = 0, 1, \ldots,\ j = 1, \ldots, K,$$

where

$$h(m,\ z) = e^{-z}\frac{z^m}{m!},\ m = 0, 1, \ldots,\ z > 0.$$

We will consider the mathematical model assuming that unknown probabilities $\{\lambda_j\}$ are independent identically distributed random variables with distribution function F from the certain class of distribution functions $\mathcal{F}$. Our problem is to get empirical Bayes estimates (see, e.g., [4]) of unknown probabilities $\{\hat{\lambda}_j\}$ from the observed number of events $\{Y_j\}$, assuming that $F \in \mathcal{F}$.

Poisson-gamma model. Given population sizes $\{N_j\}$, let random variables $\{\mathbf{Y}_j\}$ have a Poisson distribution with parameters, respectively, $\{\lambda_j N_j\}$, where $\{\lambda_j\}$ are independent identically distributed gamma random variables with shape parameter $\nu > 0$ and scale parameter $\alpha > 0$, i.e. the distribution function F has the distribution density

$$f(x) = f(x;\ \nu, \alpha) = \frac{\alpha \cdot (\alpha \cdot x)^{\nu-1}}{\Gamma(\nu)}\ e^{-\alpha x},\ 0 \le x < \infty\ .$$

Then $\mathbf{E}(\lambda_j) = \nu/\alpha$, and $\mathbf{E}(\lambda_j - \mathbf{E}(\lambda_j))^2 = \nu/\alpha^2,\ j = 1, \ldots, K$. Given observed number of events $\{Y_j\}$ and population sizes $\{N_j\}$, Bayes estimates for $\{\lambda_j\}$ are (see, e.g. [5])

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j)\ =\ \frac{Y_j + \nu}{N_j + \alpha},\ j = 1, \ldots, K. \tag{1}$$

Corresponding maximum likelihood function for parameters (ν, α) is

$$L(\nu, \alpha) = \sum_{j=1}^{K} \Big(\ln \frac{\Gamma(Y_j + \nu)}{\Gamma(\nu)} + \nu \ln(\alpha) - \\ -(Y_j + \nu)\ln(N_j + \alpha) + Y_j \ln N_j \Big). \tag{2}$$

Empirical Bayes estimates $\{\hat{\lambda}_j\}$ are obtained by maximizing (2) and replacing parameters (ν, α) in (1) with obtained parameters $(\hat{\nu},\ \hat{\alpha})$.

Poisson-Gaussian model. Alternatively, we will consider Bayes estimate $\{\tilde{\lambda}_j\}$, which is obtained under assumption that unknown probabilities are i.i.d. r.v.'s such that their logits $\alpha_j = \ln(\lambda_j/(1-\lambda_j))$, $j = 1,\ 2,\ \ldots\ , K$, are i.i.d. Gaussian r.v.'s with mean μ and variance σ^2 and corresponding distribution density φ_{μ,σ^2}.

Poisson-Gaussian model with regression variables. Additionally, let us introduce an auxiliary regression variables $\{Z_j\}^{(s)}$, $s = 1, \ldots, M$, assuming that $\mu(j) = \mu_0 + \mu_1 Z_j^{(1)} + \mu_2 Z_j^{(2)} + \ \cdots + \mu_M Z_j^{(M)}$, $j = 1,\ 2,\ \ldots\ , K$ (for our purposes we consider only simplified model without interactions of the regression variables). These variables are considered non-random, so all formulae for Poisson-Gaussian model hold also for Poisson-Gaussian model with regression variables.

In the case of both Poisson-Gaussian models conditional expectation of $\{\lambda_j\}$ has the following form:

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = D_j^{-1}(\mu(j), \sigma^2) \int\limits_{-\infty}^{\infty} \frac{1}{1+e^{-x}} h\left(Y_j, \frac{N_j}{1+e^{-x}}\right) \varphi_{\mu(j),\sigma^2}(x)\ dx,$$
$$j = 1, \ldots, K,$$

where

$$D_j(\mu(j), \sigma^2) = \int\limits_{-\infty}^{\infty} h\left(Y_j, \frac{N_j}{1+e^{-x}}\right) \varphi_{\mu(j),\sigma^2}(x)\ dx,$$
$$j = 1, \ldots, K.$$

3 Implementation of the Poisson-Gaussian regression model

To demonstrate the implementation of the Poisson-Gaussian regression model the main intention of data selection was to select freely available datasets, preferably of certain relatively large population, from the trusted databases. We have selected real data from the database of the USA National Cancer Institute, years 2011 and 2012, number of administrative territories (states) $K = 50$, 23 datasets in total. Also we have used population data by administrative territories from the United States Census Bureau.

As a basis for the regression variables we have used corresponding real data by administrative territories (states) from the Health Indicators Warehouse of the USA Center for Disease Control and Prevention. We have analysed three of possible regression variables, assuming they may have influence on the actual cancer data: (1) "Depression Medicare beneficiaries", (2) "High cholesterol Medicare beneficiaries", (3) "Toxic chemicals (pounds)".

For each dataset we have performed Monte-Carlo computer simulation of (typically) 100 independent realizations using both Poisson-gamma and Poisson-Gaussian models with corresponding parameters estimated from the real data (assuming that there are no regression). At the next stage we estimated (using maximum likelihood

method) Poisson-Gaussian model parameters without regression variables, and, alternatively, Poisson-Gaussian model parameters with selected regression variables. In the process, corresponding values of the maximum likelihood function were obtained. This procedure was applied to the real data, and to the 100 simulated realizations (either Poisson-gamma model realizations or Poisson-Gaussian model realizations).

The key point of this method is comparing difference of values of maximum likelihood function (for model with regression variables and for model without regression variables) for the real data with analogous differences for simulated realizations. Because simulated realizations have no influence of regression variables, they only have small random differences of values of maximum likelihood function, which main characteristics can be easily calculated. As a simple method, we can apply 3σ rule to detect presence of the regression variables.

As expected, the simulation results did not show significant difference of simulation using Poisson-gamma and Poisson-Gaussian models. Implementing the simple 3σ rule, we have found that for datasets 4, 9, 10, 16, 18 it is recommended to use regression variable "High cholesterol Medicare beneficiaries" for empirical Bayes estimation. For datasets 2, 9, 10, 16 it is recommended to use regression variable "Depression Medicare beneficiaries" for empirical Bayes estimation. As for regression variable "Toxic chemicals (pounds)" (combined with population size or with area size), we did not find influence of this variable.

References

[1] Abramovich M., Stegun I.A. (1968). *Handbook of Mathematical Functions*. Dover, New York.

[2] Cameron A.C., Trivedi P.K. (1998). *Regression Analysis of Count Data*. University Press, Cambridge.

[3] Cameron A.C., Trivedi P.K. (2005). *Microeconometrics. Methods and Applications*. University Press, Cambridge.

[4] Carlin B.P., Louis T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.

[5] Clayton D., Kaldor J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*. Vol. **43**, pp. 671-681.

[6] Hilbe J.M. (2011). *Negative Binomial Regression*. University Press, Cambridge.

[7] Sakalauskas L. (2010). On the Empirical Bayesian Approach for the Poisson-Gaussian Model. *Methodology and Computing in Applied Probability*. Vol. **12**, Issue **2**, pp. 247-259.

[8] Tsutakava R.K., Shoop G.L., Marienfield C.J. (1985). Empirical Bayes estimation of cancer mortality rates. *Statistics in medicine*. Vol. **4**, pp. 201-212.

STATISTICAL ANALYSIS OF DISCRETE SPATIO-TEMPORAL DATA BY CONDITIONAL AUTOREGRESSIVE MODELS

Yuriy Kharin[1], Maryna Zhurak[2]
Belarusian State University
Minsk, BELARUS
e-mail: [1]kharin@bsu.by, [2]mzhurak@gmail.com

Abstract

Poisson and Binomial conditional autoregressive model of spatio-temporal data is presented. Asymptotic properties of the maximum likelihood estimators of parameters for both conditional autoregressive models of spatio-temporal data are studied: asymptotic normality is proved and the asymptotic covariance matrix is found for the estimators, statistical tests on the values of true unknown parameters are constructed. Results of computer experiments on simulated and real data are given.

1 Introduction

Studying the probabilistic models of spatio-temporal data is a new topical scientific direction. Statistical analysis and modeling of spatio-temporal data is a challenging task [1]- [5].

Models based on spatio-temporal data become widely used for solving practical problems in meteorology, ecology, economics, medicine and other fields. In [5] spatio-temporal model is used to analyse daily precipitation for 71 meteorological stations over 60 years in Austria. Bayesian spatio-temporal model is applied to predict cancer cases in [1].

2 Conditional autoregressive models

Introduce the notation: $(\Omega, F, \mathbb{P})$ is the probability space; $S = \{1, 2, \ldots, n\}$ is the set of indexed spatial regions or space locations (let us call them sites), into which the analyzed spatial area is partitioned; n is number of sites; $t \in \mathbb{Z}$ is discrete time; T is the length of observation period; $x_{s,t}$ is a discrete random variable at time t at site s; $U(s) \subseteq S$ is a subset of neighbors of site s; $F_{<t} = \sigma\{x_{u,\tau} : u \in S, \tau < t\} \subset F$ is the σ-algebra generated by the indicated in braces random variables; $\{\varphi_k(t) : 1 \le k \le K\}$ is a given set of $K \in \mathbb{N}$ basic functions which determine a trend; $\mathcal{L}\{\cdot\}$, $\mathbb{E}\{\cdot\}$, $\mathbb{D}\{\cdot\}$ and $\mathbf{cov}\{\cdot\}$ are the symbols of probability distribution law of random variable, expectation, variance and covariance respectively; $\Pi(\lambda)$ is the Poisson probability distribution with the parameter $\lambda > 0$; $\mathrm{Bi}(N, p)$ is the binomial probability distribution with the parameters $N \in \mathbb{N}$ and $0 \le p \le 1$.

We construct the Poisson conditional autoregressive model for spatio-temporal data $\{x_{s,t}\}$, as in [3, 4]:

$$\mathcal{L}\{x_{s,t}|F_{<t}\} = \Pi(\lambda_{s,t}),$$
$$\ln \lambda_{s,t} = a_s x_{s,t-1} + \sum_{j\in U(s)} b_{s,j} x_{j,t} + \beta_s z_{s,t} + \sum_{k=1}^{K} \gamma_{s,k}\varphi_k(t),\ t \in \mathbb{Z},\ s \in S,$$
$$U(1) \equiv \emptyset, U(s) \subseteq \{1, 2, \ldots, s-1\},\ s = 2, \ldots, n,\ |U(s)| = K_s,\ K_1 \equiv 0,$$

where $x_{s,t} \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$; $z_{s,t}$ is an observed (known) level of exogenous factors at time t at site s; $a = (a_1, a_2, \ldots, a_n)' \in \mathbb{R}^n$, $b_s = \left(b_{s,j_1}, \ldots, b_{s,j_{K_s}}\right)' \in \mathbb{R}^{K_s}$, $j_k \in U(s)$, $k = 1, \ldots, K_s$, $s \in S$, $\beta = (\beta_1, \ldots, \beta_n)' \in \mathbb{R}^n$, $\gamma_s = (\gamma_{s,1}, \ldots, \gamma_{s,K})' \in \mathbb{R}^K$, $s \in S$, are the parameters of the model.

Similarly to Poisson model we construct the binomial conditional autoregressive model $\{x_{s,t}\}$: provided that prehistory $\{x_{s,\tau} : s \in S, \tau < t\}$ is fixed, random variables $x_{1,t}, \ldots, x_{n,t}$ are assumed to be conditionally independent and

$$\mathcal{L}\{x_{s,t}|F_{<t}\} = \mathrm{Bi}(N, p_{s,t}), \tag{1}$$

$$\ln \tfrac{p_{s,t}}{1-p_{s,t}} = \sum_{i=1}^{n} a_{s,i} x_{i,t-1} + \sum_{j=1}^{m} b_{s,j} z_{j,t},\ t \in \mathbb{Z},\ s \in S, \tag{2}$$

where $x_{s,t} \in A = \{0, \ldots, N\}$; $z_{j,t} \in \mathbb{R}$, $j = 1, \ldots, m$ is an observed (known) level of the j-th exogenous factor at time t which influences $x_{s,t}$; $a_s = (a_{s,1}, \ldots, a_{s,n})' \in \mathbb{R}^n$, $b_s = (b_{s,1}, \ldots, b_{s,m})' \in \mathbb{R}^m$, $s \in S$, $\theta_s = (a_s', b_s')' \in \mathbb{R}^{n+m}$, $\theta = (\theta_1', \ldots, \theta_n')' \in \mathbb{R}^{n(n+m)}$ is the composed vector of the parameters of the model; $p_{s,t}$ can be calculated as follows:

$$p_{s,t} = p_s(X_{t-1}, Z_t) ::= \exp\{\theta_s' Y_t\}/(1 + \exp\{\theta_s' Y_t\}),\ s \in S,\ t \in \mathbb{Z},$$

where $Z_t = (z_{1,t}, \ldots, z_{m,t})' \in \mathbb{R}^m$ is the column vector specifying exogenous factors at time t; $X_t = (x_{1,t}, x_{2,t}, \ldots, x_{n,t})' \in A^n$ is the column vector specifying the time slice of the process at $t \in \mathbb{Z}$; $Y_t = \left(X_{t-1}', Z_t'\right)' \in \mathbb{R}^{n+m}$, $t \in \mathbb{Z}$.

Probabilistic properties of the Poisson conditional autoregressive models are given in [3]. Here we will give probabilistic property for the binomial autoregressive model of spatio-temporal data.

Let $L = \left\{l_j = (l_{1,j}, \ldots, l_{n,j})' \in A^n : j = 1, \ldots, (N+1)^n\right\}$ be the ordered set of all admissible values of the vector X_t; $|L| = \nu = (N+1)^n$.

Theorem 1. *For the model (1), (2) the observed vector process X_t is the n-dimensional nonhomogeneous Markov chain with the finite state space L and the one-step transition probability matrix $Q(t) = (q_{I,J}(\theta, t))$, $I = (I_s), J = (J_s) \in L$:*

$$q_{I,J}(\theta, t) = \prod_{s=1}^{n} \frac{C_N^{J_s} \left(\exp\left\{a_s' I + b_s' Z_{t-1}\right\}\right)^{J_s}}{\left(1 + \exp\left\{a_s' I + b_s' Z_{t-1}\right\}\right)^{N}},\ t \in \mathbb{Z}.$$

Corollary 1. *Under conditions of Theorem 1, if vector of exogenous factors $Z_t = Z = (z_1, \ldots, z_m)' \in \mathbb{R}^m$ does not depend on t, then the one-step transition probability matrix*

does not depend on t, and Markov chain X_t is homogeneous:

$$Q = (q_{I,J}(\theta)) \in [0,1]^{\nu\times\nu}, \quad I, J \in L,$$

$$q_{I,J}(\theta) = \prod_{s=1}^{n} C_N^{J_s} \left(\exp\left\{a_s' I + b_s' Z\right\}\right)^{J_s} \left(1 + \exp\left\{a_s' I + b_s' Z\right\}\right)^{-N}.$$

Also X_t is ergodic and has the unique stationary distribution $\pi = (\pi_I) \in [0,1]^{\nu}$:

$$Q'\pi = \pi, \ \sum_{I\in A^n} \pi_I = 1.$$

Lemma 1. *For the model (1), (2) in case of any finite coefficients values $\{\theta_s\}$ and finite $\{z_{i,t}\}$ the covariance matrix $\mathbf{cov}\{X_t, X_t\}$ is positively defined and takes the form:*

$$\mathbf{cov}\{X_t, X_t\} = N\mathrm{diag}\left\{p_i(X_{t-1}, Z_t)(1 - p_i(X_{t-1}, Z_t))\right\} + D \in \mathbb{R}^{n\times n},$$

$$D = (d_{ij}) \in \mathbb{R}^{n\times n}, d_{ij} = N^2\mathbf{cov}\left\{(1 + \exp(-\theta_i' Y_t))^{-1}, (1 + \exp(-\theta_j' Y_t))^{-1}\right\}.$$

3 Statistical estimation of parameters

Theorem 2. *The loglikelihood function for the model (1), (2) under the observed spatio-temporal data $\{X_t : t = 1, 2, \ldots, T\}$ takes the additive form:*

$$l(\theta) = \sum_{s=1}^{n} l_s\left(\theta_s\right), l_s\left(\theta_s\right) = \sum_{t=1}^{T} \left(x_{s,t}\theta_s' Y_t - N\ln\left(1 + \exp\left\{\theta_s' Y_t\right\}\right) + \ln C_N^{x_{s,t}}\right). \quad (3)$$

To find the maximum likelihood estimators (MLE) $\{\hat{\theta}_s\}$ of the parameters we need to maximize the loglikelihood function (3):

$$l(\theta) \to \max_{\theta\in\mathbb{R}^{n(n+m)}}. \quad (4)$$

Theorem 3. *In case of the model (1), (2), if $m = 1$, $z_{1t} = z \neq 0$ does not depend on t and Markov chain $X_t \in L$ is stationary, then for any finite coefficients values $\{\theta_s\}$ and finite $z \in \mathbb{R}$ the Fisher information matrix is nonsingular block-diagonal matrix (with $Y_t = (X_{t-1}', z)'$):*

$$G = N\mathrm{diag}\left\{\mathbb{E}\left\{Y_t Y_t' p_i(X_{t-1}, z)(1 - p_i(X_{t-1}, z))\right\}\right\}, \ i = 1, \ldots, n. \quad (5)$$

Theorem 4. *Under Theorem 3 conditions, if $T \to +\infty$ the constructed by (4) maximum likelihood estimators $\{\hat{\theta}_s\}$ are consistent and asymptotically normally distributed:*

$$\mathcal{L}\left\{\sqrt{T}(\hat{\theta} - \theta^0)\right\} \to N_{n(n+1)}\left(0, G^{-1}\right),$$

where G is determined by (5).

Theorems 2-4 are used to construct statistical tests for testing of hypotheses on the values of true unknown parameters $\{\theta_s^0\}$:

$$H_0 : \theta^0 = \theta^*;$$
$$H_1 = \overline{H_0} : \theta^0 \neq \theta^*.$$

where $\theta^* \in \mathbb{R}^{n(n+m)}$ is some fixed (hypothetical) value of parameters. Let us consider the statistic:

$$g_T = g\left(X_1, \ldots, X_T\right) ::= T(\widehat{\theta} - \theta^*)' G (\widehat{\theta} - \theta^*) \geq 0,$$

where $\widehat{\theta}$ is estimator of model's parameters, G is determined by (5).

Theorem 5. *Under Theorem 3 conditions, if hypothesis H_0 is true and $T \to \infty$ then statistic g is asymptotically chi-square distributed with $n(n+1)$ degrees of freedom:*

$$\mathcal{L}_{H_0}\{g_T\} \to \chi^2_{n(n+1)}.$$

The decision rule based on statistic g and Theorem 5 is

$$H_0, g_T < \Delta;$$
$$H_1, g_T \geq \Delta,$$

where $\Delta = F^{-1}_{\chi^2_{n(n+1)}}(\alpha)$, α is asymptotic significance level.

Theorem 6. *Under Theorem 3 conditions for the sequence of contigual hypotheses $H_{1T} = \left\{\theta^0 = \theta^* + T^{-1/2}a\right\}$, $T \to \infty$, the test statistic g_T is asymptotically noncentral chi-square distributed with $n(n+1)$ degrees of freedom and noncentrality parameter $\Delta^2 = a'Ga$:*

$$\mathcal{L}_{H_1}\{g_T\} \to \chi^2_{\Delta^2, n(n+1)},$$

and the power of the test satisfies the asymptotics:

$$w_T \to w^* = 1 - F_{\chi^2_{\Delta^2, n(n+1)}}\left(F^{-1}_{\chi^2_{n(n+1)}}(\alpha)\right).$$

4 Results of computer experiments

We consider the model (1), (2) with the following values of parameters: $m = 1$, $z = 2$, $N = 4$, $A = \{0, 1, 2, 3, 4\}$, $n = 3$, $S = \{1, 2, 3\}$, $\theta_1 = (-0.2, 0.18, -0.15, 0.2)'$, $\theta_2 = (-0.18, 0.24, -0.05, -0.1)'$, $\theta_3 = (0.13, -0.13, -0.29, 0.3)'$.

Figure 1 plots dependence of experimental and theoretical mean square error of the parameter estimators on the observation time T (T varies from 20 to 300). Experimental mean square error was estimated by $M = 1000$ Monte-Carlo replications:

$$\hat{\delta} = \hat{E}\left\{\left\|\hat{\theta} - \theta\right\|^2\right\} = \frac{1}{M}\sum_{k=1}^{M} ||\hat{\theta}^{(k)} - \theta||^2,$$

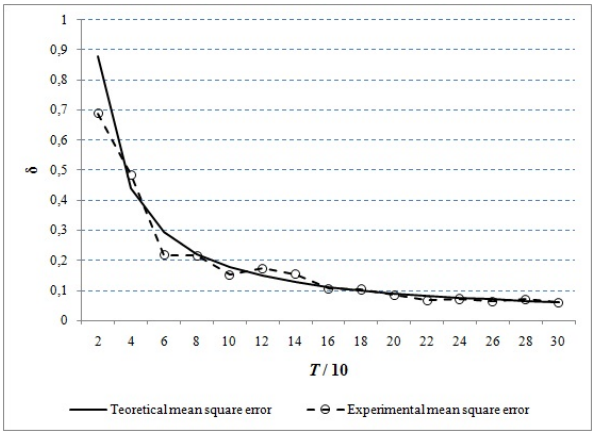


Figure 1: Mean square risk plotted against T

where $\hat{\theta}^{(k)}$ is the estimate for the kth realization. Theoretical mean square error was calculated using Theorems 3-4:

$$\delta = \frac{1}{T}\mathrm{tr}(G^{-1}),$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix.

Figure 1 illustrates the property of consistency of the MLE $\hat{\theta}$.

In [3] experiments were carried out on real data that describes the incidence rate of children leukemia in 3 sites ($n = 3$) of Republic of Belarus for 25 years ($T = 25$). Results of computer experiments on simulated and real data illustrate the theoretical results.

This research was supported by the Project financed by the Software Development Company InDataLabs.

References

[1] Kang S.Y., McGreed J., Baade P., Mengersen K. (2015). Case Study for Modelling Cancer Incidence Using Bayesian Spatio-Temporal Models. *Australian and New Zealand J. Stat.* Vol. **3**, pp. 325-345.

[2] Kharin Yu.S. (2013). *Robustness in Statistical Forecasting*. Springer, NY.

[3] Kharin Yu.S., Zhurak M.K. (2015). Statistical Analysis of Spatio-Temporal Data Based on Poisson Conditional Autoregressive Model. *INFORMATICA*. Vol. **26**(1), pp. 67-87.

[4] Mariella L., Tarantino M. (2010). Spatial temporal conditional Auto-Regressive Model: A New Autoregressive Matrix. *Austrian J. Stat.* Vol. **39**, pp. 223-244.

[5] Umlauf N., Mayr G., Messner J., Zeileis A. (2012). Why Does It Rain on Me? A Spatio-Temporal Analysis of Precipitation in Austria. *Austrian J. Stat.* Vol. **41**, pp. 81-92.

ICA AND STOCHASTIC VOLATILITY MODELS

M. Matilainen[1], J. Miettinen[2], K. Nordhausen[3], H. Oja[4], S. Taskinen[5]
[1,3,4]*Department of Mathematics and Statistics, University of Turku*
[2,5]*Department of Mathematics and Statistics, University of Jyvaskyla*
[1,3,4]*Turku and* [2,5]*Jyvaskyla, FINLAND*
e-mail: [1]markus.matilainen@utu.fi, [2]jari.p.miettinen@jyu.fi,
[3]klaus.nordhausen@utu.fi, [4]hannu.oja@utu.fi, [5]sara.l.taskinen@jyu.fi

Abstract

We consider multivariate time series where each component series is an unknown linear combination of latent mutually independent stationary time series. Multivariate financial time series have often periods of low volatility followed by periods of high volatility. This kind of time series have typically non-Gaussian stationary distributions, and therefore standard independent component analysis (ICA) tools such as fastICA can be used to extract independent component series even though they do not utilize any information on temporal dependence. In this paper we review some ICA methods used in the context of stochastic volatility models. We also suggest their modifications which use nonlinear autocorrelations to extract independent components. Different estimates are then compared in a simulation study.
Keywords: blind source separation, GARCH model, nonlinear autocorrelation, multivariate time series

1 Introduction

In this paper we assume that the observed p-variate time series $\boldsymbol{x} = (\boldsymbol{x}_t)_{t\in\mathbb{Z}}$ follows the basic *independent component (IC) model*

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \boldsymbol{\Omega}\boldsymbol{z}_t, \quad t \in \mathbb{Z},$$

where $\boldsymbol{\mu}$ is a p-variate location vector, $\boldsymbol{\Omega}$ is a full-rank $p \times p$ mixing matrix and $\boldsymbol{z} = (\boldsymbol{z}_t)_{t\in\mathbb{Z}}$ is an unobservable p-variate stationary time series such that

$$(i) \quad \mathrm{E}(\boldsymbol{z}_t) = \boldsymbol{0}, \quad (ii) \quad \mathrm{COV}(\boldsymbol{z}_t) = \boldsymbol{I}_p \text{ and}$$
$$(iii) \text{ the component series of } \boldsymbol{z} \text{ are independent.}$$

Then $\boldsymbol{x}$ is also stationary with $\mathrm{E}(\boldsymbol{x}_t) = \boldsymbol{\mu}$ and $\mathrm{COV}(\boldsymbol{x}_t) = \boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}'$. In *independent component analysis (ICA)* the goal is to find, using the observed time series $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, an estimate of an unmixing matrix $\boldsymbol{W}$ such that $\boldsymbol{W}\boldsymbol{x} = (\boldsymbol{W}\boldsymbol{x}_t)_{t\in\mathbb{Z}}$ has independent component series.

The IC model has recently achieved a lot of attention in financial time series analysis as complicated p-variate time series models can then be replaced by p simple univariate (e.g. ARMA or GARCH) models in parameter estimation and prediction problems. The model also serves as a dimension reduction tool as often only few component series

in $\boldsymbol{z}$ are relevant and the rest of the components just present noise. For some recent contributions, see [3, 6, 7, 11, 17].

In the literature standard ICA methods, such as fastICA, are often used to estimate an unmixing matrix $\boldsymbol{W}$ in a time series context although such methods only use the marginal distribution of $\boldsymbol{x}_t$ and make no use of the information on temporal dependence. On the other hand, there exist second order source separation methods, like SOBI [1], which are particularly popular for analyzing biomedical data. Such methods use autocovariances and cross-autocovariances for the estimation. They are capable of separating time series with nonzero linear autocorrelations, but they do not utilize nonlinear autocorrelations.

Volatility clustering is a common feature in economic and financial time series, i.e. there are periods of lower and higher volatility. As the transitions between such periods do not typically have any clear pattern, they are treated as random occurrences. There are a vast amount of different models that have been invented for such situations. Among stochastic volatility models, the GARCH process [2] has been the most popular one. Another popular model is the SV (Stochastic Volatility) model [20]. In our simulations we consider these two models. For further information on stochastic volatility and a recent overview of stochastic volatility models, see for example [13].

In this paper we review various independent component estimators that use nonlinear autocorrelations, and compare their performance to that of fastICA in a simulation study where the independent time series components come from GARCH and SV models. The paper has the following structure. First, in Section 2 we define the univariate stochastic volatility models. In Section 3 we discuss the ICA methods which are considered in this paper. Section 4 consists of the simulation study.

2 Stochastic volatility models for univariate series

Among stochastic volatility models, the GARCH (Generalized Autoregressive Conditional Heteroscedasticity) process [2] has been the most popular one. A univariate GARCH(p, q) process is given by

$$x_t = \sigma_t \epsilon_t,$$

where ϵ_t is an independent white noise process and σ_t^2 a deterministic conditional variance process

$$\sigma_t^2 = Var(x_t|\mathcal{F}_{t-1}) = \omega + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

with $\omega > 0$ and $\alpha_i, \beta_j \geq 0 \; \forall i, j$. For (second order) stationarity, $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$.

Another popular model is the SV (Stochastic Volatility) model [20], defined as

$$\begin{aligned} x_t &= e^{h_t/2} \epsilon_t, \\ h_t &= \mu + \phi(h_{t-1} - \mu) + \sigma \eta_t, \end{aligned}$$

where ϵ_t and η_t are two independent white noise innovation processes. Parameter μ is the level, ϕ is the persistence and $\sigma\eta_t$ is the volatility of log-variance. The process h_t is called the volatility process and it is strongly stationary with $N(0,1)$ innovations and initial state $h_0 \sim N(\mu, \sigma^2/(1-\phi^2))$. For stationarity, we require $|\phi| < 1$ and $\mu \in \mathbb{R}$.

3 Source separation for multivariate time series

Under our model assumption, the standardized multivariate series of $\boldsymbol{x}_t$ is given by $\boldsymbol{x}_t^{st} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}_t - \boldsymbol{\mu})$. One of the key results in ICA states that there exists an orthogonal matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)'$ such that $\boldsymbol{z}_t = \boldsymbol{U}\boldsymbol{x}_t^{st}$ (up to signs and order of the components) [16]. Here $\boldsymbol{z}_t$ denotes the vector of independent series. The final unmixing matrix functional is then given by $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$. The estimate of $\boldsymbol{W}$ is then obtained by replacing $\boldsymbol{\Sigma}$ and $\boldsymbol{U}$ by their sample counterparts. For finding $\boldsymbol{U}$, we next list the criterion functions in different approaches.

In the symmetric *fastICA* [9] approach and symmetric squared fastICA [15], $\boldsymbol{U}$ maximizes

$$\sum_{i=1}^{p} |\mathrm{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\right]| \quad \text{and} \quad \sum_{i=1}^{p} \left(\mathrm{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\right]\right)^2,$$

with a choice of a twice continuously differentiable, nonlinear and nonquadratic function G such that $E[G(y)] = 0$ if $y \sim N(0,1)$. Two common options are $G(z) = z^4 - 3$ and $G(z) = \log(\cosh(z)) - E[G(y)]$, where $y \sim N(0,1)$. Notice that both utilize only the stationary (marginal) distribution of $\boldsymbol{x}_t$.

The estimators presented below make use of the joint distributions of $(\boldsymbol{x}_t, \boldsymbol{x}_{t+k})$, $k = 1, 2, \ldots$. The classical *SOBI* uses only second moments and it was originally defined as a method which jointly diagonalizes several autocovariance matrices. However, SOBI can be reformulated as the maximizer of

$$\sum_{i=1}^{p}\sum_{k=1}^{K} \left(\mathrm{E}\left[(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})\right]\right)^2.$$

The solution is unique if, for all pairs $i \neq j$ there exists a k, $1 \leq k \leq K$, such that $\mathrm{E}(z_{t,i}z_{t+k,i}) \neq \mathrm{E}(z_{t,j}z_{t+k,j})$. SOBI fails to separate GARCH and SV time series as all lagged autocovariances are then zero.

The *gFOBI* procedure proposed in [12] maximizes a sum of fourth moments

$$\sum_{i=1}^{p}\sum_{k=1}^{K} \left(\mathrm{E}\left[(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})||\boldsymbol{x}_t^{st}||^2\right]\right)^2.$$

For $K = 0$, the regular ICA method FOBI [4] is obtained.

The *gJADE* procedure [12], in turn, uses a much richer sum of fourth cumulants and maximizes

$$\sum_{i=1}^{p}\sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{k=1}^{K} \left(\kappa(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st}, \boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st}, \boldsymbol{x}_{t,r}^{st}, \boldsymbol{x}_{t,s}^{st})\right)^2,$$

where

$$\kappa(z_1, z_2, z_3, z_4) = \mathrm{E}(z_1 z_2 z_3 z_4) - \mathrm{E}(z_1 z_2)\mathrm{E}(z_3 z_4) - \mathrm{E}(z_1 z_3)\mathrm{E}(z_2 z_4) - \mathrm{E}(z_1 z_4)\mathrm{E}(z_2 z_3).$$

Again, for $K = 0$, the regular ICA method JADE [5] is obtained. Both, gFOBI and gJADE, were created having stochastic volatility models in mind.

FastICA does not use any knowledge of temporal dependence, but there exist some fixed-point algorithms aimed for time series context. The *FixNA* (Fixed-point algorithm for maximizing the nonlinear autocorrelation) method was introduced in [19], and its criterion function to be maximized is

$$D_1(\boldsymbol{U}) = \sum_{i=1}^{p} \sum_{k=1}^{K} E\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st}) G(\boldsymbol{u}_i' \boldsymbol{x}_{t+k}^{st})\right],$$

where G is a twice continuously differentiable function. The G-functions suggested in [19] are $G(z) = \log(\cosh(z))$ and $G(z) = z^2$.

A similar function to be maximized is of the form

$$D_2(\boldsymbol{U}) = \sum_{i=1}^{p} \sum_{k=1}^{K} \left| E\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st}) G(\boldsymbol{u}_i' \boldsymbol{x}_{t+k}^{st})\right] - E\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st})\right]^2 \right|,$$

and we will denote it as *FixNA2*. It was first proposed in [8], however only with $G(z) = z^2$, and $K = 1$. We further similarly suggest a natural extension of SOBI with the criterion function

$$D_3(\boldsymbol{U}) = \sum_{i=1}^{p} \sum_{k=1}^{K} \left(E\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st}) G(\boldsymbol{u}_i' \boldsymbol{x}_{t+k}^{st})\right] - E\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st})\right]^2 \right)^2.$$

As a variant of SOBI, we call this estimator *vSOBI*.

To obtain the estimating equations for matrix $\boldsymbol{U}$, the Lagrangian multiplier technique can be used as in [14]. The Lagrangian function to be optimized is

$$L(\boldsymbol{U}, \boldsymbol{\Lambda}) = D_r(\boldsymbol{U}) - \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \lambda_{ij} \boldsymbol{u}_i' \boldsymbol{u}_j - \sum_{i=1}^{p} \lambda_{ii} (\boldsymbol{u}_i' \boldsymbol{u}_i - 1), \text{ for } r = 1, 2, 3,$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a symmetric matrix that contains $p(p+1)/2$ Lagrangian multipliers. Write next

$$\boldsymbol{T}_{r,i} = \boldsymbol{T}_{r,i}(\boldsymbol{U}) = \frac{\partial}{\partial \boldsymbol{u}_i} D_r(\boldsymbol{U}), \ i = 1, \ldots, p, \ r = 1, 2, 3,$$

and $\boldsymbol{T}_r = \boldsymbol{T}_r(\boldsymbol{U}) = (\boldsymbol{T}_{r,1}, \ldots, \boldsymbol{T}_{r,p})'$. Solving the optimizing problem then gives the estimating equations for $\boldsymbol{U}$, namely,

$$\boldsymbol{U}\boldsymbol{T}_r' = \boldsymbol{T}_r \boldsymbol{U}' \ \text{ and } \ \boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}_p,$$

or, equivalently,

$$\boldsymbol{U} = (\boldsymbol{T}_r \boldsymbol{T}_r')^{-1/2} \boldsymbol{T}_r.$$

For some tolerance limit ε and initial value $\boldsymbol{U}_0$, this leads to Algorithm 1.

Data: Standardized time series $\boldsymbol{x}_t^{st} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}_t - \boldsymbol{\mu})$
Result: $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$
$\boldsymbol{U}_{old} = \boldsymbol{U}_0$;
$\Delta = \infty$;
while $\Delta > \varepsilon$ **do**
  $\boldsymbol{T}_r = \boldsymbol{T}_r(\boldsymbol{U}_{old})$;
  $\boldsymbol{U}_{new} = (\boldsymbol{T}_r\boldsymbol{T}_r')^{-1/2}\boldsymbol{T}_r$;
  $\Delta = ||\boldsymbol{U}_{new} - \boldsymbol{U}_{old}||$;
  $\boldsymbol{U}_{old} = \boldsymbol{U}_{new}$;
end
$\boldsymbol{U} = \boldsymbol{U}_{new}$;

Algorithm 1: Algorithm for maximizing the criterion function $D_r,\ r = 1, 2, 3$.

4 Simulation study

The following simulations are conducted using R 3.2.2 [18] with the packages fGarch, fICA, JADE and tsBSS. In the simulation study we compare due to space limitations only the following methods:

- FixNA, FixNA2 and vSOBI with $G(z) = z^2$ and lags $1, \ldots, 12$
- symmetric fastICA and symmetric squared fastICA with $G(z) = z^4 - 3$
- gFOBI, gJADE with lags $0, 1, \ldots, 12$ and SOBI with lags $1, \ldots, 12$

The comparison is based on the Minimum Distance Index [10], which is defined as

$$\hat{D} = \hat{D}(\hat{\boldsymbol{W}}) = \frac{1}{\sqrt{p-1}} \inf_{\boldsymbol{C}\in\boldsymbol{\mathcal{C}}} ||\boldsymbol{C}\hat{\boldsymbol{W}}\boldsymbol{\Omega} - \boldsymbol{I}_p||,$$

where $\boldsymbol{\mathcal{C}}$ is the set of all matrices with exactly one non-zero element in each row and column, and $||\cdot||$ is the Frobenius (matrix) norm. The index has the range $0 \leq \hat{D} \leq 1$, where zero indicates perfect separation.

For time series of lengths $T = 100, 200, \ldots, 25600$ we report the averages $T(p-1)\hat{D}^2$ based on 2000 repetitions. Such an average represents a global measure of variation of an unmixing matrix, see [10] for details. As all the methods are affine equivariant, we choose wlog $\boldsymbol{\Omega} = \boldsymbol{I}_p$ and consider the following two 4-variate settings:

- **GARCH setting:** The sources are four GARCH$(1, 1)$ processes with normal innovations. The parameters (α_1, β_1) are chosen so that the first eight moments are finite, and are: (i) (0.05, 0.9), (ii) (0.1, 0.7), (iii) (0.1, 0.8) and (iv) (0.2, 0.5).
- **SV setting:** In the second setup the four sources are SV processes with normal innovations and (μ, ϕ, σ)-parameter vectors $(-10, 0.8, 0.1)$, $(-10, 0.9, 0.2)$, $(-10, 0.9, 0.3)$ and $(-10, 0.95, 0.4)$. Again, all the first eight moments exist.

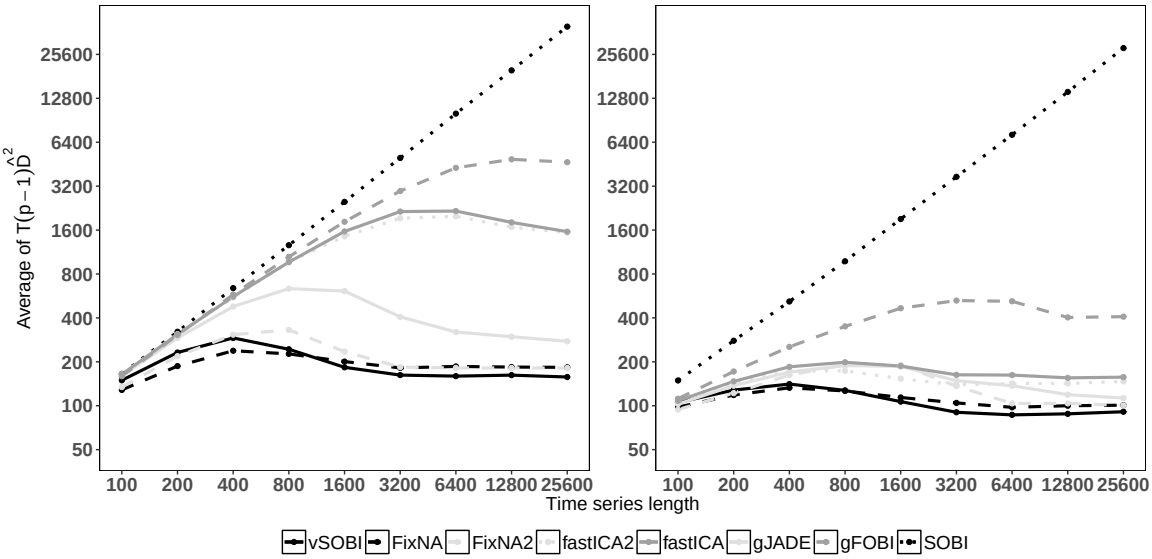


Figure 1: Comparison of performance of algorithms in the GARCH setting (left panel) and SV setting (right panel).

Figure 1 summarizes the results for both settings. As expected, SOBI does not work here. The proposed vSOBI estimator works very well in both cases and outperforms all the other estimators. Interestingly, both fastICA algorithms perform well in the SV example but not in the GARCH example. FastICA2 algorithm produces slightly better results than the fastICA algorithm. While gJADE works quite well in both cases, gFOBI has much poorer performance. FixNA and FixNA2 algorithms are among the best methods.

Convergence of FixNA2 algorithm and both fastICA algorithms is low in short time series (see Figure 2), but gets much better when the time series length increases. Convergence percentage of vSOBI is also good, and in time series of length 800 onwards very close to 100%. SOBI, gFOBI and gJADE have very few convergence issues, if any.

5 Discussion

In this paper we surveyed different blind source separation methods suitable for multivariate time series with stochastic volatility features. Such methods were earlier quite scattered in the literature. We also suggested some small modification yielding the family of vSOBI estimators which showed in our simulations the best performance. We have shown here the simulation results of vSOBI, both FixNA and both FastICA algorithms only based on G functions of the form $G(z) = z^c$. However, in an extended version of this paper we plan to have a larger simulation study, including for example also $\log(\cosh(z))$ as a nonlinearity.

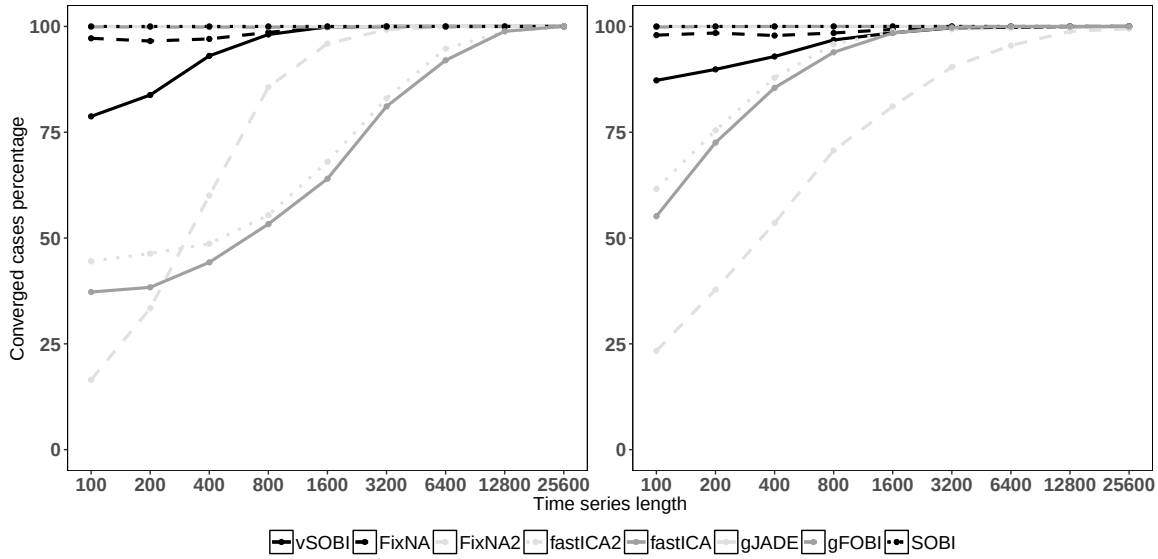


Figure 2: Comparison of convergence percentages of algorithms in the GARCH setting (left panel) and SV setting (right panel).

Acknowledgements

This work was supported by the Academy of Finland (grants 251965, 256291 and 268703).

References

[1] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE T. Signal Proc.*, 45:434–444, 1997.

[2] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31(3):307–327, 1986.

[3] S. A. Broda and M. S. Paolella. CHICAGO: A fast and accurate method for portfolio risk calculation. *J. Financ. Economet.*, 7(4):412–436, 2009.

[4] J.-F. Cardoso. Source separation using higher order moments. In *Int. Conf. Acoust. Spee.*, pages 2109–2112, 1989.

[5] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE-Proc.-F*, volume 140, pages 362–370, 1993.

[6] Y. Chen, W. Härdle, and V. Spokoiny. Portfolio value at risk based on independent component analysis. *J. Comput. Appl. Math.*, 205:594–607, 2007.

[7] A. García-Ferrer, E. González-Prieto, and D. Peña. A conditionally heteroskedastic independent factor model with an application to financial stock returns. *Int. J. Forecasting*, 28(1):70 – 93, 2012.

[8] A. Hyvärinen. Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE T. Neural Networ.*, 12(6):1471–1474, 2001.

[9] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analyis. *Neural Comput.*, 9:1483–1492, 1997.

[10] P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila. A new performance index for ICA: Properties computation and asymptotic analysis. In V. Vigneron et al., editor, *LVA/ICA 2010. LNCS*, volume 6365, pages 229–236, Heidelberg, 2010. Springer.

[11] C.-J. Lu, J.-Y. Wu, and T.-S. Lee. Application of independent component analysis preprocessing and support vector regression in time series prediction. In *International Joint Conference on Computational Sciences and Optimization*, volume 1, pages 468–471, 2009.

[12] M. Matilainen, K. Nordhausen, and H. Oja. New independent component analysis tools for time series. *Stat. Probabil. Lett.*, 105:80–87, 2015.

[13] D.S. Matteson and D. Ruppert. Time-series models of dynamic volatility and correlation. *IEEE Signal Proc. Mag.*, 28(5):72–82, 2011.

[14] J. Miettinen, K. Illner, K. Nordhausen, H. Oja, S. Taskinen, and F. Theis. Separation of uncorrelated stationary time series using autocovariance matrices. *J. Time Ser. Anal.*, 37(3): 337–354, 2016.

[15] J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, and J. Virta. The squared symmetric FastICA estimator, 2015. http://arxiv.org/abs/1512.05534.

[16] J. Miettinen, S. Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *Stat. Sci.*, 30:372–390, 2015.

[17] E. Oja, K. Kiviluoto, and S. Malaroiu. Independent component analysis for financial time series. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pages 111–116, 2000.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. R version 3.2.2.

[19] Z. Shi, Z. Jiang, and F. Zhou. Blind source separation with nonlinear autocorrelation and non-gaussianity. *J. Comput. Appl. Math.*, 223(1):908–915, 2009.

[20] S. J. Taylor. Financial returns modelled by the product of two stochastic processes – a study of daily sugar prices 1961–79. In O. D. Anderson, editor, *Time Series Analysis: Theory and Practice 1*, pages 203–216. Springer, North-Holland, Amsterdam, 1982.

MIXED POWER VARIATIONS WITH STATISTICAL APPLICATIONS

Yu.S. Mishura
Taras Shevchenko National University of Kyiv
Kyiv, UKRAINE
e-mail: myus@univ.kiev.ua

Abstract

We obtain results on both weak and almost sure asymptotic behavior of power variations of a linear combination of independent Wiener process and fractional Brownian motion. These results are used to construct strongly consistent parameter estimators in mixed models.

1 Introduction

These results are common with G. Shevchenko and M. Dozzi.

A fractional Brownian motion (fBm) is frequently used to model short- and long-range dependence. By definition, an fBm with Hurst parameter $H \in (0,1)$ is a centered Gaussian process $\{B_t^H, t \geq 0\}$ with the covariance function

$$\mathsf{E}[B_t^H B_s^H] = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right).$$

For $H > 1/2$, an fBm has a property of long-range dependence; for $H < 1/2$, it is short-range dependent and, in fact, is counterpersistent, i.e. its increments are negatively correlated. For $H = 1/2$, an fBm is a standard Wiener process.

Two important properties of an fBm are the stationarity of increments and self-similarity. However, these properties restrict applications of an fBm. So, let us consider some generalizations. A simplest approach is to consider a linear combination

$$X_t = \sum_{k=1}^{N} a_k B_t^{H_k}, t \geq 0, \tag{1}$$

of independent fBms B^{H_k} with different Hurst parameters $H_1 < H_2 < \cdots < H_N$.

We consider a particular version of the process (1) with $N = 2$ and one of the Hurst parameters equal to $1/2$. In other words, we consider a process

$$M_t^H = aB_t^H + bW_t, t \geq 0 \tag{2}$$

where a and b are some non-zero coefficients. Such process is frequently called a mixed fractional Brownian motion. Its applications were considered in many papers, see [2, 7, 10, 11].

There are only few papers concerned with parameter estimation in the mixed model, but they address questions different from the one we are interested in. In particular, [6, 9] address the estimation of drift parameter in a model with mixed fractional

Brownian motion. In [1], the authors construct several estimators based on discrete variation, so their research is quite close to ours, but they also work in the low-frequency setting, which is essentially different from the high-frequency setting we consider. In both settings, the first order difference of the observed series is a stationary sequence. However, in the low-frequency setting the covariance does not depend on the number of observations, while in the high-frequency one, the covariance structure is very different. As it was mentioned above, for $H > 1/2$, in a small scale the mixed fractional Brownian motion behaves like Wiener process. Thus, the increments of Wiener process become more and more dominating as the partition becomes finer, which makes estimation of the Hurst parameter much harder in the case where $H > 1/2$.

As it was already mentioned, our main aim is the estimation of the parameters of the process (2) based on its single observation on a uniform partition of a fixed interval. To this end, we use power variations of this process.

In the future we plan to consider more advanced techniques as those developed in [1, 4, 5, 8] to construct more efficient estimators.

2 Asymptotic behavior of mixed power variations

Let $W = \{W_t, t \geq 0\}$ be a standard Wiener process and $B^H = \{B^H_t, t \geq 0\}$ be an independent of W fBm with Hurst parameter $H \in (0, 1)$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$.

For a function $X : [0, 1] \to \mathbb{R}$ and integers $n \geq 1$, $i = 0, 1, \ldots, n-1$ we denote $\Delta^n_i X = X_{(i+1)/n} - X_{i/n}$. In this section we will study the asymptotic behavior as $n \to \infty$ of the following mixed power variations

$$\sum_{i=0}^{n-1} (\Delta^n_i W)^p \left(\Delta^n_i B^H\right)^r,$$

where $p \geq 0$, $r \geq 0$ are fixed integer numbers. Since $\Delta^n_i W$ and $\Delta^n_i B^H$ are centered Gaussian with variances $n^{-1/2}$ and n^{-H} respectively, we get that

$$\mathsf{E}\left[(\Delta^n_i W)^p \left(\Delta^n_i B^H\right)^r\right] = n^{-rH-p/2}\mu_p\mu_r,$$

where for an integer $m \geq 1$

$$\mu_m = \mathsf{E}\left[N(0,1)^m\right] = (m-1)!!\mathbf{1}_{m \text{ is even}}$$

is the mth moment of the standard Gaussian law; $(m-1)!! = (m-1)(m-3)\ldots$ is the double factorial.

In view of this, we will study centered sums of the form

$$S^{H,p,r}_n = \sum_{i=0}^{n-1} \left(n^{rH+p/2} (\Delta^n_i W)^p \left(\Delta^n_i B^H\right)^r - \mu_p\mu_r\right).$$

We start with studying the almost sure behavior of $S^{H,p,r}_n$. For brevity, the phrase "almost surely" will be omitted throughout the article.

Proposition 1. *Let $\varepsilon > 0$ be arbitrary.*

If $r = 0$, then $S_n^{H,p,r} = o(n^{1/2+\varepsilon})$, $n \to \infty$.

If p and $r \geq 2$ are even, then

- *for $H \in (0, 3/4]$ $S_n^{H,p,r} = o(n^{1/2+\varepsilon})$, $n \to \infty$.*
- *for $H \in (3/4, 1)$ $S_n^{H,p,r} = o(n^{2H-1+\varepsilon})$, $n \to \infty$.*

If p is odd and $r \geq 1$ is arbitrary, then for any $H \in (0, 1)$ $S_n^{H,p,r} = o(n^{1/2+\varepsilon})$, $n \to \infty$.

If p is even and r is odd, then

- *for $H \in (0, 1/2]$ $S_n^{H,p,r} = o(n^{1/2+\varepsilon})$, $n \to \infty$.*
- *for $H \in (1/2, 1)$ $S_n^{H,p,r} = o(n^{H+\varepsilon})$, $n \to \infty$.*

In particular, for any $H \in (0, 1)$ the following version of the ergodic theorem takes place: $S_n^{H,p,r} \to 0$, $n \to \infty$.

The following theorem summarizes the weak limit behaviour of $S_n^{H,p,r}$. We remark that some (but not all) of the results can be obtained either from the limit theorems for stationary Gaussian sequences of vectors, see e.g. [3] or from the limit theorems for arrays of Gaussian vectors, see [4]. However, we believe that our approach (using one-dimensional limit theorems) is more accessible and leads quicker to the desired results.

Denote

$$\rho_H(m) = \mathsf{E}\left[B_1^H(B_{m+1}^H - B_m^H)\right] = \frac{1}{2}\left(|m+1|^{2H} + |m-1|^{2H} - 2m^{2H}\right)$$

the covariance of the so-called fractional Gaussian noise $\{B_{k+1}^H - B_k^H\}$. It is easy to see that $\rho_H(m) \sim H(2H-1)m^{2H-2}$, $m \to \infty$.

Theorem 1. *If p and r are even, $r \geq 2$, then*

- *for $H \in (0, 3/4)$*

$$n^{-1/2} S_n^{H,p,r} \Rightarrow N(0, \sigma_{H,r}^2 \mu_p^2 + \sigma_{p,r}^2), \ n \to \infty, \tag{3}$$

where

$$\sigma_{H,r}^2 = \sum_{l=1}^{r/2} \frac{(l!)^2}{(2l)!((r-2l)!!)^2} \sum_{m=-\infty}^{\infty} \rho_H(m)^{2l}, \quad \sigma_{p,r}^2 = \mu_{2r}\left(\mu_{2p} - \mu_p^2\right);$$

- *for $H = 3/4$*

$$\frac{S_n^{3/4,p,r}}{\sqrt{n \log n}} \Rightarrow N(0, \sigma_{3/4,r}^2 \mu_p^2 + \sigma_{p,r}^2), \ n \to \infty, \tag{4}$$

where $\sigma_{3/4,r} = 3r(r-1)/4$;

- *for* $H \in (3/4, 1)$

$$n^{1-2H} S_n^{H,p,r} \Rightarrow \zeta_{H,p,r},\ n \to \infty, \tag{5}$$

where $\zeta_{H,p,r}$ *is a special "Rosenblatt" random variable.*

If p *is odd and* $r \geq 1$ *is arbitrary, then for any* $H \in (0, 1)$

$$n^{-1/2} S_n^{H,p,r} \Rightarrow N(0, \mu_{2p}\mu_{2r}). \tag{6}$$

If p *is even and* r *is odd, then*

- *for* $H \in (0, 1/2]$

$$n^{-1/2} S_n^{H,p,r} \Rightarrow N(0, \sigma_{H,r}^2 \mu_p^2 + \sigma_{p,r}^2),\ n \to \infty, \tag{7}$$

where $\sigma_{H,1} = 0$,

$$\sigma_{H,r}^2 = \sum_{l=1}^{(r-1)/2} \frac{(r!)^2}{(2l+1)!((r-2l-1)!!)^2} \sum_{m=-\infty}^{\infty} \rho_H(m)^{2l+1}, \quad r \geq 3;$$

- *for* $H \in (1/2, 1)$

$$n^{-H} S_n^{H,p,r} \Rightarrow N(0, \mu_p^2 \mu_{r+1}^2),\ n \to \infty. \tag{8}$$

Remark. For $r = 0$ we have the pure Wiener case, so for any $H \in (0, 1)$

$$n^{-1/2} S_n^{H,p,r} \Rightarrow N(0, \mu_{2p} - \mu_p^2), \quad n \to \infty.$$

3 Statistical estimation in mixed model based on quadratic variation

Now we turn to the question of parametric estimation in the mixed model

$$M_t^H = aB_t^H + bW_t,\ t \in [0, T], \tag{9}$$

where a, b are non-zero numbers, which we assume to be positive, without loss of generality. Our primary goal is to construct a strongly consistent estimator for the Hurst parameter H, given a single observation of M^H.

It is well-known (see [7]) that for $H \in (3/4, 1)$ the measure induced by M^H in $C[0, T]$ is equivalent to that of bW. Therefore, the property of almost sure convergence in this case is independent of H. Consequently, no strongly consistent estimator for $H \in (3/4, 1)$ based on a single observation of M^H exists.

In this section we denote $\Delta_i^n X = X_{T(i+1)/n} - X_{Ti/n}$ and

$$V_n^{H,p,r} = \sum_{i=0}^{n-1} (\Delta_i^n W)^p \left(\Delta_i^n B^H\right)^r.$$

Consider the quadratic variation of M^H, i.e.

$$V_n^{H,2} := \sum_{i=0}^{n-1} \left(\Delta_i^n M^H\right)^2 = a^2 V_n^{H,0,2} + 2ab V_n^{H,1,1} + b^2 V_n^{H,2,0}.$$

Note that $V_n^{H,2}$ depends only on the observed process but not on H. We use this notation to specify the distribution. Namely, we will use it to refer to the limit behavior of the quadratic variation for a specified value of the Hurst parameter H.

By Proposition 1, we have that $V_n^{H,0,2} \sim T^{2H} n^{1-2H}$, $V_n^{H,2,0} \to T$, $V_n^{H,1,1} = o(n^{1/2-H})$, $n \to \infty$. Therefore, the asymptotic behavior of $V_n^{H,2}$ depends on whether $H < 1/2$ or not. Precisely, for $H \in (0, 1/2)$,

$$V_n^{H,2} \sim a^2 T^{2H} n^{1-2H}, \ n \to \infty, \tag{10}$$

so the quadratic variation behaves similarly to that of a scaled fBm.

For $H \in (1/2, 1)$,

$$V_n^{H,2} \to b^2 T, \ n \to \infty, \tag{11}$$

so the quadratic variation behaves similarly to that of a scaled Wiener process.

Let us consider the cases $H < 1/2$ and $H > 1/2$ individually in more detail.

3.1 $H \in (0, 1/2)$

We have seen above that this case is similar to the pure fBm case. Unsurprisingly, the same estimators work, which is precisely stated below.

Theorem 2. *For $H \in (0, 1/2)$, the following statistics*

$$\widehat{H}_k = \frac{1}{2}\left(1 - \frac{1}{k}\log_2 V_{2^k}^{H,2}\right)$$

and

$$\widetilde{H}_k = \frac{1}{2}\left(\log_2 \frac{V_{2^{k-1}}^{H,2}}{V_{2^k}^{H,2}} + 1\right)$$

are strongly consistent estimators of the Hurst parameter H.

Remark. At the first sight, there is no clear advantage of $\widehat{H}_k$ or $\widetilde{H}_k$. But a careful analysis shows that $\widetilde{H}_k$ is better. Indeed, it is easy to see that

$$\widehat{H}_k = H - \frac{\log_2 a + H \log_2 T}{k} + o(k^{-1}), k \to \infty, \tag{12}$$

while

$$\widetilde{H}_k = H + O(2^{k(2H-1)}) + o(2^{k(-1/2+\varepsilon)}), k \to \infty. \tag{13}$$

Now it is absolutely clear that $\widetilde{H}_k$ performs much better (unless one hits the jackpot by having $aT^H = 1$).

References

[1] S. Achard and J.-F. Coeurjolly. Discrete variations of the fractional Brownian motion in the presence of outliers and an additive noise. *Stat. Surv.*, 4:117–147, 2010.

[2] T. Androshchuk and Y. Mishura. Mixed Brownian–fractional Brownian model: absence of arbitrage and related topics. *Stochastics*, 78(5):281–300, 2006.

[3] M. A. Arcones. Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors. *Ann. Probab.*, 22(4):2242–2274, 1994.

[4] J.-M. Bardet and D. Surgailis. Moment bounds and central limit theorems for Gaussian subordinated arrays. *J. Multivariate Anal.*, 114:457–473, 2013.

[5] A. Begyn. Asymptotic expansion and central limit theorem for quadratic variations of Gaussian processes. *Bernoulli*, 13(3), 2007.

[6] C. Cai, P. Chigansky, and M. Kleptsyna. The maximum likelihood drift estimator for mixed fractional brownian motion. 2012. Preprint, available online at http://arxiv.org/abs/1208.6253.

[7] P. Cheridito. Mixed fractional Brownian motion. *Bernoulli*, 7(6):913–934, 2001.

[8] J.-F. Coeurjolly. Estimating the parameters of a fractional Brownian motion by discrete variations of its sample paths. *Stat. Inference Stoch. Process.*, 4(2):199–227, 2001.

[9] Y. Kozachenko, A. Melnikov, and Y. Mishura. On drift parameter estimation in models with fractional Brownian motion. *Statistics*, 2012. To appear, available online at http://arxiv.org/abs/1112.2330.

[10] Y. Mishura and G. Shevchenko. Mixed stochastic differential equations with long-range dependence: Existence, uniqueness and convergence of solutions. *Comput. Math. Appl.*, 64(10):3217–3227, 2012.

[11] Y. S. Mishura. *Stochastic calculus for fractional Brownian motion and related processes*, volume 1929 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2008.

ROBUST ESTIMATION APPROACH FOR HAZARDOUS CONCENTRATION LEVELS USING SPECIES SENSITIVITY DISTRIBUTION

G. S. Monti[1], P. Filzmoser[2], R. Deutsch[2]
[1]*Dep. Economics, Management and Statistics, University of Milano-Bicocca Milan, ITALY*
[2]*Institute of Statistics and Mathematical Methods in Economics, TU Wien Vienna, AUSTRIA*
e-mail: [1]gianna.monti@unimib.it

Abstract

A commonly used tool in probabilistic risk assessment is the species sensitivity distribution (SSD) which allows to establish guidelines for setting environmental quality standards. SSD models the variation in sensitivity of species, considered representative of the ecological community they belong, to a particular toxic compound [4, 1]. We propose robust estimation approach for hazardous concentration thresholds for p% of species (HCp) in order to take into account the presence of outliers in the data or data skewness, which may occur without any ecological reason. Unusual observations are down-weighted rather than eliminated, with the advantage of not reducing the already small sample size and therefore not losing precision of the estimators [3]. Data transformations in conjunction with robust estimation methods are recommended in case of heteroscedasticity [2]. Different scenarios using real data sets as well as simulated data are presented in order to illustrate and compare the proposed approaches. As a by-product, robust methods also allow to identify data outliers, which have an important message for practitioners due to a different behavior of specific species.

References

[1] Hickey G.L., Craig P.S. (2012). Competing statistical methods for the L-fitting of normal species sensitivity distributions: Recommendations for practitioners.*Risk Analysis*, Vol. **32**(7), pp. 1232–1243.

[2] Marazzi A., Yohai V.J (2006) Robust Box-Cox transformations based on minimum residual autocorrelation. *Computational Statistics & Data Analysis*, Vol. **50**(10), pp. 2752–2768.

[3] Maronna R.A., Martin R.D., Yohai V.J. (2006). *Robust Statistics*. John Wiley & Sons, Ltd.

[4] Posthuma P., Suter G.W., Traas T.P. (2002). *Species Sensitivity Distribution in Ecotoxicology*. Lewis Publishers, Boca Raton, FL.

EXTRACTING INFORMATION FROM INTERVAL DATA USING SYMBOLIC PRINCIPAL COMPONENT ANALYSIS

M.R. Oliveira[1], M. Vilela[2], A. Pacheco[3], R. Valadas[4], P. Salvador[5]
[1,2,3]*CEMAT and* [1,2,3]*DM,* [4]*Instituto Superior Técnico, Universidade de Lisboa*
[5]*Universidade de Aveiro*
[4,5]*Instituto de Telecomunicações*
[1,2,3,4]*Lisbon and* [5]*Aveiro, PORTUGAL*
e-mail: [1]rosario.oliveira@tecnico.ulisboa.pt,
[2]margarida.azeitona@tecnico.ulisboa.pt,
[3]apacheco@math.tecnico.ulisboa.pt,
[4]rui.valadas@tecnico.ulisboa.pt, [5]salvador@av.it.pt

Abstract

We address the definition of symbolic variance and covariance for random interval-valued variables, and present four known symbolic principal component estimation methods using a common insightful framework. In addition, we provide a simple explicit formula for the scores of the symbolic principal components, equivalent to the representation by Maximum Covering Area Rectangle. Furthermore, the analysis of a real dataset leads to a meaningful characterization of Internet traffic applications.

1 Introduction

The low cost of information storage combined with recent advances in search and retrieval technologies has made huge amounts of data available, the so-called *big data* explosion. New statistical analysis techniques are now required to deal with the volume and complexity of this data. One promising technique is Symbolic Data Analysis (SDA), introduced in the late 1980s by Edwin Diday.

In conventional data analysis, the variables that characterize an object can only take single values. SDA introduces symbolic random variables which can take values over complex data structures like lists, intervals, histograms or even distributions. Symbolic data may exist on their own right or may result from the aggregation of a base dataset according to the researchers interest.

For example, suppose that our goal is to characterize the ages of university teachers. The variable that records the teachers' age will have as many observations as teachers, and these can differ among universities. Let us assume that a given university has 1000 teachers, and the values $\omega_1, \ldots, \omega_{1000}$ are the teachers' ages. SDA calls these values *micro-data*. In conventional statistical analysis, the universities would have to be characterized by single-valued variables, e.g. the mean teachers' age. SDA can deal with more complex data structures, called *macro-data*. For example, the teachers' age can be aggregated into one interval or various intervals. Our main interest in this paper is on interval-valued data, where macro-data corresponds to the interval between minimum and maximum of micro-data values: $[a, b] = [\min\{\omega_1, \ldots, \omega_{1000}\}, \max\{\omega_1, \ldots, \omega_{1000}\}]$.

The paper is organized as follows. Section 2 presents basic descriptive statistics, including symbolic variances and covariances, for interval-valued data. Section 3 introduces Symbolic Principal Component Analysis (SPCA) for interval-valued data. Section 4 uses SPCA on the analysis of Internet data produced by six different Internet applications. Finally, some conclusions are drawn in Section 5.

2 Basic descriptive statistics

There have been several proposals for definitions of symbolic versions of sample mean, variance, covariance, and correlation, according to various types of symbolic data and including interval-valued data [1].

We assume that the collected interval-valued data are realizations of random vectors. As such, we consider a random interval-valued vector $\boldsymbol{X} = (X_1, \ldots, X_p)^t$, where $X_j = [A_j, B_j]$, with A_j and B_j being random variables verifying $P(A_j \leq B_j) = 1$, denotes the j-th random interval-valued variable of $\boldsymbol{X}$. Even though this is the common representation of random interval-valued variables, we follow the approach of [2, 3, 6] and write the intervals X_j in terms of their centers, $C_j = (A_j + B_j)/2$, and their ranges, $R_j = B_j - A_j$. This choice leads to a clear interpretation of an interval in terms of its "location" on the real line along with its length; moreover it enables for the unification of several results in the literature (cf. [2, 3, 6] and references therein). Likewise, the random vector $\boldsymbol{X}$ is equivalently represented by the random vector of centers, $\boldsymbol{C} = (C_1, \ldots, C_p)^t$, and the random vector of ranges, $\boldsymbol{R} = (R_1, \ldots, R_p)^t$.

Let $(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n)^t$ and $(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n)^t$ denote the vectors of centers and ranges obtained from a random sample of size n from $\boldsymbol{X}$, where $\boldsymbol{C}_i = (C_{i1}, \ldots, C_{ip})^t$ and $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{ip})^t$ characterizes the i-th entity or object of the sample. In this setting, a natural proposal for sample symbolic mean of the interval-valued variable X_j is to use the traditional sample mean of the centers, $\overline{X}_j = \overline{C}_j$ with $\overline{C}_j = \sum_{i=1}^n C_{ij}/n$.

As concerns the sample symbolic variance of the interval-valued variable X_j, we express the proposals available in the literature as the sum of two components, the first accounting for the variability of the associated centers and the second for the size of the associated ranges, in the form

$$S_{jj}^{(\alpha)} = \sum_{i=1}^{n} \frac{\left(C_{ij} - \overline{C}_j\right)^2}{n} + \alpha \sum_{i=1}^{n} \frac{R_{ij}^2}{n}, \tag{1}$$

with the nonnegative weight α accounting for the relevance given to the ranges. In particular, we address three cases, with respective values 0, 1/4, 1/12 for the weight α. The first case ($\alpha = 0$) ignores the values of the ranges, simply turning the symbolic variance into the variance of the centers. Concerning the second case ($\alpha = 1/4$), we note that as $R_{ij}/2$ represents the radius of the interval associated with i-th entity, measured on the j-th random interval-valued variable, $\sum_{i=1}^n R_{ij}^2/(4n)$ may be interpreted as the sample second order moment of the radius of the j-th random interval-valued variable. The third case ($\alpha = 1/12$) corresponds to choosing the weight derived in [2] assuming that micro-data are uniformly distributed on the random intervals.

In the same manner, we consider proposals for the sample symbolic covariance between two interval-valued variables X_j and X_l that express it as the sum of two components, the first accounting for the sample covariance of the associated centers and the second for the size of the associated ranges, in the form

$$S_{jl}^{(\beta)} = \sum_{i=1}^{n} \frac{(C_{ij} - \overline{C}_j)(C_{il} - \overline{C}_l)}{n} + \beta \sum_{i=1}^{n} \frac{R_{ij} R_{il}}{n}, \tag{2}$$

with the nonnegative weight β accounting for the relevance given to the ranges associated to the interval-valued variables X_j and X_l.

In sequence, we may use (1)-(2) to construct a sample symbolic covariance matrix $\boldsymbol{S}^{(\alpha,\beta)}$ having on the diagonal the sample symbolic variances $S_{jj}^{(\alpha)}$, given in (1), and outside the diagonal the sample symbolic covariances $S_{jl}^{(\beta)}$, $j \neq l$, given in (2), leading to

$$\boldsymbol{S}^{(\alpha,\beta)} = \boldsymbol{S}_{CC} + (\alpha - \beta)\, \mathrm{Diag}\left(\frac{\boldsymbol{\mathcal{R}}^t \boldsymbol{\mathcal{R}}}{n}\right) + \beta\, \frac{\boldsymbol{\mathcal{R}}^t \boldsymbol{\mathcal{R}}}{n}, \tag{3}$$

with $\boldsymbol{S}_{CC}$ denoting the sample covariance matrix of the centers and $\boldsymbol{\mathcal{R}} = [R_{ij}]$ the $(n \times p)$ matrix of observed ranges. Particular cases of sample symbolic covariance matrices, $\boldsymbol{S}^{(\alpha,\beta)}$, with $\alpha \in \{0, 1/4, 1/12\}$ and $\beta = \alpha$ or $\beta = 0$, have been introduced in the literature ([2, 6] and references therein). Details about the links between these sample symbolic covariance matrices and SPCA for interval-valued data are discussed in the next section.

3 Symbolic Principal Component Analysis

Principal component analysis (PCA) is one of the most popular statistical methods to analyse real data. There have been several proposals to extend this methodology to the symbolic data analysis framework, in particular to interval-valued data. The majority of the available methods rely on a strategy called symbolic-conventional-symbolic, meaning that: (i) input data is symbolic (interval-valued, in here), (ii) the data is converted into conventional, to which the conventional PCA method is applied, and (iii) at the end, the PCA results are turned into symbolic, usually by a method called Maximum Covering Area Rectangle (MCAR), see [3, 6] and references therein for details.

We study four SPCA methods: CPCA, VPCA, CIPCA, and SymCovPCA. CPCA and VPCA corresponds to the first SPCA methods proposed in the literature and the last two are among the most recent alternatives. All these four methods rely on the symbolic-conventional-symbolic strategy, which can be specified as follows: (i) compute the associated $(p \times p)$ sample symbolic covariance matrix $\boldsymbol{S}^{(\alpha,\beta)}$ (see Table 1 and [3]); (ii) obtain the spectral decomposition of $\boldsymbol{S}^{(\alpha,\beta)}$, as in the conventional PCA, and (iii) transform the conventional scores into symbolic scores, e.g. using MCAR.

Note that $\boldsymbol{S}^{(\frac{1}{4},0)}$ and $\boldsymbol{S}^{(\frac{1}{12},0)}$ (see Table 1) are covariance matrices that use a definition of symbolic variance of an interval-valued variable that does not coincide with the definition of symbolic covariance between the same interval-valued variable and itself.

Table 1: Sample symbolic covariance matrices $\boldsymbol{S}^{(\alpha,\beta)}$, defined by the combination of several proposals for symbolic variances and covariances along with the corresponding SPCA method.

(α,β)	$\boldsymbol{S}^{(\alpha,\beta)}$	SPCA Method
(0,0)	$\boldsymbol{S}_{CC}$	CPCA
$(\frac{1}{4},\frac{1}{4})$	$\boldsymbol{S}_{CC}+\frac{1}{4}\frac{\boldsymbol{\mathcal{R}}^t\boldsymbol{\mathcal{R}}}{n}$	—
$(\frac{1}{12},\frac{1}{12})$	$\boldsymbol{S}_{CC}+\frac{1}{12}\frac{\boldsymbol{\mathcal{R}}^t\boldsymbol{\mathcal{R}}}{n}$	SymCovPCA
$(\frac{1}{4},0)$	$\boldsymbol{S}_{CC}+\frac{1}{4}\text{Diag}\left(\frac{\boldsymbol{\mathcal{R}}^t\boldsymbol{\mathcal{R}}}{n}\right)$	VPCA
$(\frac{1}{12},0)$	$\boldsymbol{S}_{CC}+\frac{1}{12}\text{Diag}\left(\frac{\boldsymbol{\mathcal{R}}^t\boldsymbol{\mathcal{R}}}{n}\right)$	CIPCA

This violates a basic rule in the conventional framework, namely that the variance of a variable equals the covariance of the variable with itself. In spite of this fact, the CIPCA's authors, who proposed $\boldsymbol{S}^{(\frac{1}{12},0)}$ [3], argue that this is an advantage of their method.

Similarly to the conventional PCA, it may be interesting to define the SPCA based on standardized interval-valued variables, and to do so we introduce the sample correlation matrix as: $\boldsymbol{P}^{(\alpha,\beta)}=\boldsymbol{U}_{(\alpha)}^{-1}\boldsymbol{S}^{(\alpha,\beta)}\boldsymbol{U}_{(\alpha)}^{-1}$, where $\boldsymbol{U}_{(\alpha)}=\text{Diag}\left(S_{11}^{(\alpha)},\ldots,S_{pp}^{(\alpha)}\right)^{1/2}$, for $\boldsymbol{S}^{(\alpha,\beta)}=[S_{jl}^{(\alpha,\beta)}]$, where $S_{jj}^{(\alpha,\beta)}=S_{jj}^{(\alpha)}$ and $S_{jl}^{(\alpha,\beta)}=S_{jl}^{(\beta)}$, for $j\neq l$. Equivalently, $\boldsymbol{S}^{(\alpha,\beta)}=\boldsymbol{U}_{(\alpha)}\boldsymbol{P}^{(\alpha,\beta)}\boldsymbol{U}_{(\alpha)}$. Thus, SPCA methods based on standardized interval-valued variables just have to use $\boldsymbol{P}^{(\alpha,\beta)}$ instead of $\boldsymbol{S}^{(\alpha,\beta)}$.

The most common way to transform conventional objects into symbolic ones for methods following the symbolic-conventional-symbolic strategy is the MCAR representation. Following the same line of work as before, in [3] we deduced an explicit formulation of the MCAR representation in terms of centers and ranges. Furthermore, the sample scores of the i-th object on the j-th symbolic principal component (SPC), according with MCAR, are:

$$\widehat{\text{SPC}}_{ij}=\left[\hat{\boldsymbol{\gamma}}_j^t(\boldsymbol{C}_i-\hat{\boldsymbol{\mu}}_C)-\frac{1}{2}|\hat{\boldsymbol{\gamma}}_j|^t\boldsymbol{R}_i,\ \hat{\boldsymbol{\gamma}}_j^t(\boldsymbol{C}_i-\hat{\boldsymbol{\mu}}_C)+\frac{1}{2}|\hat{\boldsymbol{\gamma}}_j|^t\boldsymbol{R}_i\right], \tag{4}$$

where $\hat{\boldsymbol{\gamma}}_j$ is the j-th eigenvector of $\boldsymbol{S}^{(\alpha,\beta)}$, the sample symbolic covariance matrix under consideration, $|\hat{\boldsymbol{\gamma}}_j|=(|\hat{\gamma}_{1j}|,\ldots,|\hat{\gamma}_{pj}|)^t$, and $\hat{\boldsymbol{\mu}}_C$ is the vector of center sample means.

As a direct consequence of (4), the centers of the scores, $\hat{\boldsymbol{\gamma}}_j^t(\boldsymbol{C}_i-\hat{\boldsymbol{\mu}}_C)$, are a linear combination of the centers of the original interval-valued variables, whose weights are given by the eigenvectors of the corresponding symbolic covariance matrix. Additionally, the scores ranges, $|\hat{\boldsymbol{\gamma}}_j|^t\boldsymbol{R}_i$, are also a linear combination of the original ranges, whose weights have the same magnitude as the centers but are all positive. This formulation makes clear that MCAR' score ranges are never negative.

4 Analysis of Internet Data

In this section we illustrate the use of SPCA through a dataset of Internet traffic, typically observed in backbone networks, and measured during July 2014. Specifically, the dataset contains traffic produced by six different Internet applications, namely Web browsing (produced by HTTP), file sharing (produced by Torrent), streaming, video (YouTube), port scans (produced by NMAP), and snapshots. The first four applications correspond to regular traffic and the last two to Internet attacks. The analysis usually aims at detecting the various Internet applications within a traffic aggregate and/or the separation between regular and illicit traffic.

The dataset comprises 917 traffic objects, corresponding to packet flows of specific applications, which we call *datastreams*. For each datastream, we registered five different traffic characteristics observed in 0.1 seconds intervals, during 5 minutes. The traffic characteristics registered were the following: number of upstream packets (PUp), number of downstream packets (PDw), number of upstream bytes (BUp), number of downstream bytes (BDw), and number of active TCP sessions (Ses). Thus, each object is characterized by a total of 3000 observations per traffic characteristic, which constitutes our micro-data.

The conventional approach to analyse this data is based on summary statistics of each traffic characteristic. In particular, [4, 5] used 8 summary statistics (minimum, 1^{st} quartile, median, mean, 3^{rd} quartile, maximum, standard deviation, and median absolute deviation) for the above five traffic characteristics, giving a total of 40 variables to describe the datastreams. This approach usually requires a pre-processing step to remove irrelevant and redundant variables; Pascoal [5] used a robust feature selection method based on mutual information for that purpose.

This dataset is naturally symbolic, since each traffic characteristic is multi-valued. SDA takes into consideration the complex structure of these data, and may lead to clearer interpretation and new insights. In our case, we will use interval-valued variables for each traffic characteristic (our macro-data), instead of the 8 summary statistics listed above.

Given the nature of the data and the existence of potential atypical observations among the micro-data, we decided to trim 1% of the lower and 1% of the higher values. This was only done for the regular applications given that illicit ones have few datastreams and small variability and would be completely eliminated from the dataset, even for such small trimming percentiles. Apart from that, and following the recommendations in [4, 5], data was smoothed using a logarithm transformation ($\ln(x+1)$, to overcome the existence of zeros). SPCs were estimated using the four methods under study. The conventional analysis of the eigenvalues of the various sample symbolic covariance matrices (not shown here, see [3] for details) suggests to retain two principal components, which explain between 80.3% (CIPCA) and 95.6% (SymCovPCA) of the total sample variance associated with $\boldsymbol{S}^{(\alpha,\beta)}$.

The results obtained with CPCA and SymCovPCA are similar, and so are the results obtained with VPCA and CIPCA. Moreover, these similarities are easily explained by the expressions of Table 1. For these reasons, only the estimates associated

Table 2: Eigenvectors of the sample symbolic covariance matrices for each estimation method, called loadings.

	SymCovPCA		CIPCA	
	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
$\ln(\text{PDw}+1)$	-0.264	-0.171	-0.125	-0.059
$\ln(\text{BDw}+1)$	**-0.730**	-0.043	**-0.932**	0.337
$\ln(\text{PUp}+1)$	-0.255	-0.168	-0.113	-0.070
$\ln(\text{BUp}+1)$	**-0.571**	0.075	-0.318	**-0.937**
$\ln(\text{Ses}+1)$	-0.079	**0.967**	-0.029	-0.027

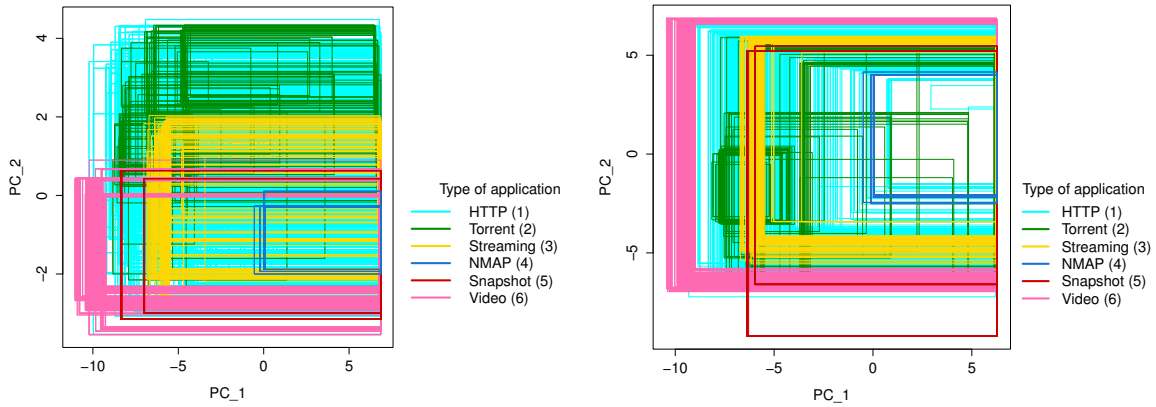
with the most recent methods (SymCovPCA and CIPCA) are shown in this paper.

Table 2 shows the loadings of the first and second SPC, obtained with SymCovPCA and CIPCA. In the case of SymCovPCA, the number of upstream and downstream bytes (BUp, BDw) have the highest loading (on absolute value) in the definition of the first SPC. Thus, the center and range of the first SPC can be interpreted as a weighted sum of the number of upstream and downstream bytes. The number of bytes is sometimes referred to as the traffic volume. For the center, the negative coefficients indicate that datastreams with high (low) number of bytes in both directions have low (high) center values on the first SPC. For the range, the coefficients are taken in absolute value, so datastreams with high (low) number of bytes in both directions have high (low) range values on the first SPC. Recall that the range expresses the inner variability of micro-data. As for the second SPC, the loading associated with number of sessions stands out. Thus, datastreams characterized by an high (low) number of sessions have high (low) center and range values on the second SPC.

The SymCovPCA scores are shown in Figure 1(a). Each datastream is represented by a rectangle, defined by the centers and ranges of the first two SPC. It can be said that the various Internet applications are, in general, well identified, since the datastreams show similar patterns for the same application. Most datastreams have a small minimum traffic volume (number of bytes), with the corresponding rectangles leaning to the right side. HTTP shows no distinctive characteristic, since the datastreams spread over all score ranges. This can be explained by the heterogeneity of user behaviours and accessed Web pages, typical of Web browsing. Torrent is concentrated on the upper part of the graph, due to its high number of sessions. The high number of sessions and large variability of the traffic volume is mostly explained by the variation on the number of available peers during traffic sharing sessions. The graph also suggests the existence of several Torrent groups, but this pattern will become clearer with the CIPCA method. The behaviour of video related with the second SPC contrasts with that of Torrent: it is concentrated in the lower part of the graph, due to its low number of sessions. Moreover, video is the application with the highest traffic volume. We may say that video datastreams are characterized by a low number of high volume sessions, and Torrent by a high number of high volume sessions. Streaming has a behaviour similar to video, but with higher number of sessions and lower traffic volume. NMAP is the application with smallest volume and variability, and has also

a relatively low number of sessions. Finally, the behaviour of snapshot is in-between video and streaming, both in terms of volume and number of sessions. Snapshot has two clear groups, that differ on the peak traffic volume, and correspond to full desktop and partial desktop uploads, respectively.

Table 2 shows that the loadings obtained with CIPCA are much higher (in absolute value) for BDw (first component) and BUp (second component). Thus, the first SPC can be interpreted as the number of bytes down (BDw) and the second one as the number of bytes up (BUp). The CIPCA scores are shown in Figure 1(b). Snapshot has the highest upstream peak traffic volume, and is now better separated from video and streaming. NMAP is again the application with smaller rectangles. However, it is now better separated from HTTP, since most HTTP datastreams have higher traffic volume range simultaneously in the upstream and downstream directions. Video and streaming are also well separated, since video datastreams have consistently higher traffic volume ranges simultaneously in both directions. Regarding Torrent, it is now possible to distinguish among three groups: the group centers occur at approximately the same upstream traffic volume; one group has small traffic range in both directions (small rectangles) and high downstream volume, another has high traffic ranges in the downstream direction but small in the upstream direction, and a third one has small downstream volumes but high upstream traffic ranges. These groups emerge from differences on the relative location of peers and the quality/stability of links. The first group corresponds to closer peers from which it is possible to download at higher speeds, the third to farther peers for which the links are less stable and unable to download at high speeds, and the third group is a mixture of the two previous ones.



(a) SymCovPCA. (b) CIPCA.

Figure 1: Symbolic scores, estimated by MCAR method.

5 Conclusion

Starting from the definition of symbolic variance and covariance for random interval-valued variables, we have used a common insightful framework to present four symbolic

principal component estimation methods that rely on a symbolic-conventional-symbolic strategy: CPCA, VPCA, CIPCA, and SymCovPCA.

The analysis of a symbolic dataset containing Internet traffic lead to a clear interpretation of the underlying Internet applications (Web browsing, file sharing, streaming, video, port scans, and snapshots). The analysis highlighted the difficulties in separating illicit traffic from regular one, suggesting the need to develop outlier detection methods for symbolic data.

.

Acknowledgements

This work has been supported by Fundação para a Ciência e Tecnologia (FCT), Portugal, through the projects UID/Multi/04621/2013 and PTDC/EEI-TEL/5708/2014.

References

[1] Billard L., Diday E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons, Chichester.

[2] Oliveira M.R., Vilela M., Pacheco A., Valadas R., Salvador P., (20XX). Population Symbolic Covariance Matrices for Interval Data. *In preparation*.

[3] Vilela M., Oliveira M.R., Pacheco A., Valadas R., Salvador P., (20XX). Population Symbolic Principal Component Analysis for Interval Data. *In preparation*.

[4] Pascoal C., Oliveira M.R., Valadas R., Filzmoser P., Salvador P., Pacheco A. (2012). Robust feature selection and robust PCA for Internet traffic anomaly detection. *In INFOCOM, 2012 Proceedings IEEE, Orlando, USA*, pp. 1755-1763.

[5] Pascoal C. (2014). Contributions to Variable Selection and Robust Anomaly Detection in Telecommunications. *PhD Thesis*, Instituto Superior Técnico, Universidade de Lisboa, Portugal.

[6] Vilela M. (2015). Classical and Robust Symbolic Principal Component Analysis for Interval Data. *Master Thesis*, Instituto Superior Técnico, Universidade de Lisboa, Portugal.

SOME FRACTIONAL EXTENSIONS OF THE POISSON PROCESS

Enzo Orsingher
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
Rome, ITALY
e-mail: enzo.orsingher@uniroma1.it

In recent years some fractional generalisations of the Poisson process appeared. The space-fractional Poisson process $N^\alpha(t)$, $t > 0$, recently studied, share with the homogeneous Poisson process $N(t)$, $t > 0$, the property of independence of increments. The space-fractional Poisson process $N^\alpha(t)$ is a time-changed Poisson process

$$N^\alpha(t) = N(S^\alpha(t)), \quad 1 < \alpha < 0, \tag{1}$$

where $S^\alpha(t)$ is a stable subordinator. The probability generating function of N^α reads

$$\mathbb{E}u^{N^\alpha(t)} = e^{-t(\lambda(1-u))^\alpha}, \quad t > 0, \lambda > 0, |u| \leq 1. \tag{2}$$

From (2) the probability distribution $p_k(t)$, $k \geq 0$ of $N^\alpha(t)$ can be extracted together with the moments and satisfies the difference-differential equation

$$\frac{dp_k}{dt} = -\lambda^\alpha (I - B)^\alpha p_k, \tag{3}$$

where B is the shift operator.
For the hitting tims T_k of levels k, namely

$$T_k^\alpha := \inf\,(s : N^\alpha(s) = k), \quad k \geq 1, \tag{4}$$

we are able to show that

$$P\{T_k^\alpha < \infty\} = \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)} \frac{1}{k!} < 1, \quad \forall\ k \geq 1, \tag{5}$$

and study its behavior with respect to k and α.
For the n-times iterated subordinator

$$N^\alpha(S^{\gamma_1}(S^{\gamma_2}(\dots S^{\gamma_n}(t)))) \stackrel{d}{=} N^{\alpha \prod_{j=1}^n \gamma_j}(t), \quad \gamma_j \in (0,1) \tag{6}$$

we study the limiting behavior. When $\prod_{j=1}^\infty \gamma_j = 0$ the limiting process of (6) is a degenerate r.v. with values 0 and ∞. If $0 < \prod_{j=1}^n \gamma_j < 1$, instead, we still have a space-fractional Poisson process. Also the space-time fractional Poisson process $N^{\alpha,\nu}(t)$, $t > 0$ is studied and we show that it has not a renewal structure. Its p.g.f. has the form

$$G_{\alpha,\nu}(t) = E_{\nu,1}(-t^\nu(\lambda(1-u))^\alpha), \quad |u| \leq 1, \nu, \alpha \in (0,1) \tag{7}$$

and for $\nu = 1$ coincides with the space-fractional Poisson while for $\alpha = 1$ gives the time-fractional Poisson process. A large class of generalized Poisson processes is obtained by considering processes with p.g.f.

$$\mathbb{E}u^{N^f(t)} = e^{-tf(\lambda(1-u))}, \tag{8}$$

where f is a Bernstein function, that is a function with integral representation

$$f(x) = \int_0^\infty (1 - e^{-xs})\nu(ds), \tag{9}$$

ν being the so-called Lévy measure on $(0, \infty)$. For $f = x^\alpha$, we have the space-fractional Poisson process while for different forms of f we have a large class of generalized Poisson processes, sharing the property of independence of increments and the characteristic that jumps have arbitrary size. Furthermore $N^f(t)$ is a time-changed Poisson process

$$N^f(t) = N(S^f(t)), \tag{10}$$

where $S^f(t)$ is a general subordinator related to the Bernstein function f. A special attention is devoted to the cases $f(x) = (x + \lambda)^\alpha - \lambda^\alpha$ and $f(x) = \ln(1 + x)$.

References

[1] Beghin L. (2015). Fractional gamma and gamma-subordinated processes. *Stoch. Anal. Appl.* Vol. **33**(5), pp. 903–926.

[2] Kumar A., Nane E., Vellaisamy P. (2011). Time-changed Poisson processes. *Statist. Prob. Lett.* Vol. **81**, pp. 1899–1910.

[3] Orsingher E., Polito F. (2012). The space-fractional Poisson process. *Statistics and Probability Letters*. Vol. **82**, pp. 852–858.

[4] Orsingher E., Toaldo B. (2015). Counting processes with Bernstein intertimes and random jumps. *J. Applied Probability*. Vol. **52**, pp. 1028–1044.

[5] Polito F., Scalas E. (2016). A Generalization of the Space-Fractional Poisson Process and its Connection to some Lévy Processes. *Electronic Communications in Probability*. Vol. **21**(20), pp. 1–14.

[6] Schilling R.L., Song R., Vondracek Z. (2012). *Bernstein functions. Theory and applications*. De Gruyter, Berlin.

TWO-SIDED INEQUALITIES FOR THE AVERAGE NUMBER OF ELEMENTS IN THE UNION OF IMAGES OF FINITE SET UNDER ITERATIONS OF RANDOM EQUIPROBABLE MAPPINGS

A. A. Serov[1], A. M. Zubkov[2]
Steklov Mathematical Institute, Russian Academy of Sciences
Moscow, RUSSIA
e-mail: [1]serov@mi.ras.ru, [2]zubkov@mi.ras.ru

Abstract

Let $\mathcal{N}$ be a set of N elements and $F_1, F_2, \ldots$ be a sequence of random independent equiprobable mappings $\mathcal{N} \to \mathcal{N}$. For a subset $S_0 \subset \mathcal{N}$, $|S_0| = n$, we consider a sequence of its images $S_k = F_k(\ldots F_2(F_1(S_0))\ldots)$, $k = 1, 2\ldots$, and a sequence of their unions $\Psi_k = S_1 \cup \ldots \cup S_k$, $k = 1, 2\ldots$ An approach to the exact computation of distribution of $|S_k|$ and $|\Psi_k|$ for moderate values of N is described. Two-sided inequalities for $\mathbf{M}|S_k|$ and $\mathbf{M}|\Psi_k|$ such that upper bound are asymptotically equivalent to lower ones for $N, n, k \to \infty, nk = o(N)$ are derived. The results are of interest for the analysis of time-memory tradeoff algorithms.

This work was supported by RFBR, grant 14-01-00318.

1 Introduction

One of the well-known time-consuming task is the search for solution of the equation

$$G(x) = a, \tag{1}$$

where G be a mapping of the finite set $\mathcal{N} = \{1, \ldots, N\}$ to itself such that the complexity of any known method to compute the value $G^{-1}(a)$ is comparable with exhaustive search over the entire set $\mathcal{N}$. The trivial method of searching the solution of the equation (1) is the sequential computation of values $G(x)$ for all $x \in \mathcal{N}$ until the solution of (1) will be found. The implementation of such method requires a memory of slowly growing size for $N \to \infty$ (necessary to calculate a value of the function G for any $x \in \mathcal{N}$), but the time (number of operations) needed this method has the order $O(N)$.

M. E. Hellman [2] proposed the universal (independent of the type of function G) method for searching the solutions of the equation (1), permitting (after the preliminary stage of the complexity $O(N)$) to find the solution of equation (1) with a high probability for a time in order less than $O(N)$ by means of tables having volume less than $O(N)$. This approach has been called the time-memory tradeoff.

We consider a simplified mathematical model of the "rainbow" table construction (this model corresponds to the version of the time-memory tradeoff method that has

been proposed in [6]). The model is as follows: an initial subset $S_0 \subset \mathcal{N}$, $|S_0| = n$, is chosen and its images

$$S_1 = F_1(S_0),\ S_2 = F_2(F_1(S_0)), \dots, S_t = F_t(F_{t-1}(\dots(F_1(S_0))\dots))$$

are calculated, where $F_1, \dots, F_t$ are independent random mappings of the set $\mathcal{N}$ to itself having uniform distribution on the set Σ_N, $|\Sigma_N| = N^N$, of all such mappings.

We propose the method to compute distributions of random variables $\varphi_k = |S_k|$ and $\zeta_t = |S_1 \cup S_2 \cup \dots \cup S_t|$ by means of Markov chains, applicable for moderate values of N, and obtain two-sided estimates for the expectation of these random variables and for the probabilities that an element $x \in \mathcal{N}$, independent of the iterated mappings $F_1, F_2, \dots$, belongs to the set S_k or to the set $S_1 \cup S_2 \cup \dots \cup S_t$. Upper and lower bounds are asymptotically equivalent for $N, n, t \to \infty$, if $nt = o(N)$. These results may be used to optimize the methods of the time-memory tradeoff.

2 Basic results

Let, as before, $F_1, F_2, \dots$ be independent random mappings of the set $\mathcal{N} = \{1, \dots, N\}$ to itself, $S_0 \subset \mathcal{N}$, $|S_0| = n$, $S_k = F_k(S_{k-1})$, $\Psi_k = \cup_{j=1}^{k} S_j$, $k \geqslant 1$. Let $\varphi_0 = |S_0|$, $\zeta_0 = 0$, $\varphi_k = |S_k|$, $\zeta_k = |\Psi_k|$, $k \geqslant 1$. Since the mappings $F_1, F_2, \dots$ are independent and identically distributed, the sequences $\{\varphi_k\}_{k\geqslant 0}$ and $\{\zeta_k\}_{k\geqslant 0}$ form the Markov chains.

Assertion 1. The transition probability matrix of the Markov chain $\{\varphi_k\}_{k\geqslant 0}$ has the form

$$P = \|p_{i,j}\|_{i,j=1}^{N},$$

$$p_{i,j} = \begin{cases} \binom{N}{j} \left(\frac{j}{N}\right)^i \sum\limits_{m=0}^{j} (-1)^m \binom{j}{m} \left(1 - \frac{m}{j}\right)^i, & 1 \leqslant j \leqslant i \leqslant N, \\ 0, & j > i. \end{cases}$$

The transition probability matrix of the Markov chain $\{(\varphi_k, \zeta_k)\}_{k\geqslant 0}$ has the form

$$Q = \|q_{(i,r),(j,s)}\|_{i,j,r,s=1}^{N},$$

$$q_{(i,r),(j,s)} = \begin{cases} p_{i,j} \frac{\binom{N-r}{s-r}\binom{r}{j-s+r}}{\binom{N}{j}} = \binom{N-r}{s-r}\binom{r}{j-s+r} \left(\frac{j}{N}\right)^i \sum\limits_{m=0}^{j} (-1)^m \binom{j}{m} \left(1 - \frac{m}{j}\right)^i, \\ \qquad \text{if } 1 \leqslant j \leqslant i \leqslant N,\ 1 \leqslant r \leqslant s \leqslant \min\{N, r+j\}, \\ 0 \qquad \text{otherwise.} \end{cases}$$

The transition probabilities of the Markov chain $\{\varphi_k\}_{k\geqslant 0}$ for k steps form the matrix $P^{(k)} = \|p_{(i,j)}^{(k)}\|_{i,j=1}^{N} = P^k$. Thus the collections of numbers $\{p_{(n,j)}^{(k)} = \mathbf{P}\{\varphi_k = j \,|\, \varphi_0 = n\},\ j = 1, \dots, N\}$ define the distributions of φ_k that allows to find the numerical values of the distribution characteristics of φ_k for the moderate values of N.

The two-sided estimates of $\mathbf{P}\{x \in S_k \,|\, \varphi_0 = n\}$, $\mathbf{P}\{x \in \Psi_k \,|\, \varphi_0 = n\}$ and the first moments of the random variables φ_k, ζ_k are contained in the following Theorem.

Theorem 1. *Let $F_1, F_2, \ldots$ be the independent equiprobable mappings of the set $\mathcal{N} = \{1, \ldots, N\}$ to itself, $S_0 \subseteq \mathcal{N}$, $|S_0| = n$,$S_k = F_k(\ldots(F_1(S_0))\ldots)$, $k \geqslant 1$. For any element $x \in \mathcal{N}$, which does not depend on F_1, $F_2, \ldots$, for all $1 \leqslant k$, $n \leqslant N$ we have*

$$\begin{gathered} \frac{n}{N} - C_n^2 \frac{k}{N^2} \leqslant \mathbf{P}\{x \in S_k \,|\, \varphi_0 = n\} < \frac{n}{N} - C_n^2 \frac{k}{N^2} + \frac{n^3 k^2}{4N^3}\,, \\ \frac{nt}{N} - C_{t+1}^2 \frac{3n^2}{2N^2} < \mathbf{P}\left\{ x \in \Psi_t \,|\, \varphi_0 = n \right\} < \frac{nt}{N} - C_n^2 C_{t+1}^2 \frac{1}{N^2} + \frac{n^3(t+1)^3}{12N^3}\,. \end{gathered} \tag{2}$$

The following inequalities are also valid:

$$n - C_n^2 \frac{k}{N} \leqslant \mathbf{M}\{\varphi_k \,|\, \varphi_0 = n\} < n - C_n^2 \frac{k}{N} + \frac{n^3 k^2}{4N^2}\,, \tag{3}$$

$$\begin{gathered} nt - C_{t+1}^2 \frac{3n^2}{2N} < \mathbf{M}\left\{ \zeta_t \,|\, \varphi_0 = n \right\} < nt - C_n^2 C_{t+1}^2 \frac{1}{N} + \frac{n^3(t+1)^3}{12N^2}\,, \\ \mathbf{D}\{\varphi_k \,|\, \varphi_0 = n\} < \frac{kn^3}{N}\left(1 + \frac{(n+2)k}{4nN}\right). \end{gathered} \tag{4}$$

References

[1] Harris B. (1960). Probability distributions related to random mappings. *Ann. Math. Statist.* Vol. **31**, No. 2, pp. 1045–1062.

[2] Hellman M.E. (1980). A cryptanalytic time–memory trade-off. *IEEE Trans. Inf. Theory*. Vol. **26**, pp. 401–406.

[3] Flajolet P., Odlyzko A.M. (1990). Random Mapping Statistics *Advances in Cryptology — Proc. Eurocrypt'89, J-J. Quisquater Ed., Lect. Notes Comp. Sci.* Vol. **434**, pp. 329–354.

[4] Kolchin V.F., Sevastyanov B.A., Chistyakov V.P. (1978). Random allocations. *Scripta Series in Math.* V. H. Winston & Sons, Washington, pp. 262.

[5] Kolchin V.F. (1986). Random mappings. *Trans. Ser. in Math. and Eng.*, Optimization Software Inc. Publications Division, New York, pp. 207.

[6] Oechslin P. (2003). Making a faster cryptanalytic time-memory trade-off. *Lect. Notes Comput. Sci.* Vol. 2729, pp. 617–630.

[7] Zubkov A.M., Serov A.A. (2014). Images of subset of finite set under iterations of random mappings. *Discrete Math. Appl.* Vol. 2015, No. 3, pp. 179–185.

PERFORMANCE STUDY OF LINFOOT'S INFORMATIONAL CORRELATION COEFFICIENT AND ITS MODIFICATION

G.L. Shevlyakov[1], N.V. Vasilevskiy[2]
[1,2]*Peter the Great St. Petersburg Polytechnic University*
[1]*Institute for Problems of Mechanical Engineering, Russian Academy of Sciences*
Saint Petersburg, RUSSIA
e-mail: [1]Georgy.Shevlyakov@phmf.spbstu.ru

Abstract

Performance of the informational correlation coefficient (Linfoot, 1957) is experimentally studied. To reduce the bias of estimation, a symmetric version of this correlation measure is proposed. This modified informational correlation coefficient outperforms Linfoot's correlation measure at the bivariate normal distribution on large samples.

1 Introduction

Pearson's correlation coefficient is a well-defined measure of the linear dependence between continuous random variables X and Y. This partially refers to closely related to it rank measures as the quadrant, Spearman and Kendall correlation coefficients. However, if one is interested either in processing discrete data or in revealing the possible nonlinear relationship between random variables, then difficulties may arise both in the implementation of those classical measures as well as in their interpretation.

In the literature, several proposals were made to solve these problems, for instance, Gebelein's (1941), Sarmanov's (1958) correlation coefficients, and the distance correlation coefficient of Szekely (2007).

In what follows, we focus on the informational measures of association between random variables [7]. Joe's dependence measure [4] exploits the concept of the relative entropy that measures the similarity of two random variables with the distributions $p(x)$ and $q(x)$ in the discrete case

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \,.$$

Silvey [8] uses the measure of dependence between two random variables defined by the ratio of their joint density and the product of their marginal densities $\varphi(x,y) = p(x,y)/[p(x)p(y)]$. The introduced measure is defined as follows: $\Delta = E[d(x)]$, where $d(x) = \int_{y:\,\varphi(x,y)>1} \big[p(y|x) - p(y)\big]\, dy$. Thus, it can be rewritten as

$$\Delta = \int\int_{(x,y):\,\varphi(x,y)>1} \big[p(x,y) - p(x)p(y)\big]\, dxdy.$$

Granger [2] introduces another measure of dependence

$$S_p = \frac{1}{2}\int\int \left[p(x,y)^{1/2} - [p(x)p(y)]^{1/2}\right]^2 dxdy.$$

Joe's measure of dependence is not symmetric, and Silvey's and Granger's measures are hard to compute. Mutual information ($I(X,Y)$) for any pair of discrete and continuous random variables X and Y is defined as follows

$$I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,, \quad I(X,Y) = \int\int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}\,dxdy\,.$$

The informational correlation coefficient (ICC), firstly introduced by Linfoot [6], is defined as follows

$$\rho_{ICC}(X,Y) = \sqrt{1 - e^{-2I(X,Y)}}\,. \tag{1}$$

Note that ICC is equal to the classical Pearson's correlation coefficient at the bivariate normal distribution: $\rho_{ICC}(X,Y) = \rho$.

2 Problem Setting

In spite of the fact that ICC was introduced more than 60 years ago, its properties as a statistical measure of correlation have not yet been studied; it was not checked how well this measure estimates the correlation coefficient based on the sample of a given size. We are going to experimentally examine the following statistical properties of ICC: (i) unbiasedness, (ii) consistency, (iii) Monte Carlo performance on small ($N \leq 20$) and large samples, and (iv) robustness. Moreover, in order to improve the performance of ICC, namely, to reduce its bias, we propose and study a modified symmetric version of ICC denoted as $MICC$.

3 Monte Carlo Experiment

3.1 Description of the computational algorithm

All numerical experiments are performed using R language, especially its "entropy" library. The first problem is how to compute mutual information, which is used in (1). This is solved by applying a shrink-algorithm [3].

There exist several different algorithms of computing $I(X,Y)$; in our work, we choose the most precise one, not the fastest (for comparative analysis, see [3]). All experiments are performed at the standard bivariate normal distribution with density $f(x,y) = N(x,y;0,0,1.1,\rho)$.

The general algorithm can be described as follows:

1. Generate a sample of the fixed size: $N = 20, 60, 100, 1000, 10000$.

2. Extract x- and y-components from the sample, which are dependent random variables with the correlation coefficient ρ.

3. Construct the table of frequencies–the discrete analog of the joint distribution–we take a rectangle $[x_{min}, x_{max}] \times [y_{min}, y_{max}]$ on plane and divide it into $n_x \times n_y$ "bins" of equal size. Thus, the table of dimension n_x, n_y is built, each element of which is equal to the number of points in the corresponding bin.

4. Mutual information $I(X, Y)$ and ICC are computed using this table of frequencies.

This sequence is repeated 1000 times, allowing us to compute Monte Carlo estimates of the mean and variance of the correlation coefficient ρ: computations are performed for $\rho = 0, 0.1, 0.2, \ldots, 0.9, 1$; the number of bins is taken equal to 400. Typical results are exhibited in Fig. 1.

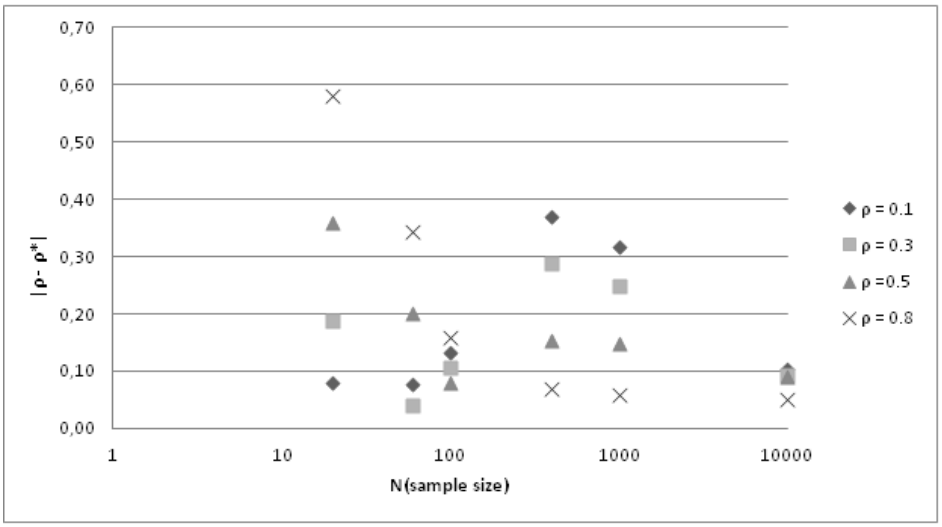


Figure 1: Monte-Carlo Biases of ICC

3.2 Monte Carlo results for *ICC*

1. From Fig. 1 it follows that estimation biases are considerably big (on small samples, they can even be greater than 0.5. Relatively small biases are observed only on large samples $N = 1000$ and$N = 10\,000$.

2. Satisfactory performance is observed in the case of a strong correlation—biases decrease with the growth of the sample size.

3. We may also add that the coefficient of variance is less than 0.2 for all examined combinations of (ρ, N).

4. *A remark on the choice of the number of bins.* The shrink-algorithm takes the table of frequencies as an input. It appeared that the algorithm performance depends on the relation N/K^2, where K is the linear dimension of the table. We observed that results are almost independent of the changes of K, as they

depend only on the coefficient $B = N/K^2$. For $\rho = 0.5$, the value $B = 7$ is optimal. Given a data sample, we can choose an appropriate value of K, which is a variable in our algorithm.

4 Main Result: A Symmetric Modification of *ICC*

Mutual entropy, also known as the Kullback-Leibler distance, has a serious disadvantage — it is not symmetric, i.e., $D_KL(p||q) \neq D_KL(q||p)$. Thus, the Kullback-Leibler divergence is used [5]:

$$Div(p||q) = D_{KL}(p||q) + D_{KL}(q||p). \tag{2}$$

Analogously, a symmetric version of mutual information can be introduced

$$J(x,y) = I(x,y) + I^*(x,y)$$
$$= \int\int f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dxdy + \int\int f(x)f(y) \log \frac{f(x)f(y)}{f(x,y)}\, dxdy\,.$$

Our idea is to repeat Linfoot's derivation of formula (1), replacing the mutual information $I(X,Y)$ with its symmetric version $J(X,Y)$. In this case, the following result holds.

Theorem 1. *A modified symmetric analog of the Linfoot's informational correlation coefficient (1) called as the modified informational correlation coefficient (MICC) is given by:*

$$\rho_{MICC} = \sqrt{1 - \frac{2}{W(2e^{2(J+1)})}}, \tag{3}$$

where $W(x)$ is the Lambert function [1], inverse w.r.t. xe^x. In particular, $\rho_{MICC} = \rho$ at the standard bivariate normal distribution.

The results of comparison of these two correlation measures are exhibited in Fig. 2– Fig. 4: from them it follows that $MICC$ outperforms ICC on all examined combinations of sample sizes and correlation coefficients. The observed improvement is more considerable on small samples and low values of the correlation coefficient — just in the most difficult cases for *ICC*.

5 Conclusions

1. The statistical performance of the Linfoot's informational correlation coefficient is studied: considerable biases of estimation are observed.
2. To reduce the biases of *ICC*, a modified symmetric version of it, namely *MICC*, is proposed, which proved to provide much lesser estimation biases as compared to the original one.
3. The proposed modified informational correlation coefficient *MICC* is recommended for processing Big Data.

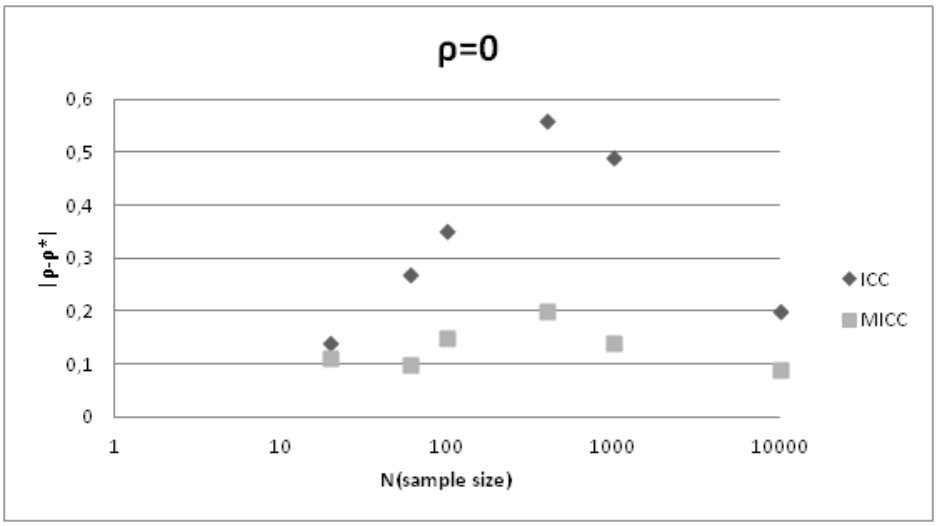


Figure 2: Monte-Carlo Biases of *ICC* and *MICC* at $\rho = 0$

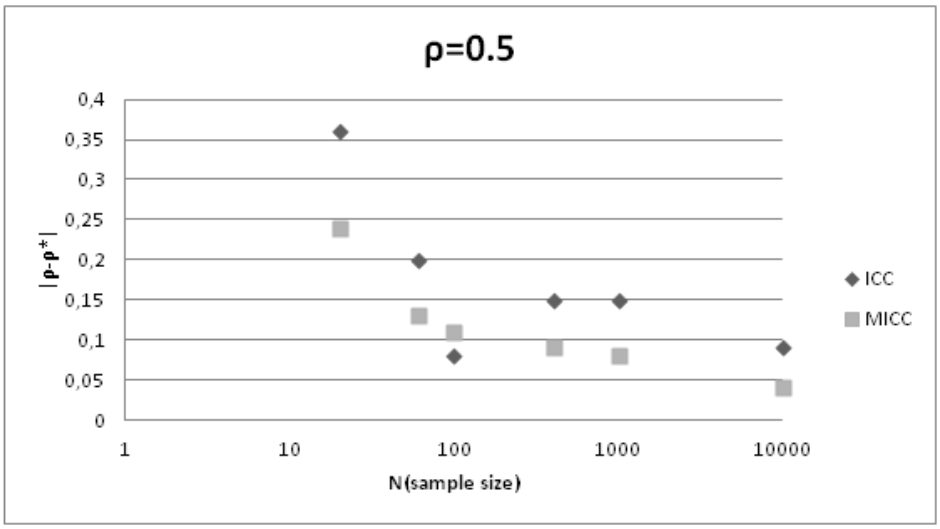


Figure 3: Monte-Carlo Biases of *ICC* and *MICC* at $\rho = 0.5$

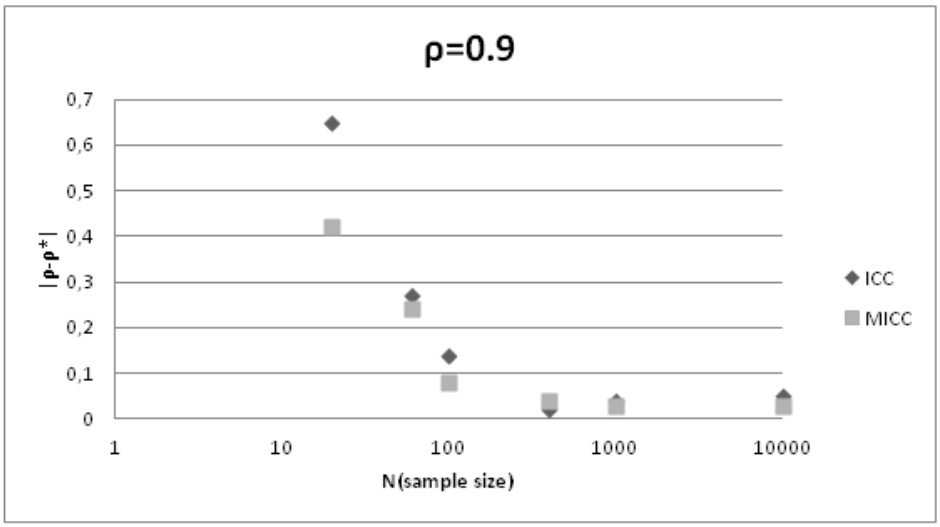


Figure 4: Monte-Carlo Biases of *ICC* and *MICC* at $\rho = 0.9$

References

[1] Corless R., Gonnet G., Hare D., Jeffrey D., Knuth, Donald. (1996). On the Lambert W Function. *Advances in Computational Mathematics*. (Berlin, New York: Springer-Verlag). Vol. **5**, pp. 329-359.

[2] Granger C.W., Maasoumi E., Racine J. (2004). A Dependence Metric for Possibly Nonlinear Processes. *Journal of Time Series Analysis*. Vol. **25**, pp. 649-669.

[3] Hausser J., Strimmer K. (2010). Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. *Journal of Machine Learning Research*. Vol. **10**, pp. 1469-1484.

[4] Joe H. (1989). Relative Entropy Measures of Multivariate Dependence. *Journal of the American Statistical Association*. Vol. **84**, pp. 157-164.

[5] Kullback S. (1959). *Information Theory and Statistics*. John Wiley.

[6] Linfoot E.H. (1957). An Informational Measure of Correlation. *Information and Control*. Vol. **1**, pp. 85-89.

[7] Shannon C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*. Vol. **27**, pp. 379-423, 623-656.

[8] Silvey S.D. (1964). A Measure of Association. *The Annals of Mathematical Statistics*. Vol. **35**, pp. 1157-1166.

AUTOMATED CLASSIFICATION OF ARTERIES AND VEINS IN THE RETINAL BLOOD VASCULATURE

G. Stabingis[1], J. Bernatavičienė[2], G. Dzemyda[3],
D. Imbrasienė[4], A. Paunksnis[5]
[1,2,3]*Institute of Mathematics and Informatics, Vilnius University*
[4,5]*Hospital of Lithuanian University of Health Sciences Kauno Klinikos*
[4]*Lithuanian Sports University*
[1,2,3]*Vilnius and* [4,5]*Kaunas, LITHUANIA*
e-mail: [2]Jolita.Bernataviciene@mii.vu.lt

Abstract

The paper deals with the automated method for early detection of eye fundus pathological changes in the main vessels, which is useful for retinal disease diagnosing. The target users are primary care physicians and optometrists that use the digital portable eye fundus imaging devices. In the experiments, the digital portable eye fundus camera is used.

1 Introduction

Retinal vessel segmentation algorithms are a major component of computer-aided retinal disease diagnosis systems [1]. Ophthalmologists have ascribed subtle changes in the diameter of arteries and veins to several diseases such as diabetic retinopathy, atherosclerosis, hypertension, choroidal neovascularization, and so on [2]. Diseases affect arteries and veins differently leading to an abnormal artery to the vein width ratio in cardio vasculature. Typical arteries and veins can be discriminated by color, illumination, width, central reflex size and topological properties, however, these properties are actual for classifying vessels only near the optic disc and depend on the quality of images and the region at eye fundus [3]. The structural characteristics of vessels could be used as rules: arteries never cross arteries and the same is true for veins. Thus, at any any crossover points, where two vessels cross each other, one of them is artery and the other is vein; the retinal blood vasculature follows the structure of a binary tree. Therefore, a vessel branches into two vessels of the same type, which means that at every bifurcation point, three vessel segments connected to each other, are from the same class of vessels [4]. Vessels in the outer regions of the image are often darker, the width of vessels change is least, branches of the same vessel lie next to each other. Also biological characteristics, e.g., changes in color of retina, are individual for each person and influence the classification results. The aim of this research is to develop a fully automated method for early detection of eye fundus pathological changes in the main vessels. It will serve as a helpful tool for diagnosing and will be used by primary care physicians and optometrists. The created method is suitable for the analysis of image, obtained by digital hand-held eye fundus imaging devices, available on the market.

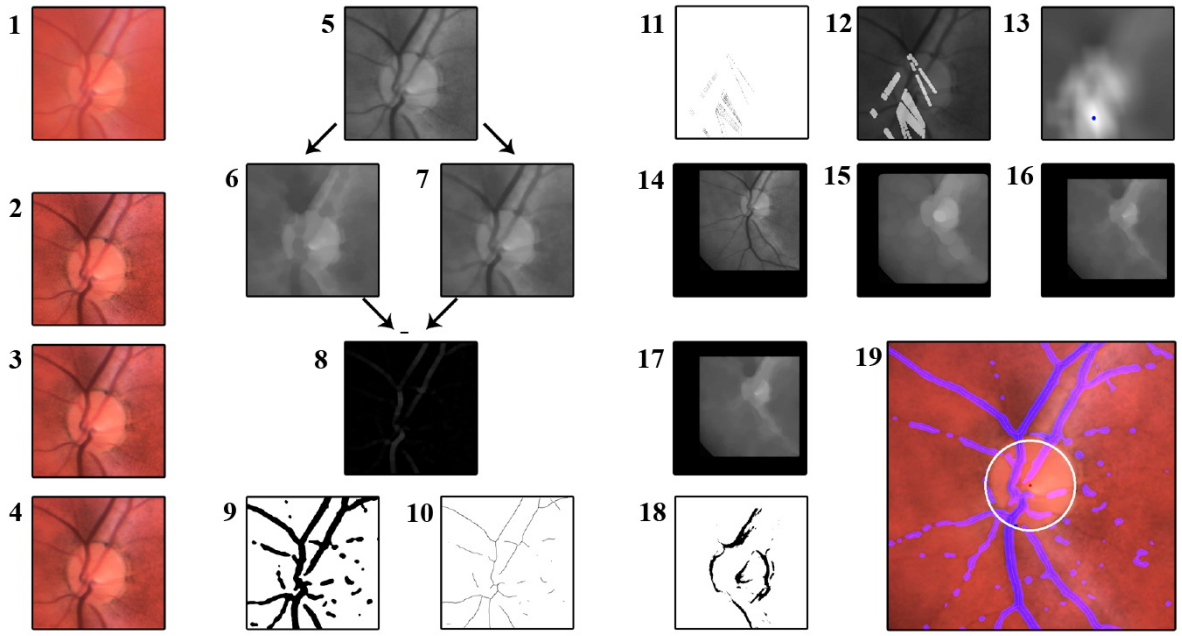


Figure 1: The main steps of the algorithm: 1 - initial image; 2-4 - preprocessing; 5 - gray color image; 6-10 - vessel extraction; 11-19 - optical disc detection.

2 Proposed Method

An automated method for segmentation and classification of arteries and veins as well as measurement of arteriolar-to-venular diameter ratio (AVR) in eye fundus images are presented. The method includes optic disc segmentation in order to determine the arteriolar-to-venular diameter ratio in the measurement zone, as well as retinal vessel segmentation, vessel classification into arteries and veins, selection of major vessel pairs, and measurement of AVRs.

As a background, we have taken the idea from two articles ([5] and [6]) that presents a high efficiency of detecting the eye fundus elements, mostly using mathematical morphological operations. All the steps of the algorithm are shown in Fig 1.

After the eye fundus image V_{init} is taken (Fig. 1.1), preprocessing of the image is performed. CLAHE (contrast limited adaptive histogram equalization) method is applied (Fig. 1.2). A median filter of size $s = 5$ (Fig. 1.3) and the Gaussian filter (Fig. 1.4) of size $s = 7$ with $\sigma = 1$ are applied. Then the preprocessed image V_{pr} is used in other calculations. The green color channel V_G is used for a further analysis (Fig. 1.5). The proportion coefficient $p = width(V_{pr})/576$ is calculated according to the width of the fundus image. This coefficient is used for the adjustment of parameters given in ([5] and [6]) in order to use full size images. The most evident blood vessels are extracted. The mathematical morphological closing operations of sizes $s = 8 \cdot p$ (Fig. 1.6) and $s = 4 \cdot p$ (Fig. 1.7) are performed on V_G. The obtained images are subtracted and the image V_{vs} is obtained (Fig. 1.8). The image V_{vs} is thresholded from $(max(V_{vs}) \cdot 0.1)$ to 255. The obtained binary image is processed with a median filter of size $s = 11$ in

order to remove smaller isolated parts and to make the image more smooth. The main blood vessels V_{vm} are obtained by this processing (Fig. 1.11). Finally, blood vessels have thinned and the resulting noise is removed by removing isolated pixels. The main thinned blood vessel net V_{vt} is obtained (Fig. 1.10).

Detection of the preliminary optic nerve disc OD_p center. V_{vt} is used for line detection using the Hough transformation. Lines are selected with more then $20 \cdot p$ votes and with slopes more then 45°. According to these lines with the slope difference larger then 1° the intersection map is created (Fig. 1.11). The line intersection map is dilated by the kernel of size $s = 5 \cdot p$ and proportionally added to the green channel V_G (Fig. 1.12). Then the resulting image is blurred using Gaussian filter of size $s = 30 \cdot p$ (Fig. 1.13). The most intense point is selected as the preliminary optic nerve disc center OD_p.

Optic nerve disc detection OD_r. A square $250 \cdot p$ wide is masked out (Fig. 1.14) at the detected preliminary optic nerve disc OD_p center. A morphological closing operation of the kernel of size $s = 25 \cdot p$ is applied to eliminate blood vessels (Fig. 1.15 - 1.16). The obtained image is blurred using the Gaussian filter with the kernel of size $s = 4 \cdot p$ (Fig. 1.17). A gradient image is received using the Sobel filter and thresholded from 7 to 255 (Fig. 1.18). The circle extraction by the Hough transformation is applied to the gradient image and the best circle with the center closest to the detected preliminary optic disc OD_p center is chosen as the real optic nerve disc OD_r (Fig. 1.19).

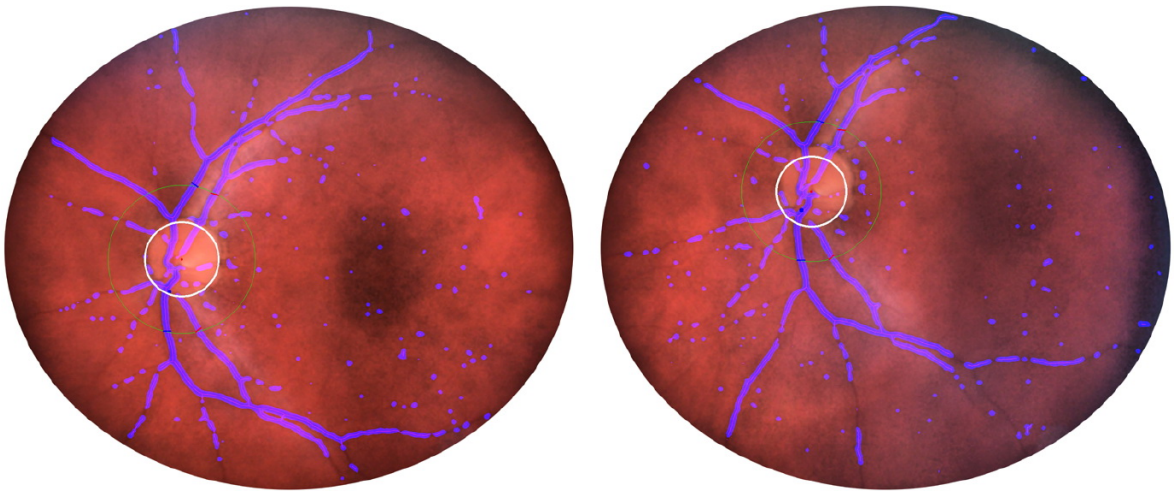


Figure 2: Results of algorithm performance. Vessels detected before (left) and after (right) the sport load. The white circle marks the detected optic nerve disc. The red line denotes the detected artery and the blue line denotes the detected vein.

Blood vessel measurements are performed at the distance of a double optic disc size from the center of OD_r. The algorithm measures in main vessels V_{vm} and finds intersections. Then it measures the minimum width of the vessels between $4 \cdot p$ and $30 \cdot p$ at these points. Next two largest top vessels and two largest bottom vessels are selected.

3 Data for Experiments

The data of 11 different men who do not practice sports actively were used. They play soccer twice a week (age=44.9±6.6 years; height=178.2±6.7 cm; body weight=85±11.4 kg; BMI=26.8 ± 3.8 kg/m^2). The eye fundus was observed using a digital portable eye fundus camera Smartscope M5-2 EY3 (Optomed OY). Taking pictures of the central part of retina with a narrow pupil, using a digital format of high resolution allowed us to measure the changes in retinal vessels. The automatically selected main vessels are classified as vein which mean intensity along the detected diameter is smaller. The other are classified as arteries (Fig. 2). The clearing of upper and lower branches next to eye nerve disc of retinal central artery and central vein was measured using the proposed algorithm.

4 Conclusions

The experiments show that preprocessing must be enhanced in order to remove the gradient noise, which is caused by extra light and causes some places of fundus image to become brighter. This fact leads to missing detection of the optical disc as well as classification of veins and arteries. A simple removal of the gradient can effect the corruption of the optical disc, because it is lighter then the surroundings. The algorithm of blood vessel crack filling algorithm must be integrated. It is especially important on the edge of the optic nerve disc. Overlying and side-by-side parallel blood vessels should be detected.

References

[1] Sim D.A. et al. (2015). Automated retinal image analysis for diabetic retinopathy in telemedicine.*Curr Diab Rep* 15:14.

[2] Kanski J.J. (2007). Clinical Ophthalmology, 6th ed., Elsevier Health Sciences, London, UK, p. 367.

[3] Kondermann C., Kondermann D., Yan Y. (2007). Blood vessel classification into arteries and veins in retinal images.*Proc. SPIE*, 6512.

[4] Mirsharifa Q., Tajeripoura F., Pourreza H. (2013) Automated characterization of blood vessels as arteries and veins in retinal images.*Comput. Med. Imaging Graphics*. Vol. **37**, pp. 607-617.

[5] Pachiyappan A., Das U., Murthy T., Tatavarti R. (2012). Automated diagnosis of diabetic retinopathy and glaucoma using fundus and OCT images. Lipids in Health and Disease. pp. 11-73.

[6] Ravishankar S., Jain A., Mittal A. (2009). Automated feature extraction for early detection of diabetic retinopathy in fundus images. Computer Vision and Pattern Recognition (CVPR 2009. IEEE Conference).

COMPARISON OF PARTIALLY RANKED LISTS

E. Stoimenova
Institute of Mathematics and Informatics
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Sofia, BULGARIA
e-mail: jeni@math.bas.bg

Abstract

In this paper we introduce a measure of closeness of partial rankings based on a metric on permutations, and we analyze some of its properties.

1 Introduction

In many situations, there are different methods for analyzing the same data. For example, several methods exist for finding differentially expressed genes using RNA-seq data. They tend to produce similar, but not identical significant genes and rankings of the gene list. When comparing different methods applied to the same data, we are interested in how close are their outputs. The main idea is to define appropriate distance of the sample space. Further, the interpretation of the rough distance between two rankings should be made on the basis of its statistical significance. That means we need to know the distribution of the distance under some common hypotheses about a sample of rankings. In recent years, many new applications appear in different areas including bioinformatics pattern recognition, information retrivial [7], [6], [1], [4], [5], etc.

In this paper we define an appropriate mathematical framework that include special cases of partially ranked lists of genes. Any ranked list can be complete, which means all n genes are ranked, or incomplete, which means some genes are not ranked. The incomplete ranking include the case where the most significant k genes are ranked, with group $k+1$ consisting of the remaining genes. Any ranking of n items corresponds a permutation $\langle \alpha(1), \ldots, \alpha(n) \rangle$ from the set of all permutations S_n. We define appropriate distance measures on S_n in order to compare full or incomplete rankings or rankings of different types. The distance can be thought of as a measure of the similarity of the two rankings.

Let α and β be two permutations from S_n corresponding to two rankings and let d be a metric on the permutation group S_n. Then $d : S_n \times S_n \to [0, \infty)$ satisfies the usual axioms: $d(\alpha, \beta) \geq 0 \quad \forall \alpha, \beta \in S_n$, $d(\alpha, \beta) = 0 \Leftrightarrow \alpha = \beta$; $d(\alpha, \beta) = d(\beta, \alpha) \quad \forall \alpha, \beta \in S_n$; and the triangle inequality $d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\gamma, \beta) \quad \forall \alpha, \beta, \gamma \in S_n$.

Invariance is natural in many problems. Right-invariance means that the distance does not depend on arbitrary labeling or reordering of the data:

$$d(\alpha, \beta) = d(\alpha\tau, \beta\tau).$$

Here $\alpha\tau$ is the product of two permutations α and τ and defined by $\alpha\tau(i) = \alpha(\tau(i))$. Right-invariant property allows to compute the distance between two permutations α and β through the the distance of $\alpha\beta^{-1}$ to the identity permutation.

Further in our analysis we are using a popular statistical measure of similarity on S_n called Spearman's ρ. For $\alpha, \beta \in S_n$ it is defined by

$$R^2(\alpha, \beta) = \sum_{i=1}^{n} (\alpha(i) - \beta(i))^2.$$

Strictly speaking, Spearman's ρ is not a metric in the above definition, however, its square root is the Euclidean metric on permutations. It is easy to see that Spearman's ρ is right-invariant. By right-invariance of a distance it is sufficient to study its statistical properties when one of the rankings is the identity permutation.

2 Complete or incomplete ranking

A ranking of n items is represented by an ordered n-tuple, which simply lists the items in their ranked order. The most preferred item is listed first, and the least preferred item appears in the n-th position. Any ranking corresponds to a permutation which is an element of the set S_n of permutations. Given a set of rankings, the problem of their comparison reduced to a problem of choosing appropriate measure of association on the set of all rankings. There are several usefull distance measures on S_n thoroughly discussed in statistical literature like Kendall's τ, Spearman's ρ, Spearman's footrule. Therefore, for two permutations $\alpha, \beta \in S_n$ the distance $d(\alpha, \beta)$ can be thought of as a measure of similarity of the two rankings. Excellent references on statistical analysis of rankings are the monographs by Diaconis [3], Critchlow [2], and Marden [8].

[**Classification into r ordered categories.**] Suppose the list of genes is splitted into several groups, so that there is a ranking between the groups and not necessarily within each group. It can be describe formally following Critchlow [2].

Let $n_1, \ldots, n_r$ be an ordered sequence of r strictly positive numbers summing to n. Such an ordered partition corresponds to a partial ranking with n_1 items in the first group, n_2 items in the second group and so on. No further information is conveyed about orderings within each group. The special case of ranking the top k items corresponds to $n_1 = \cdots = n_k = 1,\ n_{k+1} = k + 1$.

Formally, denote $N_1, \ldots, N_r$ are the following partition of $\{1, \ldots, n\}$:

$$\begin{aligned} N_1 &= \{1, \ldots, n_1\} \\ N_2 &= \{n_1 + 1, \ldots, n_1 + n_2\} \\ &\cdots \\ N_r &= \{n_1 + \cdots + n_{r-1} + 1, \ldots, n\}. \end{aligned} \tag{1}$$

Let S denote the subgroup of all rankings which permute the first n_1 items among the first n_1 ranks, and which permute the next n_2 items among the next n_2 ranks, and

so on. The equivalence class $[\alpha]$, that assigns the same set of ranks to the items from the each category as α, is the right coset $S\alpha$. There is a one-to-one correspondence between the partitioning "of type $n_1, \ldots, n_r$" and the right cosets of S.

2.1 Distances on partial rankings

In the above algebraic structure the problem of comparing of partial rankings is reduced to a problem of extending the metrics on the permutation group S_n to metrics on the corresponding coset space. We discuss an extension of the above metrics for the cases of partial rankings. One natural way of extending it is to construct the induced Hausdorff metrics. Its particular benefit is that it keeps the metric properties of the original distance.

Let the two partial rankings be of types $n_1, \ldots, n_r$. Denote n_{ij} the number of elements in the set $\{\alpha^{-1}(N_i) \cap \beta^{-1}(N_j)\}$. Then the function

$$R_{fv}(\alpha, \beta) = \sum_{i=1}^{r} \sum_{j=1}^{r} |c_i - c_j|^2 n_{ij}.$$

is a right-invariant metric on partial rankings induced by Spearman's ρ. Here $c_i = n_1 + \cdots + n_{i-1} + \frac{n_i+1}{2}$ is the average of the n_i numbers in the set N_i defined by (1).

The interpretation of this function is that it computes Spearman's ρ distance between the two rankings using the "pseudo-ranks" c_i and c_j instead the ordinary ranks to those items in $\{\alpha^{-1}(N_i) \cap \beta^{-1}(N_j)\}$. The function is called the "fixed vector" metric on S_n/S induced by Spearman's ρ. Its main advantage is that it preserves the distance properties and the right invariance as well [9]. Additionally, some useful statistical properties are known in the literature.

2.2 Comparing partial rankings of different types

We consider the most general case of comparing partial rankings of different types. Let the two partial rankings be of types $n_1, \ldots, n_r$ and $n'_1, \ldots, n'_{r'}$ respectively.

Let $N_1, \ldots, N_r$ be as defined in (1) and let $N'_1, \ldots, N'_r$ be a second partition of $\{1, \ldots, n\}$:

$$\begin{aligned} N'_1 &= \{1, \ldots, n'_1\} \\ N'_2 &= \{n'_1 + 1, \ldots, n'_1 + n'_2\} \\ &\cdot \quad \cdot \quad \cdot \\ N'_{r'} &= \{n'_1 + \cdots + n_{r'-1} + 1, \ldots, n\}. \end{aligned}$$

Let n_{ij} be the number of elements in the set $\{\alpha^{-1}(N_i) \cap \beta^{-1}(N'_j)\}$. Then

$$R_*(\alpha, \beta) = \sum_{i=1}^{r} \sum_{j=1}^{r'} |c_i - c'_j|^2 n_{ij}.$$

is a right-invariant metric on partial rankings. Here $c'_j = n'_1 + \cdots + n'_{j-1} + \frac{n'_j+1}{2}$ is the average of the n'_j numbers in the set N'_j defined by (2).

3 Large sample approximation of a distance distribution

Now, we estimate the mean and the variance of R^{2*} and find approximations of its distribution.

Definition 1. The metric d^* on S_n/S is asymptotically normally distributed if for partial rankings α^* and β^* the following limit distribution is valid

$$\lim_{n\to\infty} P\left(\frac{d^*(\alpha^*,\beta^*) - E\ d^*(\alpha^*,\beta^*)}{\sqrt{var(d^*(\alpha^*,\beta^*))}} \le x\right) = \Phi(x)$$

for all real numbers x, where Φ, is the standard normal cumulative distribution function.

The significance of the distance is useful to estimate the similarity between the two partial rankings. For this one needs to calculate the probability that d^* is less than or equal to the observed value $d^*(\alpha^*,\beta^*)$. This probability is the p-value for α^* and β^*. Smaller values of p indicate stronger evidence that α^* and β^* are "similar". To compute the p-value, Critchlow [2] finds the probability distribution of some popular metrics on permutations under the appropriate uniformity assumption.

The mean and the variance of R^{2*} are given by [2]:

$$E(R^{2*}) = \sum_{i=1}^{r}\sum_{j=1}^{r} |c_i - c_j|^2 \frac{n_i n_j}{n}$$

$$var(R^{2*}) = \frac{1}{n^2(n-1)} \sum_{i=1}^{r}\sum_{j=1}^{r}\sum_{k=1}^{r}\sum_{l=1}^{r} n_i n_j n_k n_l (|c_i - c_j|^2 + |c_k - c_l|^2 - 2|c_k - c_j|^2),$$

where $c_i = n_1 + \cdots + n_{i-1} + \frac{n_i+1}{2}$ is the average of the n_i numbers in the set $N_i = \{n_1 + \cdots + n_{i-1} + 1, \ldots, n_1 + \cdots + n_{i-1} + n_i\}$.

For equal partition sizes these reduce to

$$E(R^{2*}) = n^3 \frac{(r^2-1)}{6r^2}$$

$$var(R^{2*}) = \frac{n^6}{n-1} \frac{(r^2-1)^2}{6r^5}.$$

Normal approximation is valid for the distance distribution under the assumption that they were generated, independently, from a uniform distribution on all possible partial rankings. For equal partition sizes Gamma distribution with shape parameter $= \mu^2/\sigma^2 = n-1$ gives better approximation.

Example (Palejev & Stoimenova [10]). A simulation study is based on one million repetitions of gene sequences of size 13932. Each of them contains data for the significance of gene expression. Further, the genes are splitted into six groups by values

according the size of the p-values. Intervals reasonable for application are determined by $0, 10^{-4}, 10^{-3}, 10^{-2}, 0.05, 10^{-1}, 1$. For this unbalance case the distances between any two of the partial rankings are calculated. The distributions of the distances is shown on Figure 1. Gamma distribution approximation is also suitable for this case.

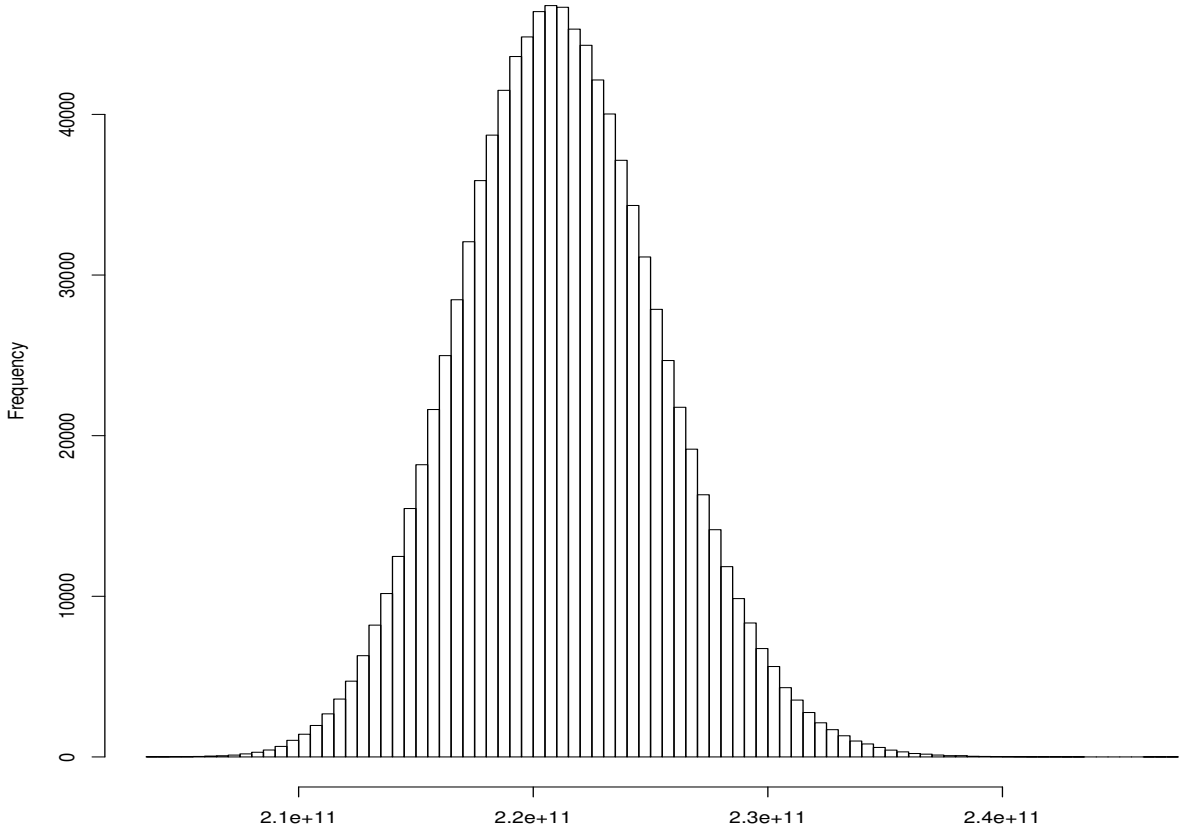


Figure 1: Distribution of distances between 2 random permutations

Acknowledgments. The author acknowledge funding by the Bulgarian fund for scientific investigations Project I02/19.

References

[1] Chan, C. H., Yan, F., Kittler, J., and Mikolajczyk, K. (2015). Full ranking as local descriptor for visual recognition: A comparison of distance metrics on S_n. *Pattern Recognition*, 48(4):134–160.

[2] Critchlow, D. E. (1985). *Metric methods for analyzing partially ranked data.* Lecture Notes in Statistics, 34. Berlin etc.: Springer-Verlag.

[3] Diaconis, P. (1988). *Group representations in probability and statistics.* IMS Lecture Notes-Monograph Series, 11. Hayward, CA: Institute of Mathematical Statistics.

[4] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. (2006). Comparing partial rankings. *SIAM J. Discrete Math.*, 20(3):628–648.

[5] Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. *SIAM J. Discrete Math.*, 17(1):134–160.

[6] Jurman G.,Riccadonna S. (2009). Canberra distance on ranked lists. In: *Proceedings of Advances in Ranking NIPS 09 Workshop*, pages 22–27.

[7] Jurman, G., Merler, S., Barla, A., Paoli, S., Galea, A., and Furlanello, C. (2007). Algebraic stability indicators for ranked lists in molecular profiling . *Bioinformatics*, 24(2):258–264.

[8] Marden, J. I. (1995). *Analyzing and modeling rank data.* Monographs on Statistics and Applied Probability. 64. London: Chapman.

[9] Rukhin, A. L. (1970). Certain statistical and probability problems on groups. 111:52–109.

[10] Palejev, D., Stoimenova, E. Comparison of incomplete ranked lists with application to RNA-seq differential expression methods. Working paper.

DATA DEPTH AND ITS APPLICATIONS IN CLASSIFICATION

Ondrej Vencalek
Department of Mathematical Analysis and Applications of Mathematics
Palacky University
Olomouc, CZECH REPUBLIC
e-mail: ondrej.vencalek@upol.cz

Abstract

Concept of data depth provides one possible approach to the analysis of multivariate data. Among other it can be also used for classification purposes. The present paper summarizes main ideas how the depth can be used in classification. Two step scheme of the typical depth-based classifiers is discussed. Different choices of the depth function and classification procedure used in this scheme lead to different classifiers with different properties. Different distributional assumptions might lead to the preference of either global or local depth functions and classification procedures.

1 Introduction

Depth function is basically any function which provides ordering (or quasi-ordering) of points in multidimensional space. Existence of ordering enables generalization of many nonparametric techniques proposed for univariate variables and thus the data depth creates one possible basis of nonparametric multivariate data analysis.

Data depth has been also applied in classification. The aim of classification is to create a rule for allocation of new observations into one of two (or more) groups. Several such rules, known as classifiers, based on data depth were proposed since 2000. The aim of the present paper is to summarize main ideas how the depth can be used in classification and to present new trends in this area.

2 Concept of data depth

There are several depth functions commonly used in applications – halfspace depth, projection depth, spatial depth, Mahalanobis depth, zonoid depth, simplicial depth and some others. Their survey can be found in [8]. We recall here the first three of the depth functions listed above since they are used most frequently in context of classification:

- The *halfspace depth* of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P is defined as the minimum probability mass carried by any closed halfspace containing $\boldsymbol{x}$, that is

$$D(\boldsymbol{x}, P) = \inf_{\mathbb{H}} \left\{ \mathrm{P}(\mathbb{H}) : \mathbb{H} \text{ a closed halfspace in } \mathbb{R}^d : \boldsymbol{x} \in \mathbb{H} \right\}.$$

- The projection depth of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P is defined as

$$D(\boldsymbol{x}, P) = \frac{1}{1 + O(\boldsymbol{x}, P)}, \quad \text{where} \quad O(\boldsymbol{x}, P) = \sup_{\|\boldsymbol{u}\|=1} \frac{|\boldsymbol{u}^T \boldsymbol{x} - \mu_{P_{\boldsymbol{u}}}|}{\sigma_{P_{\boldsymbol{u}}}},$$

  where $\mu_{P_{\boldsymbol{u}}}$ is some location and σ_{P_u} is some scale measure of distribution of random vector $\boldsymbol{u}^T \boldsymbol{X}$, usually the median and MAD.

- The spatial depth (also called L_1-depth) of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P with variance matrix $\boldsymbol{\Sigma}$ is defined as

$$D(\boldsymbol{x}, P) = 1 - \mathrm{E} \left\| \frac{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{X})}{\left\| \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{X}) \right\|} \right\|.$$

All three depth functions listed above have desirable properties like affine invariance, maximality at a point of symmetry (if the distribution is symmetric in some sense, e.g. angularly), monotonicity on rays from the point with the maximal depth – so called deepest point, which can be considered as multivariate analogy to median. See [10] for the discussion of desirable properties of the depth functions.

The very important difference among the considered depth functions consists in different behaviour of their empirical versions. While the empirical halfspace depth is equal to zero for any point which lies outside of the convex hull of the data, the empirical projection depth (as well as empirical spatial depth) are nonzero everywhere.

The depth of a point is a characteristic of the point specifying its centrality or outlyingness with respect to the considered distribution. Since the whole distribution is considered, the depth is said to be a "global" characteristic of the point. However, in recent years there have been attempts to "localize" depth, see [7] or [3]. Later we will discuss importance of these attempts for classification purposes.

3 Maximal depth classifier

Let us first shortly recall the two-classes classification problem. We consider two unknown absolutely continuous probability distributions P_1 and P_2 on $\mathbb{R}^d$. Independent random samples from these distributions are available. Together they constitute so called training set. Empirical distributions based on the training set are denoted $\widehat{P}_1$ and $\widehat{P}_2$. The class to which the observation $\boldsymbol{x}$ is assigned is denoted by $d(\boldsymbol{x})$.

The maximal depth classifier – the very first depth-based classifier – is simple:

$$d(\boldsymbol{x}) = \arg\max_{i=1,2} D(\boldsymbol{x}; \widehat{P}_i) \tag{1}$$

The points close to the centre (multivariate median) of some distribution have high depth with respect to this distribution and it seems to be natural to classify them to

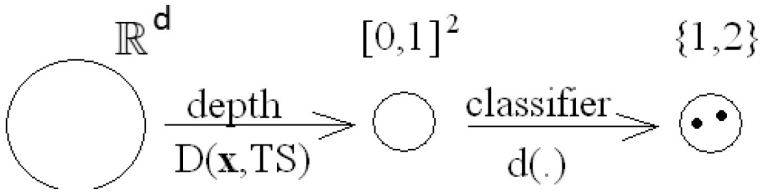


Figure 1: Scheme of typical depth-based classifier.

this distribution. The idea of the maximal depth classifier is thus in accordance to common sense.

However Ghosh and Chaudhuri [2] proved that the maximal depth attains minimal possible probability of misclassification (known as Bayes risk) only in very special cases – they showed optimality when the considered distributions are elliptically symmetric with the density decreasing from the centre, differing only in location and having equal prior probabilities. The optimality is lost even if only one of these assumptions is not fulfilled.

4 Advanced depth-based classifiers

The paper by Ghosh and Chaudhuri [2] started the search for depth-based classifiers which would be applicable in broader class of distributional settings. Typical depth-based classifier can be described as a two-steps procedure:

1. The first step consists in computation of depths of the new observation $\boldsymbol{x}$ with respect to both parts of the training set. Each point is characterized by a pair of depths, these pairs lies in so called DD-space (depth-versus-depth space). Typically the DD-space is subset of $[0,1] \times [0,1] \subset \mathbb{R}^2$ and thus the first step can be usually considered as reduction of dimensionality – from $\mathbb{R}^d$ to the compact subset of $\mathbb{R}^2$. This step is connected with the question "Which depth function should be used?"

2. The second step consists in application of some classification procedure in the DD-space. This step is connected with the question "Which classification procedure should be applied in the DD-space?"

The scheme of typical depth-based classification procedure is shown in Figure 1.

The difference among the classifiers proposed in literature consists mainly in different answers to the two questions connected with the two steps – different depth functions can be applied in the first step and different classification procedures can be applied in the second step.

4.1 Which depth function should be used

The classifiers which use any global depth function perform well only if the considered distributions have some global properties like symmetry or unimodality. In more gen-

eral settings, for example in the case of multimodality or nonconvexity of levelsets of density, some local depth should be used - see e.g. [1] or [4].

4.2 Which classification procedure should be applied in the DD-space

Procedures that are applied in the DD-space can be also either "global" or "local" in nature. As the global we denote the procedures that take into account all points of the training set when constructing the classifier while the local procedures take into account only points of the training set close to the point which is classified. Let us mention at least some of the both global and local procedures:

- The *DD-classifier* proposed by Li et al. [6] belongs to the global depth-based classifiers. The idea of the classifier is to separate the groups in DD-space by line (or more generally by a polynom) to minimize number of errors when classifying points from the training set.

- The *DD-alpha* procedure proposed by Lange et al. [5] is another global depth-based classifier. Instead of the pair $[D(\boldsymbol{x}, \widehat{P}_1), D(\boldsymbol{x}, \widehat{P}_2)]$, it works with a vector

$$\boldsymbol{z} := [D(\boldsymbol{x}, \widehat{P}_1), D(\boldsymbol{x}, \widehat{P}_2), D(\boldsymbol{x}, \widehat{P}_1) \cdot D(\boldsymbol{x}, \widehat{P}_2), D(\boldsymbol{x}, \widehat{P}_1)^2, D(\boldsymbol{x}, \widehat{P}_2)^2].$$

  Lange et al. proposed a heuristic for finding proper parameters which specify separating hyperplane given by the equation $aD(\boldsymbol{x}, \widehat{P}_1) + bD(\boldsymbol{x}, \widehat{P}_2) + cD(\boldsymbol{x}, \widehat{P}_1)D(\boldsymbol{x}, \widehat{P}_2) + dD(\boldsymbol{x}, \widehat{P}_1)^2 + eD(\boldsymbol{x}, \widehat{P}_2)^2 = 0$. The procedure was successfully tested on many real datasets leading usually to low misclassification rates. The procedure was implemented in the R-package `ddalpha`.

- The *classifier which uses kernel density estimation* proposed by Ghosh and Chaudhuri [2] is a local depth-based classifier. In the case of elliptically symmetric distribution the classifier which minimizes probability of misclassification can be expressed as

$$d(\boldsymbol{x}) = \arg\max_{i=1,2} \pi_i \theta_i(D(\boldsymbol{x}, \widehat{P}_i)),$$

  where $\theta_i, i = 1, 2$ are some unknown real functions that can be estimated by kernel density estimation. This procedure can be used even if the distributions are nonelliptical.

- The *k depth-nearest neighbour classifier* used by Vencalek [9] is another local depth-based procedure. The idea is quite simple – to use the well known k-nearest neighbour procedure in the DD-space. The question is which metric should be used to measure distances between distinct points.

5 Conclusion

The data depth provides basis for nonparametric inference on multidimensional data. It can be also applied in classification. Although one can expect broad applicability of

the nonparametric depth-based classifiers their optimality can be guaranteed usually only under some restrictive assumptions. Global depth functions and global classification techniques applied on the DD-space lead to good results only if the considered distributions have some global properties like unimodality. In more general settings localization is needed – one can use local depth functions or local classifiers used in the DD-space.

Acknowledgement

The research was supported by the grant of Czech Science Foundation GA15-06991S.

References

[1] Dutta, S., Chaudhuri, P., Ghosh, A. K. (2012). Classification using Localized Spatial Depth with Multiple Localization. *Communicated for publication.*

[2] Ghosh, A. K., Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics.* Vol. **32**, pp. 327-350.

[3] Hlubinka, D., Kotik, L., Vencalek, O. (2010). Weighted data depth. *Kybernetika*, Vol. **46**, No. 1, pp. 125-148.

[4] Hlubinka, D., Venclek, O. (2013). Depth-Based Classification for Distributions with Nonconvex Support. *Journal of Probability and Statistics.* 2013.

[5] Lange, T., Mosler, K., Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers.* Vol. **55**, No. 1, pp. 49-69.

[6] Li, J., Cuesta-Albertos, J. A., Liu, R. (2012). DD-classifier: nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association.* Vol. **107**, No. 498, pp. 737-753.

[7] Paindaveine, D., Van Bever, G. (2013). From depth to local depth: a focus on centrality. *Journal of the American Statistical Association.* Vol. **108**, No. 503, pp. 1105-1119.

[8] Vencalek, O. (2011). Weighted data depth and depth based classification. *PhD thesis.* URL: `http://artax.karlin.mff.cuni.cz/~venco2am/DataDepth.html`

[9] Vencalek, O. (2014). New depth-based modification of the k-nearest neighbour method. *SOP Transactions on Statistics and Analysis*, 2014.

[10] Zuo, Y., Serfling, R. (2000). General notion of statistical depth function. *Annals of Statistics.* Vol. **28**, pp. 461-482.

OPTIMAL CHOICE OF ORDER STATISTICS UNDER CONFIDENCE REGION ESTIMATION IN CASE OF LARGE SAMPLES

A. Zaigrajew[1], M. Alama-Bućko[2]
[1] *Nicolaus Copernicus University of Toruń*
[2] *University of Technology and Life Sciences of Bydgoszcz*
[1] *Toruń and* [2] *Bydgoszcz, POLAND*
e-mail: [1]alzaig@mat.umk.pl

Abstract

Let $x = (x_1, \ldots, x_n)$ be a sample from a distribution $P_\theta, \theta = (\theta_1, \theta_2)$, where $\theta_1 \in R$ is a location parameter and $\theta_2 > 0$ is a scale parameter. To estimate θ strong two-dimensional confidence regions of given confidence level $\alpha \in (0, 1)$ are considered. The quality of a Borel confidence set $B(x)$ is characterized by the risk function defined as $R(\theta, B) = \mathrm{E}_\theta \lambda_2(B(x))$, where $\lambda_2(B(x))$ is the Lebesgue measure of $B(x)$. Among confidence regions we distinguish those having the minimal risk and call them optimal. The method for construction of an optimal confidence region is well-known (see, e.g., [1]) and is based on using a pivot. Let $x_{i:n}$ represents the ith order statistic of the sample x for $i = 1, \ldots, n$. To construct a pivot two statistics t_1 and t_2 are taken; both statistics depend on given $k \leq n$ order statistics, say $t_1(x) = \sum_{i=1}^{k} a_i x_{m_i:n},\ t_2(x) = \sum_{i=1}^{k} b_i x_{m_i:n}$, where $1 \leq m_1 < m_2 < \ldots < m_k \leq n$. The case $k = 2$ was considered in [4]. If $k > 2$, then the problem of choosing $\{a_i, b_i\}$ is appeared. Here given $\{m_i\}$ the coefficients $\{a_i, b_i\}$ are taken in such a way that t_1 and t_2 are the asymptotically best linear estimators of θ_1 and θ_2, respectively (see, e.g., [3]). The main goal of the paper is to make the best choice of order statistics, that is the best choice of $\{m_i\}$, to minimize the risk function, as $n \to \infty$, under the assumptions that $m_i/n \to p_i,\ i = 1, \ldots, k,\ 0 \leq p_1 < p_2 < \ldots < p_k \leq 1$. It turns out that such a problem is quite close to that considered in e.g. [2], Section 10.4. In the paper the problem of choice the value of k is also discussed. Several examples of location-scale families of distributions are presented.

References

[1] Alama-Bućko M., Nagaev A.V., Zaigraev A. (2006). Asymptotic analysis of minimum volume confidence regions for location-scale families. *Applicationes Mathematicae (Warszawa)*. Vol. **33**, pp. 1-20.

[2] David H.A., Nagaraja H.N. (2003). *Order Statistics*. Wiley, New York.

[3] Masoom Ali M., Umbach D. (1998). Optimal linear inference using selected order statistics in location-scale models. In: *Handbook of Statistics*. Vol. **17**, pp. 183-213. North-Holland, Amsterdam.

[4] Zaigraev A., Alama-Bućko M. (2013). On optimal choice of order statistics in large samples for the construction of confidence regions for the location and scale. *Metrika*. Vol. **76**, pp. 577-593.

Section 1

ROBUST AND MULTIVARIATE DATA ANALYSIS

ESTIMATION OF TWO-DIMENSIONAL SURVIVAL FUNCTION BY RANDOM RIGHT CENSORED DATA

A.A. Abdushukurov[1], R.S. Muradov[2]
National University of Uzbekistan
Institute of Mathematics
Tashkent, UZBEKISTAN
e-mail: [1]a_abdushukurov@rambler.ru, [2]r_muradov@myrambler.ru

Abstract

At present time there are several approaches to estimation of survival functions of vectors of life times. However, some of these estimators either are inconsistent or not fully defined in range of joint survival functions and hence not applicable in practice. Almost all of these estimators have an exponential or product structures. In this work we present absolutely other estimator of power structure for bivariate survival function of bivariate lifetime vector, which is censored from the right by censoring vector of random variables. We prove weak convergence and strong consistency results for estimators. Moreover, propose estimators of bivariate survival function from random censored observations in the presence of covariate and study the large sample properties of estimators.

1 Introduction

The problem of estimation of multivariate distribution (or survival) function from incomplete data is considered with beginning of 1980's (Campbell (1981), Campbell & Földes (1982), Hanley & Parnes (1983), Horváth (1983), Tsay, Leurgang & Crowley (1986), Burke (1988), Dabrowska (1988, 1989), Gill (1992), Huang (2000), Abdushukurov (2004) etc.)(see, [1-20]). In the special bivariate case, there are the numerous examples of paired data representing the times to death of individuals (twins or married couples), the failure times of components of system which are subjected to random censoring with possible dependence between the two censoring variables. At present time there are several approaches to estimation of survival functions of vectors of life times. However, some of these estimators either are inconsistent or not fully defined in range of joint survival functions and hence not applicable in practice. Almost all of these estimators have an exponential or product structures. In this work we present also the estimator of power structure for bivariate survival function $F(t,s)$ and present some large sample properties of estimators.

2 Random right censoring model

Let $\{X_i = (X_{1i}, X_{2i})\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed two-dimensional random vectors with a common continuous survival function $F(s;t) =$

$P(X_{11} > s, X_{21} > t)$,$(s,t) \in R^2$. This sequence is censored from the right by sequence $\{Y_i = (Y_{1i}, Y_{2i})\}_{i=1}^{\infty}$ of independent two-dimensional random vectors with survival functions $\left\{G_{(i)}(s;t) = P\left(Y_{1i} > s, Y_{2i} > s\right)\right\}_{i=1}^{\infty}$,$(s,t) \in R^2$. The statistical model is such that at the n-th stage of the experiment the observation is available the sample $V^{(n)} = \{(Z_i, \Delta_i), 1 \le i \le n\}$, where $Z_i = (Z_{1i}, Z_{2i})$,$\Delta_i = (\delta_{1i}, \delta_{2i})$, $Z_{ki} = \min(X_{ki}, Y_{ki})$ and $\delta_{ki} = I(Z_{ki} = X_{ki})$,$k = 1, 2$. The problem is consist in estimating of F from the sample $V^{(n)}$. Let $H_{(i)}(s;t) = P(Z_{1i} > s, Z_{2i} > t)$,$(s,t) \in R^2$. The model is a generalization of two-dimensional non-homogeneous random right censorship, where the vectors X_i and Y_i may be dependent. Note that this type of two-dimensional random right censoring is not considered by other authors.

The proposed estimators for F will be constructed by using the two-dimensional cumulative hazard function $-\log F(s;t) = L(s;t)$. We introduce the average functions

$$G^{(n)}(s;t) = \frac{1}{n}\sum_{i=1}^{n} G_{(i)}(s;t),\ H^{(n)}(s;t) = \frac{1}{n}\sum_{i=1}^{n} H_{(i)}(s;t),$$

$$M^{(n)}(s;t) = \frac{1}{n}\sum_{i=1}^{n} M_{(i)}(s;t),\ N^{(n)}(s;t) = \frac{1}{n}\sum_{i=1}^{n} N_{(i)}(s;t),$$

$$\widetilde{M}^{(n)}(s;t) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{M}_{(i)}(s;t),\ \widetilde{N}^{(n)}(s;t) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{N}_{(i)}(s;t),$$

where

$$M_{(i)}(s;t) = P\left(Z_{1i} \le s, Z_{2i} > t\right),\ N_{(i)}(s;t) = P\left(Z_{1i} > s, Z_{2i} \le t\right),$$

$$\widetilde{M}_{(i)}(s;t) = P\left(Z_{1i} \le s, Z_{2i} > t, \delta_{1i} = 1\right),$$

$$\widetilde{N}_{(i)}(s;t) = P\left(Z_{1i} > s, Z_{2i} \le t, \delta_{2i} = 1\right).$$

Let's consider also the cumulative hazard functions corresponding to previously defined functions:

$$\Lambda_1^{(n)}(s;t) = \int\limits_{(-\infty;s]} \frac{M^{(n)}(du;t)}{H^{(n)}(u-;t)},\ \Lambda_2^{(n)}(s;t) = \int\limits_{(-\infty;t]} \frac{N^{(n)}(s;dv)}{H^{(n)}(s;v-)},$$

$$\tilde{\Lambda}_1^{(n)}(s;t) = \int\limits_{(-\infty;s]} \frac{\tilde{M}^{(n)}(du;t)}{H^{(n)}(u-;t)},\ \tilde{\Lambda}_2^{(n)}(s;t) = \int\limits_{(-\infty;t]} \frac{\tilde{N}^{(n)}(s;dv)}{H^{(n)}(s;v-)},$$

and their estimators from the sample $V^{(n)}$:

$$\Lambda_{1n}(s;t) = \int\limits_{(-\infty;s]} \frac{M_n(du;t)}{H_n(u-;t)},\ \Lambda_{2n}(s;t) = \int\limits_{(-\infty;t]} \frac{N_n(s;dv)}{H_n(s;v-)},$$

$$\widetilde{\Lambda}_{1n}(s;t) = \int\limits_{(-\infty;s]} \frac{\widetilde{M}_n(du;t)}{H_n(u-;t)},\ \widetilde{\Lambda}_{2n}(s;t) = \int\limits_{(-\infty;t]} \frac{\widetilde{N}_n(s;dv)}{H_n(s;v-)}.$$

Here

$$M_n(s;t)=\frac{1}{n}\sum_{i=1}^{n} I(Z_1\le s, Z_{2i}>t),\quad N_n(s;t)=\frac{1}{n}\sum_{i=1}^{n} I(Z_{1i}>s, Z_{2i}\le t),$$

$$\widetilde{M}_n(s;t)=\frac{1}{n}\sum_{i=1}^{n} I(Z_{1i}\le s, Z_{2i}>t, \delta_{1i}=1),$$

$$\widetilde{N}_n(s;t)=\frac{1}{n}\sum_{i=1}^{n} I(Z_{1i}>s, Z_{2i}\le t, \delta_{2i}=1)$$

- empirical analogues of functions $M^{(n)}(s;t)$, $N^{(n)}(s;t)$, $\widetilde{M}^{(n)}(s;t)$ and $\widetilde{N}^{(n)}(s;t)$. Let's introduce

$$\Lambda^{(n)}(s;t)=\Lambda_1^{(n)}(s;-\infty)+\Lambda_2^{(n)}(s;t),\widetilde{\Lambda}^{(n)}(s;t)=\widetilde{\Lambda}_1^{(n)}(s;-\infty)+\widetilde{\Lambda}_2^{(n)}(s;t),$$

$$\Lambda_n(s;t)=\Lambda_{1n}(s;-\infty)+\Lambda_{2n}(s;t),\widetilde{\Lambda}_n(s;t)=\widetilde{\Lambda}_{1n}(s;-\infty)+\widetilde{\Lambda}_{2n}(s;t).$$

For a function of two arguments $\psi(s;t)$, put

$$\psi(\Delta s;t)=\psi(s;t)-\psi(s-;t),\ \psi(s;\Delta t)=\psi(s;t)-\psi(s;t-).$$

Consider the problem of estimation of the function $F(s,t)$ by previously estimating the following functionals:

$$F_1^{(n)}(s;t)=\exp\left(-\widetilde{\Lambda}^{(n)}(s;t)\right)=$$

$$=\exp\left(-\left(\widetilde{\Lambda}_1^{(n)}(s;-\infty)+\widetilde{\Lambda}_2^{(n)}(s;t)\right)\right),$$

$$F_2^{(n)}(s;t)=\exp\left(-\left(\tilde{\Lambda}_1^{(n)^c}(s;-\infty)\right)\right)\cdot\prod_{u\le s}\left(1-\tilde{\Lambda}_1^{(n)}(\Delta u;-\infty)\right)\cdot$$

$$\cdot\exp\left(-\left(\tilde{\Lambda}_2^{(n)^c}(s;t)\right)\right)\prod_{v\le t}\left(1-\tilde{\Lambda}_2^{(n)}(s;\Delta v)\right),\qquad(1)$$

$$F_3^{(n)}(s,t)=\left[H^{(n)}(s,t)\right]^{R^{(n)}(s,t)},$$

where $R^{(n)}(s;t)=\frac{\widetilde{\Lambda}^{(n)}(s;t)}{\Lambda^{(n)}(s;t)}$. The plug-in estimates of functionals (1) are

$$F_{1n}(s;t)=\exp\left(-\widetilde{\Lambda}_n(s;t)\right)=\exp\left(-\left(\widetilde{\Lambda}_{1n}(s;-\infty)+\widetilde{\Lambda}_{2n}(s;t)\right)\right),$$

$$F_{2n}(s;t)=\prod_{u\le s}\left(1-\widetilde{\Lambda}_{1n}(\Delta u;-\infty)\right)\cdot\prod_{v\le t}\left(1-\widetilde{\Lambda}_{2n}(s;\Delta v)\right),\qquad(2)$$

$$F_{3n}(s;t)=\left[H_n(s;t)\right]^{R_n(s;t)},$$

where $R_n(s;t)=\frac{\widetilde{\Lambda}_n(s;t)}{\Lambda_n(s;t)}$. We give some properties of the estimates (2). The following result allows to estimate the difference between the estimates. We propose some results from [1-4].

Theorem 1. ***[1-4].*** *For all* $(s,t) \in \left(-\infty, Z_{1(n)}\right) \times \left(-\infty, Z_{2(n)}\right)$ *we have*
(I) $0 \leq F_{1n}(s,t) - F_{2n}(s,t) \overset{a.s.}{=} O(\frac{1}{n})$;
(II) $|F_{1n}(s,t) - F_{3n}(s,t)| \leq \pi_n(s,t)$, a.s.,
(III) $|F_{3n}(s,t) - F_{2n}(s,t)| \leq \pi_n(s,t) + O(\frac{1}{n})$, a.s.,
where $\pi_n(s,t) = |-\log H_n(s,t) - \Lambda_n(s,t)|$ *and* $Z_{m(1)} \leq ... \leq Z_{m(n)}$ *are order statistics corresponding to the* Z_{mj}, $m = 1,2$; $j = 1,...,n$.

In the following theorem we give conditions of strong uniform consistency of estimates (2), when the censors are identically distributed, although not necessarily independent from the censoring random vectors.

Theorem 2. ***[1-4].*** *Let the pairs* $\{(X_i;Y_i),\ i \geq 1\}$ *are identically distributed and in the case of* $m = 3$ *function* $G(s;t) = P(Y_{11} > s, Y_{21} > t)$ *is a continuous. For* $m = 1,2,3$ *equality*

$$P\left(\lim_{n\to\infty} \sup_{(s;t)\in Q} |F_{mn}(s;t) - F(s;t)| = 0\right) = 1, \tag{3}$$

occurs if and only if for all $(s;t) \in Q = Supp(N) \cap Supp(M) \cap Supp\left(\widetilde{N}\right) \cap Supp\left(\widetilde{M}\right) \neq \emptyset$. *:*

$$\begin{cases} P(Y_{11} \geq s/X_{11} = s) = P(Y_{11} \geq s/X_{11} > s), \\ P(Y_{11} > s, Y_{21} \geq t/X_{11} > s, X_{21} = t) = P(Y_{11} > s, Y_{21} \geq t/X_{11} > s, X_{21} > t). \end{cases} \tag{4}$$

Corollary 1. *It is easy to see that if the pair*(X_{11}, Y_{11}) *of the first components of the vectors* $X_1 = (X_{11}, X_{21})$ *and* $Y_1 = (Y_{11}, Y_{21})$ *is independent from the pair* (X_{21}, Y_{21}) *of second components, then the system (4) can be written as follows:*

$$\begin{cases} P(Y_{11} \geq s/X_{11} = s) = P(Y_{11} \geq s/X_{11} > s), \\ \\ P(Y_{21} \geq t/X_{21} = t) = P(Y_{21} \geq t/X_{21} > t). \end{cases} \tag{5}$$

Note also that for estimators (2) we prove results of weak convergence to the appropriate Gaussian processes.

Example. Let the joint distribution function of the pairs $\{(X_{k1}, Y_{k1}), k = 1,2\}$ is the two-dimensional exponential distribution of Marshall-Olkin with parameters $\{(\lambda_{1,k}, \lambda_{2,k}, \lambda_{3,k}),\ k = 1,2\}$:

$$P(X_{k1} > s, Y_{2k} > t) = \exp\{-\lambda_{1,k}s - \lambda_{2,k} - \lambda_{3,k}\max\{s,t\}\},\ (s,t) \in (0,\infty)^2.$$

then it is easy to see that system (5) is hold.

It should be noted that all these results authors generalized to the case of random Poisson sample size. In [6-8, 14] the authors consider the problem of estimation of multivariate survival functions in dependent models of random censoring using copula functions and in the presence of covariates.

References

[1] Abdushukurov A.A. (2004). Nonparametrical estimators of survival function on the plane by censored observations. *Uzbek mathematical journal*. Vol. **2**, pp. 311.(In Russian)

[2] Abdushukurov A.A. (2009). *Statistics incomplete observations*. NUUz University press, Tashkent.(In Russian)

[3] Abdushukurov A.A. (2011). *Estimates of unknown distributions from incomplete observations and their properties*. LAMBERT Academic Publishing, Germany.(In Russian)

[4] Abdushukurov A.A. (2016). Estimation of joint survival function from censored observations. *Zavod. Lab. Diagn. Mater.*. Vol. **82**, pp. 80-84.(In Russian)

[5] Abdushukurov A.A., Muradov R.S. (2014). On the estimates of the distribution function in a random censorship model. *Zavod. Lab. Diagn. Mater.*. Vol. **80**, pp. 62-67.(In Russian)

[6] Abdushukurov A.A., Muradov R.S. (2014). On Estimation of Conditional Distribution Function under Dependent Random Right Censored Data. *Journal of Siberian Federal University*. Vol. **7**, pp. 409-416.

[7] Abdushukurov A.A., Muradov R.S. (2014). Estimation of survival functions from dependent random right censored data. *Intern. J. Innovation in Science and Mathematics*. Vol. **2**, pp. 280-287.

[8] Abdushukurov A.A., Muradov R.S. (2016). Some algebraic properties of Archimedean copula functions and their applications in the statistical estimation of the survival function. *New Trends in Math. Sci.*. (to appear).

[9] Burke M.D. (1988). Estimation of a bivariate distribution function under random censorship. *Biometrika*. Vol. **75**, pp. 379-382.

[10] Campbell G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika*. Vol. **68**, pp. 417422.

[11] Campbell G. Földes A. (1982). Lange sample properties of nonparametric bivariate estimators with censored data. *Colloquia Mathematica-Societatis János Bolyai*. Vol. **32**, pp. 103-122.

[12] Dabrowska D.M. (1988). Kaplan Meier estimate on the plane. *Ann. Statist.*. Vol. **16**, pp. 1475-1489.

[13] Dabrowska D.M. (1989). Kaplan Meier estimate on the plane: weak convergence, LIL and the bootstrap *J. Multivar. Anal.* . Vol. **29**, pp. 308-325.

[14] Gill R.D. (1992). Multivariate Survival Analysis I. *Theory Probab. Appl.*. Vol. **37**, pp. 1935.(In russian)

[15] Gill R.D. (1992). Multivariate Survival Analysis II. *Theory Probab. Appl.*. Vol. **37**, pp. 307328.(In russian)

[16] Hanley J.A., Parnes M.N. (1983). Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrica*. Vol. **39**, pp. 129-139.

[17] Horváth L. (1983). The rate of strong uniform consistency for the multivariate product-limit estimator. *J. Multivar. Anal.* . Vol. **13**, pp. 202209.

[18] Huang Y. (2000). Twosample multistate accelerated sojourn times model. *J.A.S.A.* . Vol. **95**, pp. 619627.

[19] Muradov R.S., Abdushukurov A.A. (2011). *Estimation of multivariate distributions and there mixes by incomplete data*. LAMBERT Academic Publishing, Germany.(In russian)

[20] Tsai W.Y., Leurgans S., Crowley J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *Ann. Statist.* . Vol. **14**, pp. 1351-1365.

GROUP DETECTION IN THE CONTEXT OF IMBALANCED DATA

S. Brodinova[1], M. Zaharieva[2], P. Filzmoser[3],
T. Ortner[4], C. Breiteneder[5]
[1,2,4,5]*Interactive Media Systems Group, TU Wien*
[3,4]*Institute of Statistics and Mathematical Methods in Economics, TU Wien*
[2]*Multimedia Information Systems Groups, University of Vienna*
Vienna, AUSTRIA
e-mail: [1]sarka.brodinova@tuwien.ac.at

Abstract

The problem of group detection with no prior knowledge, i.e clustering, is one of the most important tasks in data analysis. It has been addressed in many applications in various fields. Data clustering becomes challenging when the group sizes are very different–this is called imbalanced data–with different densities and shapes. This task is even more difficult in the context of high-dimensional data since it is very hard do state any assumption about specific characteristics of groups (sizes, densities, or shapes). However, many clustering techniques are built upon some of these assumptions. For instance, the most popular *k-means* method [3] can be shown as a particular case of the EM algorithm for data generated by Gaussian mixtures [1]. In addition, many clustering algorithms (also *k-means*) require the ad-hoc specification of parameters, especially the number of clusters. This is almost impossible to know beforehand. Unfortunately, the final clustering solution usually depends on the choice of the predefined parameters.

We propose an algorithm which identifies the clusters in imbalanced high-dimensional data. Our procedure incorporates an existing clustering method in order to detect the homogeneous set of initial clusters. These initial clusters are successively merged in order to build final clusters. Merging a pair of initial clusters is based on Local Outlier Factor [2] (LOF) which captures the final clusters of arbitrary sizes without assumptions on cluster characteristics. The fact of small group sizes in imbalanced data makes the observations of those groups atypical. Therefore, our special focus is towards the ability of finding these interesting groups next to the description of the data structure. The usefulness of our approach is demonstrated with imbalanced media data sets, and it is shown that state-of-the-art methods are outperformed.

References

[1] Aggarwal C.C., Reddy C.K. (2013). *Data clustering: algorithms and applications.* CRC Press.

[2] Breunig M.M., Kriegel H.-P., Ng R.T., Sander J. (2000). *LOF: identifying density-based local outliers.* Proc. 2000 ACM SIGMOD, New York. pp. 93–104.

[3] Jain A.K. (2010). *Data clustering: 50 years beyond K-means.* Pattern recognition letters, **31**(8), pp. 651–666.

ERROR PROBABILITIES IN SEQUENTIAL TESTING OF SIMPLE HYPOTHESES FOR DEPENDENT OBSERVATIONS

S. Y. Chernov
Belarusian State University
Minsk, BELARUS
e-mail: chernovsy@tut.by

Abstract

The problem of the error probabilities evaluation for the sequential probability ratio test (SPRT) if observations have the Markov dependence is considered.

1 Introduction

The sequential method of hypotheses testing is widely used for information processing in medicine, statistical quality control, biology [7] and finance [6]. A profit of the sequential procedures is that the average number of observations is less than the fixed number of observations for the equivalent tests procedures [1].

The sequential test proposed by A.Wald [1] is considered in the paper.

One of the sequential approach disadvantages is that the probabilities of types I and II errors are influenced greatly by the distortions in observations. Tukey-Huber outliers in sequential testing when observations are independent random variables are considered in [3]. Functional distortions in L_1-, L_2- and C-metrics are considered in [4] and [5]. It should be noted that the probabilities of types I and II errors of the sequential tests could not be calculated exactly in general case. Therefore providing robustness analysis one should use not the values of error probabilities, but their analytical approximations. So the construction of such approximations is actual problem. In [8] and [3] the approach for approximate calculation of error probabilities is described when observations are independent random variables. In this paper the case where observations form Markov sequence is considered.

2 Mathematical Model

Let $\lambda_1, \lambda_2, \ldots$ be the homogeneous Markov chain on the measurable space $(\Omega, \mathcal{F})$. Let D be the state space of (λ_n), $D \subseteq \mathbf{R}$, $|D| < \infty$, $p(x|y)$ – the transition function and $f_1(x)$ – stationary probability density function. It is clear that the random sequence $(\lambda_{n+1}, \lambda_n)$ is homogeneous Markov chain as well. Let $f_2(x, y)$ be stationary distribution of $(\lambda_{n+1}, \lambda_n)$. Suppose that $f_1(x)$ satisfies $\sup\{f_1(x)\} \leq M < +\infty$ if $x \in D$.

As it is done in [8] let us try to transform the random sequence λ_n with continuous state space into the Markov chain ξ_n with finite (enumerable) state space. As D is the state space of λ_n, then $a_1, \ldots, a_m$, $a_k \in D$, is the set space of ξ_n, where m is the parameter of transformation λ_n into ξ_n. Divide D by m subsets A_1, $A_2, \ldots, A_m$, i.e.

$D = \bigsqcup_{k=1}^{m} A_k$, and a_k corresponds to A_k. Let $\phi_m(\cdot)$ be the function of such correspondence, i.e. $\xi_n = \phi_m(\lambda_n)$.

The example of described function $\phi_m(\cdot)$ is $\phi_m(x) = A + \left[\frac{x-A}{h}\right] h$, where $\lambda_n \in [A; B] = D$. In this case $a_k = A + (k-1)h$, $A_k = [A + (k-1)h; A + kh)$, $|A_k| \to 0$ at $m \to \infty$.

Let us investigate when constructed random sequence ξ_n forms the Markov chain.

Theorem 1. *Under specified assumptions the random sequence ξ_n is the first order Markov chain iff observations λ_n are independent.*

Now let ξ_n not to be a Markov chain, but only "similar" to one. What can we say about random sequence λ_n, if transition probabilities of ξ_n are "almost" Markovian?

Theorem 2. *Transition probabilities of ξ_n satisfy*

$$P\{\xi_{n+1} = a_j \mid \xi_n = a_i, \xi_{n-1} = a_k\} = P\{\xi_{n+1} = a_j \mid \xi_n = a_i\} + \varepsilon_m,$$

where $\varepsilon_m \to 0$ at $m \to \infty$ iff observations λ_n are independent.

Due to given theorems it is impossible to approximate the Markov sequence with continuous space by the Markov sequence with finite space, so the method in [8] of approximation of error probabilities of the sequential test with independent observations is not suitable for the case when observations have Markov dependence.

References

[1] Wald A. (1947) *Sequential Analysis*. Wiley, NY.

[2] Kharin A. (2002) An approach to performance analysis of the SPRT for simple hypotheses testing. *Proc. Belarusian State Univ.* Vol. **1**, pp. 92-96. (In Russian)

[3] Kharin A., Kishilau D. (2005) Robust Sequential Testing of Hypotheses on Discrete Probability Distributions. *Austrian J. Stat.* Vol. **32**(2), pp. 153-162.

[4] Chernov S., Kharin A. (2014) An Approach to Robustness Evaluation for Sequential Testing under Functional Distortions in L_1-metric and C-metric. *Austrian J. Stat.* Vol. **43**(3), pp. 195-203.

[5] Chernov S., Kharin A. (2013) Error Probabilities of the Sequential Test under Functional Distortions in L_2-metric. *Stat. Meth. Estim. and Hyp. Testing*. Vol. **25**. pp. 64-72. (in Russian, translated by Springer)

[6] Lai T.L. (2001) Sequential analysis: some classical problems and new challenges. *Statistica Sinica*. Vol. **11**. pp. 303-408.

[7] Mukhopadhyay N. et al. (2004) *Applied Sequential Methodologies*. M.Dekker, NY.

[8] Woodall W., Reynolds M. (1983). A Discrete Markov Chain Representation of the Sequential Probability Ratio Test. *Comm. in Stat. – Seq. Analysis*. Vol. **2**(1), pp. 27-44.

ON STOCHASTIC PERTURBATION METHOD FOR ESTIMATION OF HIGH DIMENSIONAL MATRIX

H. S. Hoang[1], R. Baraille
DOPS/HOM/REC, SHOM
Toulouse, FRANCE
e-mail: [1]hhoang@shom.fr

Abstract

A simple stochastic algorithm is proposed for estimating the elements of a matrix as well as its decomposition under the condition that only the matrix-vector product is accessible. Theoretical results on convergence of the algorithm are presented.

1 Introduction

Consider the following linear system of equations

$$\Phi x = b, \tag{1}$$

where $\Phi \in R^{m\times n}, b \in R^m, x \in R^n$. In (1) b is a known vector, Φ is unknown, but $b' = \Phi x'$ is known once x' is given. The problem we are interested in is to estimate the matrix Φ and to decompose it into the product of two matrices, $\Phi = \Phi_1\Phi_2$.

The motivation for solving the aforementioned problems arises in many engineering inverse problems. As an example, consider the filtering problem

$$x_{k+1} = \Phi x_k + w_k, z_{k+1} = Hx_{k+1} + v_{k+1}, \tag{2}$$

$\Phi \in R^{n\times n}$, $H \in R^{p\times n}$. Based on the set of observations $z_l, l = 1, 2, ..., k$, under standard conditions, a minimum mean squared estimate $\hat{x}_k$ can be obtained by the Kalman filter (KF) [5]. In many engineering problems, Φ is unknown and its dimension n is very high, for example, of orders $O(10^6) - O(10^7)$. In such situations, it is very important to have a possibility to estimate numerically Φ, to store and manipulate it. Here the product $b' = \Phi x'$ is known once x' is given. For moderate n, a component-wise integration method is of common use (see [1]).

2 Estimation of matrix

For $b := (b_1, ..., b_m)'$, the derivatives of b_i with respect to (w.r.t.) to the vector x is defined as

$$db_i/dx = (\partial b_i/\partial x_1, ..., \partial b_i/\partial x_n) = (\phi_{i1}, ..., \phi_{in}),\ i = 1, \ldots, m,$$

where ϕ_{ij} are the ij element of Φ. We can write then

$$db/dx = [(db_1/dx)', \ldots, (db_m/dx)']' = \Phi.$$

In what follows we present a low-cost algorithm for approximating derivatives of b w.r.t. x, independently of the dimension of x, and to estimate Φ. The idea on estimation of high dimensional Φ on the basis of stochastic simultaneous perturbation (SSP) has been first briefly presented in [3]. Remember that in [7] Spall proposes a simultaneous perturbation stochastic approximation (SPSA) algorithm for seeking optimal parameters by minimizing some objective function. The main feature of this algorithm resides in the way to approximate the gradient vector : a sample gradient vector is estimated by SSP of all components of the unknown vector. This method requires only two or three measurements of the objective function, regardless of the dimension of the vector of unknown parameters.

Let $\bar{\Delta} := (\Delta_1, ..., \Delta_n)'$, $\Delta_i, i = 1, ..., n$ be Bernoulli independent and identically distributed (i.i.d.) variables assuming two values +/- 1 with equal probabilities 1/2. Introduce $[\bar{\Delta}]^{-1} := (1/\Delta_1, ..., 1/\Delta_n)'$, $\bar{\Delta}_c := c\bar{\Delta}, c > 0$ is a small positive value.

In the context of estimating Φ, the proposed algorithm looks as follows:

Algorithm 2.1. Suppose it is possible to compute the product $\Phi x = b(x)$ for a given x. At the beginning let $l = 1$. Let the value u be assigned to the vector x, i.e. $x := u$, L be a (large) fixed integer number.

Step 1. Generate $\bar{\Delta}^{(l)}$ whose components are l^{th} samples of the Bernoulli i.i.d. variables assuming two values +/- 1 with equal probabilities 1/2;

Step 2. Compute $\delta b^{(l)} = \Phi(u + \bar{\Delta}_c^{(l)}) - \Phi u$, $\bar{\Delta}_c^{(l)} = c\bar{\Delta}^{(l)}$, c is a small positive value;

Step 3. Compute $g_i^{(l)} = \delta b_i^{(l)}[\bar{\Delta}_c^{(l)}]^{-1}$, δb_i is the i^{th} component of δb, $g_i^{(l)}$ is the column vector consisting of derivative of $b_i(u)$ w.r.t. to u, $i = 1, ..., m$.

Step 4. Go to Step 1 if $l < L$. Otherwise, go to Step 5.

Step 5. Compute

$$\hat{\Phi}(L) := D_x b = [\hat{g}_1, ..., \hat{g}_m]', \ \hat{g}_i = L^{-1} \sum_{l=1}^{L} g_i^{(l)}, \ i = 1, \ldots, m.$$

Theorem 1. *Consider Algorithm 2.1 for estimation of the elements of the matrix Φ. Then this algorithm will yield the estimates for the elements of Φ with the mean squared error (MSE) $O(1/L)$ where L is the number of samples used in the estimation procedure.*

3 Estimation of decomposition of Φ

For very high dimensional Φ, it is impossible to store all the elements of Φ. This difficulty can be overcome by approximating Φ by some matrix in a subspace of fewer dimensions (for example, the class of matrices of given rank).

Let $\Phi \in R^{m \times n}, m \leq n$ with $\text{rank}(\Phi) = m$. We want to find a best approximation for Φ within the class of matrices:

$$\Phi_e = AB', \ A \in R^{m \times r}, \ B \in R^{n \times r}, \ m \geq r = \text{rank}(AB'). \tag{3}$$

Under the condition (3), the optimization problem is formulated as

$$J(A,B) = ||\Phi - \Phi_e||_F^2 = ||\Phi - AB'||_F^2 \to \min_{(A,B)}, \tag{4}$$

where $||.||_F$ denotes the Frobenius matrix norm.

Consider Φ and let $U\Sigma V'$ be SVD (singular value decomposition) of Φ [2], i.e.

$$\begin{aligned} \Phi = U\Sigma V', U \in R^{m\times m}, V \in R^{n\times n}, \Sigma = [\Sigma_m|0], \\ \Sigma_m = \text{diag}[\sigma_1, ..., \sigma_m], \sigma_1 \geq \sigma_2 \geq ... \geq \sigma_m \geq 0. \end{aligned} \tag{5}$$

Theorem 2. *Suppose A_oB_o' is a solution to the problem (4),(3). Then*

$$J(A_o, B_o) = \sum_{k=r+1}^{m} \sigma_k^2 \tag{6}$$

Theorem 2 implies that $\Phi_e^o := A_oB_o'$ is equal to the matrix formed by truncating the SVD of Φ (5) to its first r singular vectors and singular values.

By the same way as seen in Algorithm 2.1, one can write out the algorithm for estimating A and B. The vector of unknown parameters θ consists of all the elements of A and B. We have to perturb simultaneously all the components of θ stochastically, compute the gradient of $J(A,B)$ w.r.t. to θ and iterate the estimation procedure.

4 Simulation experiment

It is important to emphasize that for a finite number of samples L, the estimation error for $\hat{\Phi}^{(L)}$ depends on the number of non-zero elements of Φ. That is why an appropriate assumption on sparsity of Φ plays the important role on quality of the estimate $\hat{\Phi}^{(L)}$. To see the impact of this assumption, consider the nonlinear transport problem (page 109, [6]). Applying upwind difference scheme, the numerical model obtained has the state vector $x(t_k)$ of dimension $n = 51$ with $\delta x = 1/(n-1), \delta t = 0.00833$. The observation vector has the form $z_i(k) := x_j(k) + v_i, i = 1, ..., 25, j = 2i$. Mention that Φ of the linearized system has a diagonal structure with non-zero diagonal $\Phi(i,i)$ and up-diagonal elements $\Phi(i,i+1)$. Different Extended KFs (EKFs) are used to estimate the system state subject to different estimates for Φ obtained by: (A0) exact linearization; (A1) Algorithm 2.1 without any assumption on structure of Φ; (A2) Algorithm 2.1 with the assumption that all elements of Φ are zero except for $\phi(i,j)$, $|i-j| \leq 1$; (A3) Algorithm 2.1 with exact structure of Φ. Figure 1 shows performances of different EKFs.

5 Conclusion

A simple algorithm for estimating an unknown matrix as well as the way to decompose it into a product of two matrices, have been proposed. Based on this algorithm, different numerical problems like SVD decomposition, Nearest Kronecker Problem (NKP) in

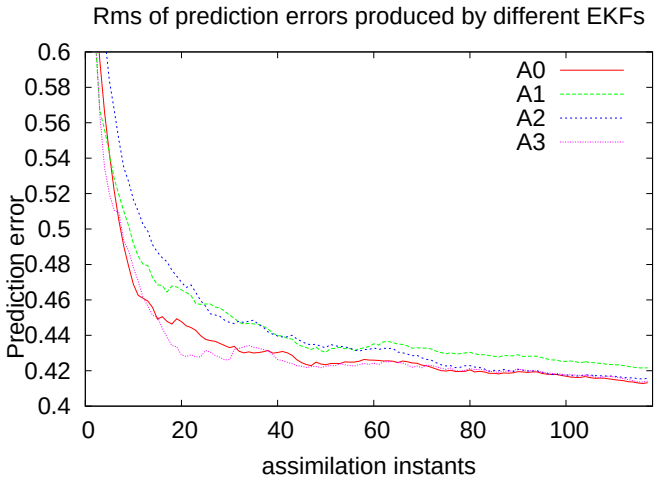


Figure 1: Rms of prediction error (PE) resulting from different EKFs

high dimensional setting can be solved in a simple and efficient way, compared to classical algorithms (see [2], for example). This algorithm constitutes also a basis for estimation of the error covariance matrix using the hypothesis on separation of vertical and horizontal variables [4] for geophysical systems.

References

[1] Fukumori I., Malanotte-Rizzoli P. (1995) An Approximate Kalman Filter for Ocean Data Assimilation: An Example with an Idealized Gulf Stream Model, *J. Geophys. Resear.*, Vol. **100**(C4), pp. 6777–6793.

[2] Golub G.H., van Loan C.F. (1996) *Matrix Computations*, Cambridge Univ. Press.

[3] Hoang H.S., Baraille R. (2015) Stochastic Simultaneous Perturbation as Powerful Method for State and Parameter Estimation in High Dimensional Systems. In *Adv. in Math. Research*. Ed. A.R. Baswell, Nova Pub., Vol. **20**, pp. 117-148.

[4] Hoang H.S. and Baraille R. (2014) A Low Cost Filter Design for State and Parameter Estimation in Very High Dimensional Systems, *Proc. 19th IFAC World Congress, Cap Town, South Africa*. Vol. **19**(1), pp. 3256–3261.

[5] Kalman R.E. (1960) A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME-J. Basic Engineering*, Vol. **82**(D), pp. 35–45.

[6] Sewell G. (1988) *The numerical solution of ordinary and partial differential equations*, Acad. Press.

[7] Spall J.C. (1992) Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, *IEEE Trans. Autom. Contr.*, Vol. **37**(3), pp. 332–341.

ROBUST AND SPARSE MULTICLASS CLASSIFICATION BY THE OPTIMAL SCORING APPROACH

I. Hoffmann[1], P. Filzmoser, C. Croux
Institute of Statistics and Mathematical Methods in Economics, TU Wien
Vienna, AUSTRIA
e-mail: [1]irene.hoffmann@tuwien.ac.at

Abstract

We present a robust and sparse linear classifier for multiclass problems. Regression methods with the desirable properties of outlier detection and variable selection have got a lot of attention recently. The optimal scoring approach enables a propagation of such regression methods to classification problems.

1 Introduction

Linear discriminant analysis (LDA) is a very simple and popular method for classification, but the model can be distorted by a single anomalous observation. Several robust methods for linear classification have been introduced (see for an overview [3]) to address this issue. These methods are restricted to settings with more observations than variables.

One popular approach in regression analysis for data with a large number of variables compared to the number of observations is Lasso regression. An L_1 norm penalty on the coefficient estimate favours exact zero entries and so excludes uninformative variables from the model. This idea has been incorporated in least trimmed squares (LTS) regression and yields a sparse and robust regression method [1].

2 Optimal Scoring

Optimal scoring is an approach where the class labels are modelled as continuous values and so the classification problem is recast into a regression framework. Let $\boldsymbol{X}$ be an $n \times p$ data matrix, $\boldsymbol{Y}$ an $n \times G$ matrix of dummy variables coding the class membership of the observations, G be the number of classes and $H = G - 1$. For $h = 1, ..., H$

$$\min_{\boldsymbol{\beta}_h, \boldsymbol{\theta}_h} \{\|\boldsymbol{Y}\boldsymbol{\theta}_h - \boldsymbol{X}\boldsymbol{\beta}_h\|^2\} \quad \text{s.t.} \quad \frac{1}{n}\boldsymbol{\theta}_h^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{\theta}_h = 1, \quad \boldsymbol{\theta}_h^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{\theta}_l = 0 \quad \forall l < h. \tag{1}$$

Adding an L_1 penalty for $\boldsymbol{\beta}_h$ to the minimization problem leads to sparse discriminant analysis as proposed by [2].

3 Methodology

We propose to propagate the robustness and sparsity properties of sparse LTS to classification problems via the optimal scoring approach. The optimal scoring problem is solved iteratively for $\boldsymbol{\beta}_h$ and $\boldsymbol{\theta}_h$. For fixed $\boldsymbol{\theta}_h$ we replace the least squares minimization in (1) by trimmed least squares with L_1 penalty and solve it with a fast sparse LTS algorithm. This leads to a sparse and robust estimation of $\boldsymbol{\beta}_h$ for $h = 1, \ldots, H$. Then LDA with robustly estimated scatter matrix and centre is applied to $(\boldsymbol{X}\hat{\boldsymbol{\beta}}_1, ..., \boldsymbol{X}\hat{\boldsymbol{\beta}}_H)$.

4 Evaluation

A simulation study is conducted to illustrate the properties of the proposed algorithm. Its performance is compared to classical sparse discriminant analysis by means of correctly selected variables and the ratio of misclassified observations. For simulation settings with more observations than variables further comparison is made with LDA and robust LDA methods.

5 Acknowledgements

This work is supported by the Austrian Science Fund (FWF), project P 26871-N20.

References

[1] Alfons A., Croux C., Gelper S. (2013). Sparse least trimmed squares regression for analysing high-dimensional large data sets. *The Annals of Applied Statistics*. Vol. **7**(1), pp. 226-248.

[2] Clemmensen L., Hastie T., Witten D., Ersboll B. (2011). Sparse discriminant analysis. *Technometrics*. Vol. **53**, pp. 406-413.

[3] Todorov V., Pires A.M. (2007). Comparative performance of several robust linear discriminant analysis methods. *REVSTAT Statistical Journal*. Vol. **5**, pp. 63-83.

EVALUATION OF SEQUENTIAL TEST CHARACTERISTICS FOR TIME SERIES WITH A TREND

Alexey Kharin[1], Ton That Tu[2]
Belarusian State University
Minsk, BELARUS
e-mail: [1]KharinAY@bsu.by, [2]tthattu@gmail.com

Abstract

The problem of sequential testing of simple hypotheses for time series with a trend is considered. Analytical expressions and asymptotic expansions of error probabilities and expected numbers of observations are obtained. The result is illustrated numerically.
Keywords: sequential test, time series, trend, error probability, expected number of observations

1 Introduction

The sequential approach to test parametric hypotheses was proposed by Wald (see [6]) and is applied in many practical problems of statistical data analysis. The problem of sequential test characteristics (error probabilities and expected number of observations) evaluation is well studied for the case of identical distribution of observations (see [1] – [6]). In this paper, the model of non-identical distribution is considered.

Let $x_1, x_2, ...$ be observations of time series with a trend:

$$x_t = \theta^T \psi(t) + \xi_t,\ t = 1, 2, 3, ..., \tag{1}$$

where $\psi(t) = (\psi_1(t), \psi_2(t), ..., \psi_m(t))^T,\ t \geq 1$, are the vectors of basic functions of trend, $\theta = (\theta_1, \theta_2, ..., \theta_m)^T \in \mathbb{R}^m$ is an unknown vector of coefficients, and $\{\xi_t, t \geq 1\}$ is the sequence of independent identically distributed random variables, $\xi_t \sim N(0, \sigma^2)$.

Consider two simple hypotheses:

$$H_0 : \theta = \theta^0, H_1 : \theta = \theta^1, \tag{2}$$

where $\theta^0, \theta^1 \in \mathbb{R}^m$ are known vectors.

Denote the accumulated log-likelihood ratio statistic:

$$\Lambda_n = \Lambda_n(x_1, x_2, ..., x_n) = \sum_{t=1}^{n} \lambda_t, \tag{3}$$

where $\lambda_t = ln\left(\dfrac{p_t(x_t, \theta^1)}{p_t(x_t, \theta^0)}\right)$ is the log-likelihood ratio calculated on the observation x_t, and $p_t(x, \theta)$ is the probability density function of x_t provided the parameter value is θ.

To test these hypotheses, after n observations one makes the decision:

$$d = \mathbf{1}_{[C_+,+\infty)}(\Lambda_n) + 2 \cdot \mathbf{1}_{(C_-,C_+)}(\Lambda_n). \tag{4}$$

The thresholds C_- and C_+ are the parameters of the test. Decisions $d = 0$ and $d = 1$ mean stopping of the observation process and acceptance of H_0 or H_1 correspondently. According to Wald (see [6]) we use $C_+ = ln\left(\frac{1-\beta_0}{\alpha_0}\right)$ and $C_- = ln\left(\frac{\beta_0}{1-\alpha_0}\right)$, where α_0, β_0 are the given values for probability errors of types I and II respectively.

2 Main results

Introduce the notation: $E^{(k)}(\cdot), D^{(k)}(\cdot)$ are conditional expected value and variance provided hypothesis H_k is true $(k = 0, 1)$; for $n \geq 1$,

$$\sigma_n^2 = \frac{(\theta^0 - \theta^1)^T \psi(n)\psi^T(n)(\theta^0 - \theta^1)}{\sigma^2}, \mu_n^{(k)} = \frac{(-1)^{k+1}\sigma_n^2}{2}, s_n^2 = \sum_{t=1}^{n} \sigma_t^2,\ m_n^{(k)} = \frac{(-1)^{k+1}s_n^2}{2},$$

$$A_n = \{a_{ij}\}_{n\times n}, \quad a_{ij} = \begin{cases} 1, & i \geq j, \\ 0, & otherwise; \end{cases} \quad X_n = (\lambda_1, \lambda_2, ..., \lambda_n)^T,$$

$$T_n = (\Lambda_1, \Lambda_2, ..., \Lambda_n)^T = A_n X_n, \quad \mu_{T_n}^{(k)} = A_n E^{(k)}(X_n), \Sigma_{T_n} = A_n Cov(X_n, X_n) A_n^T;$$

$\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution $N(0, 1)$. Put $N = \inf\{n \in \mathbb{N} : \Lambda_n \notin (C_-, C_+)\}, \Gamma = (\theta^0 - \theta^1)(\theta^0 - \theta^1)^T$ and $H_n = \sum_{i=1}^{n} \psi(i)\psi^T(i)$. Let α, β be the factual values of the error type I and II probabilities for test (3), (4).

Theorem 1. *If the trace of the matrix ΓH_n tends to $+\infty$ when $n \to +\infty$, then the test terminates finitely with probability 1.*

Proof. The proof is derived from the fact that $P_k(N > n) \leq P_k(\Lambda_n \in (C_-, C_+))$. □

Corollary 1. *If $tr\{\Gamma H_n\}$ is bounded, then there exists a positive constant L so that $s_n^2 \to L$ when $n \to +\infty$. In this case, we have:*

$$\lim_{n\to+\infty} P_k(\Lambda_n \in (C_-, C_+)) = \Phi\left(\frac{2C_+ - (-1)^{k+1}L}{2\sqrt{L}}\right) - \Phi\left(\frac{2C_- - (-1)^{k+1}L}{2\sqrt{L}}\right) > 0.$$

Theorem 2. *Under the Theorem 1 condition following expressions are valid for the characteristics of test (2):*

$$E^{(k)}(N) = 1 + \sum_{i=1}^{+\infty} \int_{C_-}^{C_+} ds_i \int_{C_-}^{C_+} ds_{i-1} ... \int_{C_-}^{C_+} n_i(s, \mu_{T_i}^{(i)}, \Sigma_{T_i}) ds_1,\ k = 0, 1;$$

$$\alpha = \int_{C_+}^{+\infty} n_1(s_1, \mu_1^{(0)}, \sigma_1^2) ds_1 + \sum_{i=2}^{+\infty} \int_{C_+}^{+\infty} ds_i \int_{C_-}^{C_+} ds_{i-1} ... \int_{C_-}^{C_+} n_i(s, \mu_{T_i}^{(0)}, \Sigma_{T_i}) ds_1,$$

$$\beta = \int_{-\infty}^{C_-} n_1(s_1, \mu_1^{(1)}, \sigma_1^2) ds_1 + \sum_{i=2}^{+\infty} \int_{-\infty}^{C_-} ds_i \int_{C_-}^{C_+} ds_{i-1} ... \int_{C_-}^{C_+} n_i(s, \mu_{T_i}^{(1)}, \Sigma_{T_i}) ds_1.$$

Proof. The results above are proved directly by using the properties of multivariate normal distributions. □

Corollary 2. *Under the Theorem 1 condition, the following inequalities hold:*

$$E^{(k)}(N) \le 1 + \sum_{i=1}^{+\infty} \int_{iC_-}^{iC_+} n_1(x, \bar{m}_i^{(k)}, \bar{s}_i^2)dx,\ k = 0, 1;$$

$$\alpha \le 1 - \Phi\left(\frac{C_+ - \mu_1^{(0)}}{\sigma_1}\right) + \sum_{i=2}^{+\infty} \int_{C_+}^{+\infty} \int_{C_-}^{C_+} n_1(x, m_{i-1}^{(0)}, s_{i-1}^2) n_1(y, x + \mu_i^{(0)}, \sigma_i^2)dxdy,$$

$$\beta \le \Phi\left(\frac{C_- - \mu_1^{(1)}}{\sigma_1}\right) + \sum_{i=2}^{+\infty} \int_{-\infty}^{C_-} \int_{C_-}^{C_+} n_1(x, m_{i-1}^{(1)}, s_{i-1}^2) n_1(y, x + \mu_i^{(1)}, \sigma_i^2)dxdy,$$

where $\bar{m}_i^{(k)} = \dfrac{(-1)^{k+1}}{2} \sum\limits_{j=1}^{i} (i + 1 - j)\sigma_j^2$, $\bar{s}_i^2 = \sum\limits_{j=1}^{i} (i + 1 - j)^2 \sigma_j^2$.

To construct asymptotic expansions, split the state space of Λ_n into $K + 2$ cells:

$$A_0 = (-\infty, C_-),\ A_i = [C_{i-1}, C_i),\ i = \overline{1, K},\ A_{K+1} = [C_+, +\infty)$$

$$C_- = C_0 < C_1 < C_2 < ... < C_K = C_+,\ C_i = C_- + ih,\ h = \frac{C_+ - C_-}{K}, i = \overline{1, K}.$$

Denote $f_{C_-}^{C_+}(x) = \left(\left[\dfrac{x - C_-}{h}\right] + 1\right) \cdot \mathbf{1}_{(C_-, C_+)}(x) + (K + 1) \cdot \mathbf{1}_{[C_+, +\infty)}(x)$.
For the random sequence Λ_n, let us introduce the discrete random sequence Z_n with the finite state space $V = \{0, 1, ..., K + 1\}$. Put $Z_1 = f_{C_-}^{C_+}(\Lambda_1)$ and for $n \ge 2$:

$$Z_n = \begin{cases} 0, & \text{if } Z_{n-1} = 0, \\ K + 1, & \text{if } Z_{n-1} = K + 1, \\ f_{C_-}^{C_+}(\Lambda_n), & \text{otherwise.} \end{cases}$$

In this case, Z_n is an inhomogeneous Markov chain with a finite state space $\{0, ..., K + 1\}$, in which 0 and $K + 1$ are absorbing states. In order to simplify the notation, let us renumerate the states space of Z_n: $V = \{\{0\}, \{K + 1\}, \{1\}, ..., \{K\}\}$.

Introduce the notation:

$$P^{(n)}(\theta^i) = \left(\begin{array}{c|c} I_2 & \mathbf{O}_{2\times K} \\ \hline R_n(\theta^i) & Q_n(\theta^i) \end{array}\right), \quad i = 0, 1; P^{(n)}(\theta^i) = \{p_{kl}^{(n)}(\theta^i)\}_{(K+2)\times(K+2)},$$

$$p_{kl}^{(n)}(\theta^i) = \frac{\int_{A_k} n_1(y, m_{n-1}^{(i)}, s_{n-1}^2) \int_{A_l} n_1(x, y + \mu_n^{(i)}, \sigma_n^2)dxdy}{\int_{A_k} n_1(y, m_{n-1}^{(i)}, s_{n-1}^2)dy},$$

$$S(\theta^i) = I_K + \sum_{k=1}^{+\infty} \prod_{j=1}^{k+1} Q_j(\theta^i), \quad B(\theta^i) = R_2(\theta^i) + \sum_{k=2}^{+\infty} \prod_{j=1}^{k} Q_j(\theta^i) R_{k+1}(\theta^i);$$

$B_{(j)}(\cdot)$ is the j^{th}-column of matrix $B(\cdot), \pi(\theta^i)$ is the probability distribution of Z_1, $\mathbf{1}_K$ is the vector of size K with all components equal to 1, $t(\theta^i) = E(N|\theta^i), i = \overline{0, 1}$.

Theorem 3. *If* $\inf_n tr(\Gamma\psi(n)\psi^T(n)) \geq C, C = const > 0$, *then the characteristics of the test* (2) *satisfy the following expansions:*

$$t(\theta^i) = 1 + (\pi(\theta^i))' S(\theta^i) \cdot \mathbf{1}_K + O(h),\ i = \overline{0,1};$$

$$\alpha = (\pi(\theta^0))' B_{(2)}(\theta^0) + \pi_{K+1}(\theta^0) + O(h), \quad \beta = (\pi(\theta^1))' B_{(1)}(\theta^1) + \pi_0(\theta^1) + O(h).$$

Proof. The approximations are derived from properties of inhomogeneous Markov chains. □

3 Numerical results

The probability model (1) was considered and the hypotheses (2) were tested by (3), (4) with the following values of parameters:

$$m = 4, \sigma = 2, \psi(t) = (1, t/10, t^2/100, t^3/1000)^T, \theta^0 = (1, 2, 3, 0.9)^T, \theta^1 = (1, 1, 1, 1)^T.$$

The infinite sum was limited to 1000 summands. The thresholds C_-, C_+ were calculated according to Wald. Denote the sample estimate of a characteristic γ with Monte-Carlo method by $\hat{\gamma}$. The number of runs used in this method was 100 000. The results of Corollary 2 are given in Table 1, where $t_i = E(N|\theta^i), i = 0, 1$.

α_0	β_0	$\alpha \leq$	$\beta \leq$	$\hat{\alpha}$	$\hat{\beta}$	$E^{(0)}(N) \leq$	$E^{(1)}(N) \leq$	$\hat{t_0}$	$\hat{t_1}$
0.1	0.1	0.0545	0.0545	0.0477	0.0480	12.674	12.674	9.428	9.434
0.05	0.05	0.0230	0.0230	0.0207	0.0216	13.685	13.685	10.275	10.266
0.01	0.01	0.0037	0.0037	0.0034	0.0036	15.359	15.359	11.532	11.538

Table 1. Upper bounds and Monte-Carlo estimates

References

[1] Kharin A. (2002). An approach to performance analysis of the SPRT for simple hypotheses testing. *Proc. of the Belarusian State University*. Vol. **1**, pp. 92-96.

[2] Kharin A. (2008). Robustness evaluation in sequential testing of composite hypotheses. *Austrian Journal of Statistics*. Vol. **37**(1), pp. 51-60.

[3] Kharin A. (2013). *Robustness of Bayesian and Sequential Statistical Decision Rules*. BSU, Minsk.

[4] Kharin A. (2013). Robustness of sequential testing of hypotheses on parameters of M-valued random sequences. *Journal of Mathematical Sciences*. Vol. **189**(6), pp. 924-931.

[5] Kharin A. (2016). Performance and robustness evaluation in sequential hypotheses testing. *Communications in Statistics - Theory and Methods*. Vol. **45**(6), pp. 1693-1709.

[6] Wald A. (1947). *Sequential analysis*. John Wiley and Sons, New York.

LEE DISTANCE IN TWO-SAMPLE RANK TESTS

Nikolay I. Nikolov
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Sofia, BULGARIA
e-mail: n.nikolov@math.bas.bg

Abstract

In nonparametric statistic there are various procedures to constructing rank tests via metrics on the permutation group. In this paper Critchlow's unified approach is applied to Lee distance. The two-sample location problem is considered and the distribution of the test statistic under null hypothesis is derived and studied.

1 Introduction

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent random samples with continuous distribution functions $F(x)$ and $G(x)$, respectively. We consider rank tests for the two-sample location problem of testing the null hypothesis H_0 against the alternative H_1

$$H_0 : F(x) \equiv G(x)$$
$$H_1 : F(x) \geq G(x),$$

with strict inequality for some x. Let $\alpha(i)$ be the rank of X_i for $i = 1, 2, \ldots, m$ and $\alpha(m+j)$ be the rank of Y_j for $j = 1, 2, \ldots, n$ among $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$. Then $\alpha = (\alpha(1), \alpha(2), \ldots, \alpha(m+n))$ is the rank vector of all observations and $\alpha \in S_{m+n}$, where S_{m+n} is the permutation group generated by the first $m+n$ natural integers. The class of permutations, which are most in agreement with the alternative H_1 is $E = S_m \times S_n = \{\pi \in S_{m+n} : \pi(i) \leq m, \forall i \leq m\}$. The left coset $[\alpha] = \alpha(S_m \times S_n) = \{\alpha \circ \pi : \pi \in S_m \times S_n\}$ consists of all permutations in S_{m+n} which are equivalent to α. Many rank statistics could be obtained by using distances between sets of permutation. Critchlow [2] proposed a unified approach to constructing nonparametric tests which produces many well-known rank statistics. The method is based on finding the minimum interpoint distance between the class of equivalence $[\alpha]$ and the extremal set E:

$$d([\alpha], E) = \min_{\substack{\pi \in [\alpha] \\ \sigma \in E}} d(\pi, \sigma), \tag{1}$$

where d is an arbitrary metric on S_{m+n}. The proposed test rejects the null hypothesis H_0 for small values of the statistic $d([\alpha], E)$. This contrasts with the structure of some parametric test, where H_0 is rejected if the distance from H_0 is large. Since the minimal value of the proposed test statistic is zero and $d([\alpha], E) = 0$ if and only if $d(\alpha, \sigma) = 0$ for some $\sigma \in E$, the strongest evidence for rejecting H_0 occurs if and only if the observed permutation α is equivalent to some extremal permutation $\sigma \in E$.

2 Lee distance on S_N

The goal of this paper is to derive and study the rank test statistic in (1) induced by the Lee distance function on S_N:

$$L(a,b) = \sum_{i=1}^{N} \min\left(\mid a(i) - b(i) \mid, N - \mid a(i) - b(i) \mid\right).$$

In nonparametric statistics the right-invariance of a metric is necessary requirement since it means that the distance between rankings does not depend on the labelling of the observations.

Definition 1. The metric d on S_N is called right-invariant, if and only if $d(\alpha,\beta) = d(\alpha\circ\gamma, \beta\circ\gamma)$ for all $\alpha,\beta,\gamma \in S_N$.

Deza and Huang [4] includes extensive discussion of some metrics on the permutation group S_N which are widely used in applied scientific and statistical problems. Critchlow [2] obtained the minimal value defined by (1) for four basic distance functions: Spearman's footrule, Ulam distance, Kendall's tau and Hammning distance, and proved that the induced test statistics are equivalent to some familiar rank statistics. Stoimenova [6] derived the test statistic induced by Chebyshev metric. More properties of these distances can be found in Critchlow [1, 3], Deza [4] and Diaconis [5].

Since $L(a,b)$ is right-invariant it follows

$$\begin{aligned} L([\alpha], E) &= \min_{\substack{\pi\in[\alpha]\\ \sigma\in E}} L(\pi,\sigma) = \min_{\pi\in[\alpha]} L(\pi, e) \\ &= \min_{\pi\in[\alpha]} \left\{ \sum_{i=1}^{m+n} \min\left(\mid a(i) - i \mid, m+n - \mid a(i) - i \mid\right) \right\}, \end{aligned} \tag{2}$$

where $e = (1,2,\ldots,m+n)$ is the identity permutation. After solving the optimal problem (2), $L([\alpha], E)$ can be expressed as

$$\begin{aligned} L([\alpha], E) = 2 \sum_{i\in K_m} \min\big(&\mid \alpha(i) - \gamma_n^{-1}(k+1-\gamma_m(\alpha(i))) \mid, \\ &m+n - \mid \alpha(i) - \gamma_n^{-1}(k+1-\gamma_m(\alpha(i))) \mid\big) \end{aligned} \tag{3}$$

where

$$\begin{aligned} K_m &= \{i \in \{1,2,\ldots,m\} : \alpha(i) > m\} \quad , \\ K_n &= \{i \in \{m+1, m+2, \ldots, m+n\} : \alpha(i) \le m\} \quad , \end{aligned} \tag{4}$$

k is the number of elements of K_m $(k = \mid K_m \mid = \mid K_n \mid)$, $\gamma_m(\alpha(i))$ is the rank of $\alpha(i)$ among $\{\alpha(i) : i \in K_m\}$, $\gamma_n(\alpha(i))$ is the rank of $\alpha(i)$ among $\{\alpha(i) : i \in K_n\}$ and γ^{-1} is the inverse of γ, i.e. $\gamma^{-1}(\gamma(\alpha(i))) = \alpha(i)$. The statistic $L([\alpha], E)$ is equivalent to

$$L := \frac{L([\alpha], E)}{2}. \tag{5}$$

There is an interpretation of the rank test statistic L in terms of graph theory. Let C be a simple cycle graph with vertices $\{i\}_{i=1}^{m+n}$ and edges $\bigcup_{i=1}^{m+n-1}\{i, i+1\}$ and $\{m+n, 1\}$. Then L is the minimum sum of distances over C between the elements of K_m and the elements of K_n. An example when $m = 6$, $n = 4$, $K_m = \{3, 5\}$ and $K_n = \{8, 9\}$ is illustrated on Figure 1. In this case $L = (10 - \mid 3 - 9 \mid) + \mid 5 - 8 \mid = 4 + 3 = 7$.

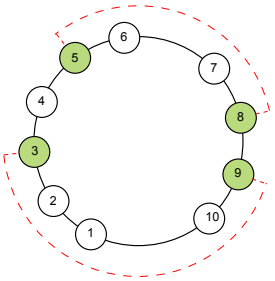


Figure 1: Lee distance on C.

The value of L depends not only on the elements in K_m and K_n, but also on the way in which their elements are paired. Formula (3) gives that the minimal sum of distances between pairwise elements of K_m and K_n is obtained when the smallest element of K_m is combined with the largest element of K_n, the second smallest element of K_m is combined with the second largest element of K_n, ..., the largest element of K_m is combined with the smallest element of K_n. Using this fact the distribution of the test statistic could be calculated for fixed number k of elements in K_m and K_n, $k = \mid K_m \mid = \mid K_n \mid$. Let $[K_m \times K_n]^*$ be the described above set of pairs and $s-1$ be the number of pairs $(x, y) \in [K_m \times K_n]^*$ for which the shortest path on C goes over the edge $\{m, m+1\}$. Obviously, s is between 1 and $k+1$. If for some pair $(x, y) \in [K_m \times K_n]^*$ the paths over $\{m, m+1\}$ and over $\{m+n, 1\}$ are with the same length, then the path over $\{m+n, 1\}$ is considered to be the shortest. For $i = 0, 1, \ldots, s-1$ let $a_i^{(m)}$ be the number of elements in $\{1, 2, \ldots, m\} \backslash K_m$ which are in the shortest path of exactly i pairs $(x, y) \in [K_m \times K_n]^*$ connected by the edge $\{m, m+1\}$. For $j = 1, 2, \ldots, k-s+1$ let $b_j^{(m)}$ be the number of elements in $\{1, 2, \ldots, m\} \backslash K_m$ which are in the shortest path of exactly j pairs $(x, y) \in [K_m \times K_n]^*$ connected by the edge $\{m+n, 1\}$. Similarly the numbers $\{a_i^{(n)}\}_{i=0}^{s-1}$ and $\{b_j^{(n)}\}_{j=1}^{k-s+1}$ are defined for the set $\{m+1, m+2, \ldots, m+n\} \backslash K_n$. An illustration of the used notation is shown on Figure 2.

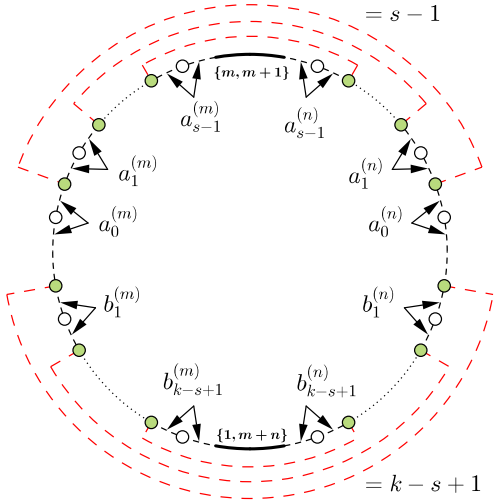


Figure 2: Notations.

For the considered example on Figure 1, $m = 6$, $n = 4$, $[K_m \times K_n]^* = \{(3, 9), (5, 8)\}$, $s = 2$, $a_0^{(m)} = 1 = \mid \{4\} \mid$, $a_1^{(m)} = 1 = \mid \{6\} \mid$, $b_1^{(m)} = 2 = \mid \{1, 2\} \mid$, $a_0^{(n)} = 0$, $a_1^{(n)} = 1 = \mid \{7\} \mid$ and $b_1^{(n)} = 1 = \mid \{10\} \mid$.

Theorem 1. *Let L be defined by (5) and $K = \mid K_m \mid$ be the number of elements of the set K_m, defined by (4). Then the joint distribution of L and K under H_0 is given by*

$$P(L = l, K = k) = \begin{cases} \dfrac{m!n!}{(m+n)!} & , \text{ for } l = 0 \text{ and } k = 0 \\ \displaystyle\sum_{s} \sum_{a,b} \frac{m!n!}{(m+n)!} & , \text{ for } 1 \leq k \leq \min(m, n) \text{ and} \end{cases} \tag{6}$$

$\left[\frac{k^2+1}{2}\right] \leq l \leq \left[\frac{(m+n-k)\,k+1}{2}\right]$, *where* $[x]$ *is the integer part of* x.

The first summation in (6) is taken over all s such that $(s-1)^2+(k-s+1)^2 \leq l$. The second summation is over all nonnegative integers $\{a_i^{(m)}\}_{i=0}^{s-1}$, $\{a_i^{(n)}\}_{i=0}^{s-1}$, $\{b_j^{(m)}\}_{j=1}^{k-s+1}$ and $\{b_j^{(n)}\}_{j=1}^{k-s+1}$ that satisfy:

(i) $\sum_{i=0}^{s-1} a_i^{(m)} + \sum_{j=0}^{k-s+1} b_j^{(m)} = m-k$ **(ii)** $\sum_{i=0}^{s-1} a_i^{(n)} + \sum_{j=0}^{k-s+1} b_j^{(n)} = n-k$

(iii) $l = (s-1)^2 + (k-s+1)^2 + \sum_{i=0}^{s-1} i\left(a_i^{(m)} + a_i^{(n)}\right) + \sum_{j=0}^{k-s+1} j\left(b_j^{(m)} + b_j^{(n)}\right)$

(iv) $2\,(s-1) + \sum_{i=0}^{s-1}\left(a_i^{(m)} + a_i^{(n)}\right) \geq 2\,(k-s) + \sum_{j=0}^{k-s+1}\left(b_j^{(m)} + b_j^{(n)}\right)$, if $s \in \{1,2,\ldots,k\}$

(v) $2\,(s-2) + \sum_{i=1}^{s-1}\left(a_i^{(m)} + a_i^{(n)}\right) < 2\,(k-s+1) + a_0^{(m)} + a_0^{(n)} + \sum_{j=0}^{k-s+1}\left(b_j^{(m)} + b_j^{(n)}\right)$,

if $s \in \{2,3,\ldots,k+1\}$. The indexes $b_0^{(m)}$ and $b_0^{(n)}$ are defined to be zeros, $b_0^{(m)} := 0$, $b_0^{(n)} := 0$, for completeness in conditions **(i)-(v)**.

Given the joint distribution of L and K the marginal distribution of L under H_0 can be easily derived.

Acknowledgements: This work was supported by the grant I02/19 of the Bulgarian National Science Fund.

References

[1] Critchlow D. (1985). *Metric Methods for Analyzing Partially Ranked Data*. Lecture Notes in Statist.,No. **34**, Springer, New York.

[2] Critchlow D. E. (1986). A Unified Approach to Constructing Nonparametric Tests, Tech. Report. No. **86-15**, Dept. of Statistics, Purdue University, Indiana.

[3] Critchlow D. E. (1992). On rank statistics: An Approach via Metrics on the Permutation Group, *J. Statist. Plann. Inference*, Vol. **32**, pp. 325-346.

[4] Deza M., Huang T.(1998). Metrics on Permutations, a Survey, *Journal of Combinatoric, Information and System Sciences*, Vol. **23**, pp. 173-185.

[5] Diaconis P. (1988). *Group Representations in Probability and Statistics*. IMS Lecture Notes - Monograph Series, Vol. **11**, Hayward, Carifornia.

[6] Stoimenova E.(2000). Rank Tests Based on Exceeding Observations, *Ann. Inst. Stat. Math.*, Vol. **52-2**, pp. 255-266.

FORWARD PROJECTION FOR HIGH-DIMENSIONAL DATA

T. Ortner[1], P. Filzmoser[2], S. Brodinova[3],
M. Zaharieva[4], C. Breiteneder[5]
[1,3,4,5]*Interactive Media Systems Group, TU Wien*
[1,2]*Institute of Statistics and Mathematical Methods in Economics, TU Wien*
[4]*Multimedia Information Systems Group, University of Vienna*
Vienna, AUSTRIA
e-mail: [1]thomas.ortner@tuwien.ac.at

Abstract

We provide a novel view on group structure in data. Projecting observations onto a subspace spanned by a small selection of observations, we calculate orthogonal distances as a measure for dissimilarity. Sequentially exchanging the observations, used to span the subspace, we recieve a series of distances. Observations, taken from a similar group structure will behave similar along those projections.

This leads to a visualisation of high dimensional data providing some basic diagnostic on group structures and outliers. The series of distances can be further utilized to perform cluster algorithms, leading to significant improvement when facing clusters located in different subspaces.

ON THE SEQUENTIAL CHI-SQUARE TEST

M. P. Savelov
Lomonosov Moscow State University
Moscow, RUSSIA
e-mail: savelovmp@gmail.com

Abstract

Chi-square test based on the Pearson statistics is used to check whether frequencies of the finite number of outcomes correspond with their hypothetical probabilities. A sequential version of this test is based on several Pearson statistics computed for nested samples; this version was considered in several papers. Formulas for the joint distributions of the Pearson statistics for nested samples are very cumbersome. Here we present exact formulas for the covariance of two Pearson statistics computed for nested samples and asymptotic relations connecting the error probabilities of one- and two-dimensional chi-square tests.

1 Introduction

Suppose that independent identically distributed trials with m outcomes having probabilities $p_1, \dots, p_m$ are performed. Denote by $\nu_i(n)$ the frequency of the j-th outcome in the first n trials. Pearson statistics $\chi^2(n) := \sum_{j=1}^{m} \frac{(\nu_i(n)-np_j)^2}{np_j}$ is widely used to test the hypothesis $H(\mathbf{p})$: "outcomes have law $\mathbf{p} = (p_1, \dots, p_m)$", because $\chi^2(n)$ converges in distribution to the standard χ^2_{m-1} at $n \to \infty$. So, if $\pi_{m-1}(\alpha)$ is the $(1-\alpha)$-quantile of χ^2_{m-1}, then the rule, accepting $H(\mathbf{p})$, if $\chi^2(n) < \pi_{m-1}(\alpha)$, has type I error $\approx \alpha$.

A sequential version of the chi-square test is based on the statistics $(\chi^2(n_1), \dots, \chi^2(n_r))$ for $n_1 < \dots < n_r$; it was studied by Zakharov et al. [5] and others [2, 3]. In the sequential version $H(\mathbf{p})$ is rejected if and only if $\chi^2(n_k) > \pi_{m-1}(\alpha_k)$ for all $k = 1, \dots, r$.

For true hypothesis $H(\mathbf{p})$ we find (Theorems 1, 2) the covariance $\operatorname{cov}(\chi^2(n_1), \chi^2(n_2))$ and an asymptotic relation between error probabilities $\mathbf{P}\{A_1\} \sim \alpha_1$, $\mathbf{P}\{A_2\} \sim \alpha_2$ and $\mathbf{P}\{A_1A_2\}$, $A_j = \{\chi^2(n_j) > \pi_{m-1}(\alpha_j)\}$, as $n_1, n_2 \to \infty$, $n_1/n_2 \to \text{const}$, $\alpha_1, \alpha_2 \to 0$.

2 Main results

Theorem 1. *If $n_1 \le n_2$, then*

$$\operatorname{cov}(\chi^2(n_1), \chi^2(n_2)) = \Big(2(n_1-1)(m-1) - m^2 + \sum\nolimits_{k=1}^{m} p_k^{-1}\Big) \Big/ n_2.$$

Corollary 1. $\operatorname{cov}(\chi^2(n_1), \chi^2(n_2))$ *is nonincreasing in n_2 for fixed $n_1 \le n_2$.*

Note that if $\frac{n_2}{n_1} \to \infty$, then the covariance tends to 0. It follows from the Theorem 1 that the variance $\mathbf{D}\chi^2(n) = 2(m-1) + \frac{1}{n}\left(2 - 2m - m^2 + \sum_{k=1}^{m} \frac{1}{p_k}\right)$, this expressions was obtained in [4].

Zakharov, Sarmanov and Sevastyanov [5] have derived asymptotic formulas for the error probabilities of the sequential chi-square test in the form of integrals containing Infeld functions and exponential functions. By means of these formulas we obtain explicit asymptotic expressions for the error probability in the case $r = 2$.

Theorem 2. *Let* $m \geq 3$, $n_1, n_2 \to \infty$, $\frac{n_1}{n_2} \to c^2$, $c \in (0,1)$, $\alpha_1 := \lim \mathbf{P}\{\chi^2(n_1) > \pi_{m-1}(\alpha_1)\}$, $\alpha_2 := \lim \mathbf{P}\{\chi^2(n_2) > \pi_{m-1}(\alpha_2)\}$ *and* $\alpha = \lim \mathbf{P}\{\chi^2(n_1) > \pi_{m-1}(\alpha_1), \chi^2(n_2) > \pi_{m-1}(\alpha_2)\}$. *If* $\alpha_1, \alpha_2 \to +0$ *and* $\sqrt{\frac{\ln \alpha_2}{\ln \alpha_1}} = P \in (c, \frac{1}{c})$, *then*

$$\alpha = \frac{(1-c^2)^{\frac{3}{2}} \cdot P^{\frac{m}{2}-1} \cdot (-\ln \alpha_1)^{\frac{m}{2}-2} \cdot Q^{-\frac{1}{2(1-c^2)}}}{2c^{m-\frac{1}{2}}\sqrt{\pi}\Gamma\left(\frac{m-1}{2}\right)(P-c)(1-cP)} \cdot (1+o(1)),$$

where

$$Q = \frac{P^{2(m-3)(1-\frac{c}{P})}}{\left(\Gamma\left(\frac{m-1}{2}\right)\right)^{4(1-Pc)}} \cdot (-\ln \alpha_1)^{(m-3)(2-c(P+P^{-1}))} \cdot \alpha_1^{-2(P^2-2Pc+1)}.$$

Corollary 2. *If conditions of the Theorem 2 hold and* $\alpha_2 = \alpha_1$ *(i.e.* $P = 1$*), then*

$$\alpha = \frac{(1-c^2)^{\frac{3}{2}}\left(\Gamma(\frac{m-1}{2})\right)^{\frac{1-c}{1+c}}}{2c^{m-\frac{1}{2}}\sqrt{\pi}(1-c)^2} \cdot \left(-\ln \alpha_1\right)^{\frac{m}{2}-\frac{m-3}{1+c}-2}\left(\alpha_1\right)^{\frac{2}{1+c}}(1+o(1)).$$

Note that $\frac{2}{1+c} \in (1,2)$ since $c \in (0,1)$.

3 Acknowledgments

The author is grateful to A.M. Zubkov for problem statement and constant attention.

References

[1] Germogenov A.P., Ronzhin A.F. (1985). A sequential chi-square test. *Theory Probab. Appl.* Vol. **29**(2), pp. 397–403.

[2] Ronzhin A.F. (1985). The limit distribution for a chi-square process with disorder. *Theory Probab. Appl.* Vol. **29**(3), pp. 613–617.

[3] Selivanov B.I., Chistyakov V.P. (1997). The sequential chi-square test based on s-tuples of states of a Markov chain. *Discrete Math. Appl.* Vol. **7**(5), pp. 523–532.

[4] Tumanyan S.Kh. (1956). Asymptotic distribution of the χ^2 criterion when the number of observations and number of groups increase simultaneously. *Theory Probab. Appl.* Vol. **1**(1), pp. 117–131.

[5] Zakharov V.K., Sarmanov O.V., Sevastyanov B.A. (1969). Sequential χ^2 criteria. *Math. USSR-Sbornik.* Vol. **8**(3), pp. 419–435.

ON APPROXIMATION OF THE Q_n-ESTIMATE OF SCALE BY HIGHLY ROBUST AND EFFICIENT M-ESTIMATES

P. O. Smirnov[1], I. S. Shirokov[2], G. L. Shevlyakov[3]
[1,2,3]*Peter the Great St. Petersburg Polytechnic University*
[3]*Institute for Problems of Mechanical Engineering, Russian Academy of Sciences*
Saint Petersburg, RUSSIA
e-mail: [3]Georgy.Shevlyakov@phmf.spbstu.ru

Abstract

Low-complexity and computationally fast Huber M-estimates of scale are proposed to approximate the highly robust and efficient Q_n-estimate of scale of Rousseeuw and Croux (1993). The parameters of approximation are chosen to provide high robustness and efficiency of the proposed M-estimates of scale at an arbitrary underlying data distribution. A special attention is payed to the particular cases of the Gaussian and Cauchy distributions.

1 Introduction and Problem Set Up

The problem of estimation of a scale parameter is one of most important in statistical analysis. In present, the commonly used highly robust and efficient estimate of scale is given by the Q_n-estimate [4]. This estimate is defined as the first quartile of the distance between observations: $Q_n = c\{|x_i - x_j|\}_{(k)}$, where c is a constant that provides the consistency of estimation, $k = C_h^2$, $h = [n/2] + 1$.

The Q_n-estimate is highly robust with the highest breakdown point $\varepsilon^* = 0.5$ possible and high efficiency 82% at the Gaussian. Its drawback is the high asymptotic computational complexity: generally, it requires $O(n^2)$ of computational time and memory.

On the contrary, Hubers' robust M-estimates of scale are of low-complexity having a potential for enhancing their efficiency. Thus, the main goals of our work are:

1. to construct a low-complexity, computationally fast and highly robust approximation to the Q_n-estimate,

2. to adapt this approximation to data distributions of a general shape.

In what follows, we consider the class of Hubers' M-estimates $\widehat{S}$ of scale given by the implicit estimating equation [3]

$$\sum \chi(x_i/\widehat{S}) = 0, \tag{1}$$

where $\chi(x)$ is a score function commonly even and nondecreasing for $x > 0$.

2 Approximation of the Q_n-estimate: General Case

An important tool for the statistical analysis of estimation in robustness is given by the influence function $IF(x; S, F)$, which defines a measure of the resistance of an estimate functional $S = S(F)$ at a distribution F to gross errors at a point x [2]. Further, the asymptotic variance of the estimate $\widehat{S}$ is given by

$$AV(\widehat{S}, F) = \int IF(x; S, F)^2 \, dF(x) \, .$$

The class of Huber M-estimates of scale (1) has a convenient feature: the influence function $IF(x; S, F)$ is proportional to the score function $\chi(x)$: $IF(x; S, F) \propto \chi(x)$. Thus, it is possible to construct an M-estimate with any admissible influence function, and accordingly, efficiency.

It is known that the influence function of the Q_n-estimate is given by [4]

$$IF(x; Q, F) = c \cdot \left(\frac{1}{4} - F(x + c^{-1}) + F(x - c^{-1}) \right) \Big/ \left(\int f(y + c^{-1}) f(y) dy \right). \quad (2)$$

Since the score χ in Equation (1) is defined up to an arbitrary factor, the normalization integral in the denominator of (2) can be omitted. Then, the Q_n-estimate corresponds to the M-estimate generated by the score function

$$\chi_Q(x) = \frac{c}{4} - c \cdot (F(x + c^{-1}) - F(x - c^{-1})), \quad (3)$$

hence the influence function $IF(x; \chi_Q, F)$ is identical with $IF(x; Q, F)$, ensuring the match of the derivatives of its characteristics.

Now we transform Equation (3). At first, let us make the substitution $\alpha = c^{-1}$, generally not fixing α and considering it as an estimate parameter. Then we expand the distribution function F in a Taylor series, leaving only the first three terms:

$$F(x \pm \alpha) = F(x) \pm \alpha f(x) + \frac{1}{2}\alpha^2 f'(x) \pm \frac{1}{6}\alpha^3 f''(x) + o(\alpha^3). \quad (4)$$

Combination of (3) and (4) leads to the following.

Definition 1. Let f be an analytic probability density function on $\mathbf{R}$. One-parametric family of M-estimates with score functions

$$\chi_\alpha(x) = c_\alpha - 2f(x) - \frac{1}{3}\alpha^2 f''(x), \quad (5)$$

is called f-based MQ_n-family (of f-based MQ_n-estimates). The scalar constant c_α in Equation (5) provides consistency of defined MQ_n-estimates.

3 Approximation of the Q_n-estimate: Gaussian Case

In this section we use the recent results of [5].

Consider the proposed M-estimate in the case of the Gaussian distribution density: $f(x) = \varphi(x) = 2\pi^{-1/2}\exp(-x^2/2)$. Then $\varphi''(x) = (x^2-1)\varphi(x)$, and the score function takes the form

$$\chi_\alpha(x) = c_\alpha - \frac{1}{3}(6 + \alpha^2(x^2-1))\varphi(x), \qquad c_\alpha = \frac{12-\alpha^2}{12\sqrt{\pi}}. \tag{6}$$

In the important special case when $\alpha = 0$, the expression takes the following form

$$\chi_0(x) = \frac{1}{\sqrt{\pi}} - 2\varphi(x). \tag{7}$$

This score is similar to a Welsh generalized error score [1] given by

$$\chi(x) = \sqrt{\frac{d}{d+2}} - \exp\left(-\frac{x^2}{d}\right), \qquad d > 0.$$

For $d = 2$ this score yields the same M-estimate of scale as the score given by (7). The highest possible asymptotic efficiency of estimates defined by (7) is 95.9%.

In the Gaussian case, the following result holds.

Theorem 1. *The Gaussian-based MQ_n-estimates for $\alpha \in [0;\sqrt{2}]$ at the Gaussian distribution are B-robust with the bounded influence function of the form*

$$IF(x; MQ, \Phi) = \frac{2(12-\alpha^2) - 8\sqrt{\pi}(6+\alpha^2(x^2-1))\varphi(x)}{3(4-\alpha^2)}.$$

The asymptotic efficiency of the fast low-complexity MQ_n-estimate with the score function (7) is 81%, just 1% less than that of the Q_n-estimate at the Gaussian.

4 Approximation of the Q_n-estimate: Cauchy Case

Now we consider the Cauchy-based MQ_n-estimates with the score (5) derived from the heavy-tailed Cauchy distribution density

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

For the sake of low-complexity, take the MQ_n-estimate with $\alpha = 0$, as other parameter values are of no interest because of worse performance. In this case, the following result holds.

Theorem 2. *The Cauchy-based MQ_n-estimate with $\alpha = 0$ and score function*

$$\chi_0(x) = \frac{1}{\pi} \cdot \frac{x^2 - 1}{x^2 + 1}$$

coincides with the MLE estimate of scale for the Cauchy distribution.

The highest possible asymptotic efficiency of this estimate of scale is 100% at the Cauchy distribution but the asymptotic relative efficiency at the Gaussian is 50%.

5 Conclusions

1. A class MQ_n of low-complexity, computationally fast and highly robust M-estimates of scale close in efficiency to the highly efficient and robust Q_n-estimate is proposed.

2. The important Gaussian and Cauchy distribution particular cases are thoroughly studied both theoretically in asymptotics and experimentally on small samples—the obtained results confirm effectiveness of the proposed approach.

3. In our talk, we plan to exhibit the theoretical and Monte Carlo results of application of the proposed approach in the parametric families of t- and exponential-power distributions.

References

[1] Genton M. G. (1998). Asymptotic Variance of M-estimators for Dependent Gaussian Random Variables. *Statistics and Probability Letters*. Vol. **38**, pp. 255-261.

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

[3] Huber P. J. (1981). *Robust Statistics*. John Wiley.

[4] Rousseeuw P. J., Croux C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*. Vol. **88**, pp. 1273-1283.

[5] Smirnov P.O., Shevlyakov G.L. (2014). Fast Highly Efficient and Robust One-Step M-Estimators of Scale Based on Qn. *Computational Statistics and Data Analysis*. Vol. **78**, pp. 153-158.

A PAIRWISE LOG-RATIO METHOD FOR THE IDENTIFICATION OF BIOMARKERS

J. Walach[1], P. Filzmoser, K. Hron, B. Walczak
Institute of Statistics and Mathematical Methods in Economics, TU Wien
Vienna, AUSTRIA
e-mail: [1]jan.walach@tuwien.ac.at

Abstract

One of the main goals in metabolomics is the identification of biomarkers metabolites which are capable of distinguishing between groups of, e.g., healthy and unhealthy patients. There are various methods for identifying biomarkers in the statistical field. Difficulties arise by facing the so-called size effect, which occurs due to different sample volume or concentration. In that case, the true signal is hidden in the data structure, and it can be revealed only after a special treatment. One possibility is to normalize the data first, other possibilities include certain transformations, see e.g. [1].

Here we propose a method that makes use of the log-ratio approach [2]. We use the elements of the variation matrix, which are defined as the variance of $\log(x_i/x_j)$, for all pairs of variables x_i and x_j. The advantage of log-ratios is that the absolute concentration is irrelevant, which is appropriate in this context. The variation matrix is computed for the joint data, as well as for the single groups separately. A statistic is then constructed, involving all three sources of information. Since the distribution of the statistic is unknown, we use the bootstrap technique; biomarkers are then considered as variables where most of their pairwise log-ratios are significantly different.

The method has been tested on simulated data as well as on real data sets. The simulations have been carried out according to the scheme outlined in [1]. In both the low-dimensional (9 variables) and the high-dimensional (500 variables) situation, the new proposal shows excellent behavior with respect to the true positives, false discovery and false negative rates. These simulations reveal slight advantages over PQN normalization, the method which turned out in [1] as the best among all considered options. The new pairwise log-ratio method has the big advantage that it can easily be robustified against outliers in the data, by simply using a robust estimator of the variance.

References

[1] Filzmoser P., Walczak B. (2014). What can go wrong at the data normalization step for identification of biomarkers? *J. Chromatography A*, Vol. 1362, p. 194.

[2] Pawlowsky-Glahn V., Egozcue R., Tolosana-Delgado J.J. (2015). *Modeling and Analysis of Compositional Data*, Wiley, Chichester, UK.

ASSIGNMENT OF ARBITRARILY DISTRIBUTED RANDOM SAMPLES TO THE FIXED PROBABILITY DISTRIBUTION AND ITS RISK

E.E. Zhuk[1], D.D. Dus[2]
Belarusian State University
Minsk, BELARUS
e-mail: [1]zhukee@mail.ru, [2]dzianisdus@gmail.com

Abstract

The problem of statistical assignment of arbitrarily distributed random samples to the fixed probability distribution is considered. The decision rule based on the maximum likelihood method is proposed and its efficiency is analytically examined. The case of two samples of the same size and the Fisher model is studied.

1 Introduction

Let $m \geq 2$ random samples $X^{(1)}, \ldots, X^{(m)}$ be determined in the observation space R^N ($N \geq 1$) and the following conditions be satisfied.

1. Each sample $X^{(i)} = \{x_t^{(i)}\}_{t=1}^{n_i}$ consists of independent and identically distributed random vectors $x_t^{(i)} \in R^N$, $t = \overline{1, n_i}$ (n_i is the sample size) with the same probability density $p_i(x)$:

$$p_i(x) \geq 0, \quad x \in R^N : \quad \int_{R^N} p_i(x)dx = 1, \quad i = \overline{1, m}. \tag{1}$$

2. Samples $X^{(1)}, \ldots, X^{(m)}$ are independent in total.

Suppose that all densities $\{p_i(x)\}_{i=1}^m$ from (1) are unknown and distinguished from the fixed probability density function, which is often referred as hypothetical density function [1, 2]:

$$p(x) \geq 0, \quad x \in R^N : \quad \int_{R^N} p(x)dx = 1. \tag{2}$$

The problem is to choose the one of samples $\{X^{(i)}\}_{i=1}^m$ that is closer to the hypothetical density (2) in terms of the distribution simillarity.

Note, that the declared problem differs from so-called "goodness of fit testing" problem [1, 2]: samples $\{X^{(i)}\}_{i=1}^m$ are obtained from corresponding probability densities (1), but not from the hypothetical density (2). Also the problem differs from the classification problem [3, 4]: there is the only one class, determined by the density (2), to which one of samples $\{X^{(i)}\}_{i=1}^m$ should be assigned.

The problem is to construct the decision rule (DR):

$$d = d(X^{(1)}, \ldots, X^{(m)}) \in M, \quad M = \{1, \ldots, m\}, \tag{3}$$

to solve the specified assignment problem.

2 Maximum likelihood method and its risk

As it earlier was proposed in [4], the maximum likelihood method [1, 2, 3, 4] can be used to solve the assignment problem:

$$d = d(X^{(1)}, \ldots, X^{(m)}) = \arg\max_{i \in M} P(X^{(i)}); \tag{4}$$

$$P(X^{(i)}) = \prod_{t=1}^{n_i} p(x_t^{(i)}), \quad i \in M,$$

where $P(X^{(i)})$ is the hypothetical likelihood function [1, 2] evaluated for the sample $X^{(i)}$.

Theorem. *Let the following integrals be finite:*

$$\int_{R^N} |\ln(p(x))| p_i(x) dx < +\infty, \quad i \in M, \tag{5}$$

where $\{p_i(x)\}_{i \in M}$, $p(x)$ *are densities from (1), (2).*

If for values

$$H_i = H(p_i(\cdot), p(\cdot)) = \int_{R^N} \ln(p(x)) p_i(x) dx, \quad i \in M, \tag{6}$$

the condition

$$\exists d^0 \in M : H_{d^0} > H_i, \forall i \neq d^0, \quad i \in M,$$

is satisfied, and all samples $\{X^{(i)}\}_{i=1}^m$ *have the same size:*

$$n_i = n, \quad i \in M, \tag{7}$$

then for the decision rule (4) the following statement is true:

$$d = d(X^{(1)}, \ldots, X^{(m)}) \xrightarrow{a.s.} d^0, \quad n \to +\infty; \tag{8}$$

$$d^0 = \arg\max_{i \in M} H_i.$$

Analytical results described above allow us to introduce the generalization of the traditional risk (like as in [4]) as the measure of efficiency of the decision rule (4):

$$r = r(d(X^{(1)}, \ldots, X^{(m)})) = P\{d(X^{(1)}, \ldots, X^{(m)}) \notin D^0\}; \tag{9}$$

$$D^0 = \{k: \ H_k = \max_{j \in M} H_j\}.$$

Here establishment of set D^0 allows us to deal with the situation when some of values H_i may be the same.

The risk (9) means the probability not to assign to hypothetical distribution (4) those samples of $\{X^{(i)}\}_{i \in M}$ that are closer to (4) in terms of the distribution similarity expressed in values (6).

If all values $\{H_i\}_{i \in M}$ are distinguished then the risk (9) is simplified:

$$r = r(d(X^{(1)}, \ldots, X^{(m)})) = P\{d(X^{(1)}, \ldots, X^{(m)}) \neq d^0\}; \tag{10}$$

$$d^0 = \arg\max_{i \in M} H_i.$$

3 The asymptotical investigation of the risk in the case of two samples of the same size. The Fisher model.

Now let us assume the situation when there are only two ($m = 2$) samples $X^{(1)} = \{x_t^{(1)}\}_{t=1}^n$, $X^{(2)} = \{x_t^{(2)}\}_{t=1}^n$ of the same size ($n_1 = n_2 = n$) given for assignment to the hypothetical distribution (2). Then it becomes possible to rewrite DR (4) in the form:

$$d(X^{(1)}, X^{(2)}) = \begin{cases} 1, & if \quad \overline{\xi}_n(X^{(1)}, X^{(2)}) \leq 0; \\ 2, & if \quad \overline{\xi}_n(X^{(1)}, X^{(2)}) > 0, \end{cases} \tag{11}$$

where

$$\overline{\xi}_n(X^{(1)}, X^{(2)}) = \frac{1}{n} \sum_{t=1}^{n} \ln \frac{p(x_t^{(2)})}{p(x_t^{(1)})} \tag{12}$$

and $p(\cdot)$ is the hypothetical probability density from (2).

Also the risk r (9), (10) of the decision rule (11), (12) takes form:

$$r = \begin{cases} P\{\overline{\xi}_n(X^{(1)}, X^{(2)}) \leq 0\}, & if \quad H_1 < H_2; \\ 1 - P\{\overline{\xi}_n(X^{(1)}, X^{(2)}) \leq 0\}, & if \quad H_1 > H_2; \\ 0, & if \quad H_1 = H_2, \end{cases} \tag{13}$$

where H_1, H_2 are values from (6).

Theorem. *Let us consider the assignment problem of two samples ($m = 2$) of the same size ($n_1 = n_2 = n$) and let the following conditions be true:*

$$G_i = \int_{R^N} (\ln(p(x)))^2 p_i(x) dx < +\infty, \quad G_i - H_i^2 \neq 0, \quad i = 1, 2, \tag{14}$$

where $p_1(\cdot)$, $p_2(\cdot)$ and $p(\cdot)$ are densities from (1), (2).

Then the risk (13) can be calculated asymptotically (assuming $H_1 \neq H_2$):

$$\frac{r}{\tilde{r}} \to 1, \quad n \to +\infty; \quad \tilde{r} = \Phi\left(-\sqrt{n}\frac{|H_1 - H_2|}{\sqrt{G_1 + G_2 - (H_1^2 + H_2^2)}}\right), \tag{15}$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{w^2}{2}\right) dw, \quad z \in R,$$

is the standard Gaussian distribution function.

For further results let us assume that all densities $p_1(\cdot)$, $p_2(\cdot)$ and $p(\cdot)$ are multivariate Gaussian with the same covariance matrix. Such assumption is often used in various applications and it is known as the Fisher model [1, 3, 4]:

$$p_i(x) = n_N(x|\mu_i, \Sigma), \quad i = 1, 2; \tag{16}$$

$$p(x) = n_N(x|\mu, \Sigma);$$

$$n_N(x|\mu, \Sigma) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)\right), \quad x \in R^N,$$

where

$$\mu_i = \int_{R^N} x\, p_i(x) dx, \quad i = 1, 2; \quad \mu = \int_{R^N} x\, p(x) dx$$

are appropriate mathematical mean N-vectors and

$$\Sigma = \mathbf{E}\{(x - \mu_i)(x - \mu_i)' | d^o = i\}, \quad i \in S,$$

is the common non-singular covariance $(N \times N)$-matrix.

Under the Fisher model (16) the asymptotical risk $\tilde{r}$ (15) takes the form:

$$\tilde{r} = \Phi\left(-\sqrt{n}\frac{|\rho^2(\mu, \mu_1) - \rho^2(\mu, \mu_2)|}{2\sqrt{N + \rho^2(\mu, \mu_1) + \rho^2(\mu, \mu_2)}}\right), \tag{17}$$

where $\rho(\mu, \mu_i) = \sqrt{(\mu - \mu_i)'\Sigma^{-1}(\mu - \mu_i)}$ is the Mahalanobis distance [1, 3, 4] between μ and μ_i $(i = 1, 2)$.

References

[1] Kharin Yu.S., Zuev N.M., Zhuk E.E. (2011). *Probability Theory, Mathematical and Applied Statistics*. BSU, Minsk.

[2] Borovkov A.A. (1984). *Mathematical Statistics*. Nauka, Moskow.

[3] Aivazyan S.A., Buchstaber V.M., Yenyukov I.S., Meshalkin L.D. (1989). *Applied Statistics: Classification and Dimensionality Reduction*. Finansy i Statistika, Moskow.

[4] Zhuk E.E. (2013). Assignment of multivariate samples to the fixed classes by the maximum likelihood method and its risk. *Computer Data Analysis and Modeling: Theoretical and Applied Stochastics : Proc. of the Tenth Intern. Conf.* Vol. **1**, pp. 185-188.

Section 2

STATISTICAL ANALYSIS OF TIME SERIES AND SPATIAL DATA

FORECASTING OF REGRESSION MODEL UNDER CLASSIFICATION OF THE DEPENDENT VARIABLE

H. Ageeva
Belarusian State University
Minsk, BELARUS
e-mail: helenaageeva@gmail.com

Abstract

Regression model under classification of the dependent variable is considered. Asymptotic properties of plug-in predictive statistic are obtained.

1 Introduction

In this paper we consider a regression model with incompletely observed dependent variable: instead of its true value we observe only one of the given intervals (classes) in which the true value falls. We denote this type of distortion by classification. Classification is a special case of grouping [2].

In discriminant function analysis [3] we use previous observations to predict the class numbers for a future moment. However, in this paper we give a point prediction for the dependent variable.

2 Regression time series under classification of the dependent variable

Let

$$Y_t = F(X_t; \theta^0) + \xi_t,\ t = 1, \dots, T, \tag{1}$$

be a multiple regression time series defined on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$, where T is the sample size; $\theta^0 = (\theta^0_1, \dots, \theta^0_m)' \in \Theta \subset \mathbb{R}^m$ is the unknown regression vector parameter; $X_t = (X_{t,1}, \dots, X_{t,N})' \in \mathbf{X} \subseteq \mathbb{R}^N$ is the observed N-dimensional vector of predictors; $Y_t \in \mathbb{R}^1$ is the nonobservable dependent variable; $\xi_t \in \mathbb{R}^1$ is the normally distributed random error with mean $\mathbf{E}\{\xi_t\} = 0$ and unknown variance $0 < \mathbf{D}\{\xi_t^2\} = (\sigma^0)^2 < +\infty$; $\{\xi_t\}_{t=1}^n$ are jointly independent. The true model parameter is a composite vector-column $\delta^0 = (\theta^{0\prime}, (\sigma^0)^2)' \in \Xi \subseteq \mathbb{R}^{m+1}$.

Let the set of real numbers $\mathbb{R}$ be divided into K nonintersecting intervals ($2 \le K < +\infty$):

$$A_k = (a_{k-1}, a_k],\ k \in \mathbf{K} = \{1, 2, \dots, K\},\ -\infty = a_0 < a_1 < \dots < a_K = +\infty. \tag{2}$$

This set of intervals defines classification of the dependent variable Y_t:

$$Y_t \text{ belongs to class } \nu_t \in \mathbf{K},\ \text{if } Y_t \in A_{\nu_t}. \tag{3}$$

Instead of exact values of $Y_1, \dots, Y_T$ we observe only corresponding class (interval) numbers $\nu_1, \dots, \nu_T \in \mathbf{K}$. Our aim is to construct a forecast of the dependent variable Y_{T+1} for some future predictor X_{T+1}.

3 Maximum likelihood estimator

Introduce the notation:

$$P(k;\delta,X)=\Phi\left(\frac{a_k-F(X;\theta)}{\sigma}\right)-\Phi\left(\frac{a_{k-1}-F(X;\theta)}{\sigma}\right),$$

where $k\in\mathbf{K}$, $\delta=(\theta',\sigma^2)'\in\Xi$, $X\in\mathbf{X}$, $\Phi(\cdot)$ is the standard normal distribution function. Model assumptions (1), (2), (3) determine the probability distribution of the random observations $\nu_t\in\mathbf{K}$:

$$\mathbf{P}_{X_t,\delta}\{\nu_t=k\}=\mathbf{P}_{X_t,\delta}\{Y_t\in A_k\}=P(k;\delta,X_t),\ t=1,\ldots,T;$$

observations $\{\nu_t\}_{t=1}^n$ are jointly independent.

Lemma 1. *Under model assumptions (1), (2), (3) the log-likelihood function is*

$$l(\delta;H,\mathcal{X})=\sum_{t=1}^{T}\ln\left(\Phi\left(\frac{a_{\nu_t}-F(X_t;\theta)}{\sigma}\right)-\Phi\left(\frac{a_{\nu_t-1}-F(X_t;\theta)}{\sigma}\right)\right),\qquad(4)$$

where $\mathcal{X}=\{X_1,\ldots,X_T\}$ is the experimental design, $H=\{\nu_1,\ldots,\nu_T\}$ is the set of classified observations.

Maximum likelihood estimator (MLE) $\hat{\delta}^T$ of the model parameter δ^0 is determined by maximization of the log-likelihood function (4):

$$\hat{\delta}^T=(\hat{\theta}^T,(\hat{\sigma}^T)^2)':\quad l(\hat{\delta}^T;\mathcal{H},\mathcal{X})=\max_{\delta\in\Xi}l(\delta;\mathcal{H},\mathcal{X}).\qquad(5)$$

The following theorems present asymptotic properties of MLE $\hat{\delta}^T$ [1].

Theorem 1. *Let the following conditions hold:*

SC1. $K>2$.

SC2. Regression coefficient space Θ is a closed bounded subset of $\mathbb{R}^m$; there are known bounds $\bar{\sigma}^2>0$ and $\bar{\bar{\sigma}}^2>0$, that $\bar{\sigma}^2\le(\sigma^0)^2\le\bar{\bar{\sigma}}^2$.

SC3. Regressors space $\mathbf{X}\subseteq\mathbb{R}^N$ is a compact space.

SC4. Function $F(X;\theta)$ is continuous on $\mathbf{X}\times\Theta$.

SC5. For any $\varepsilon>0$ there exists $\gamma=\gamma(\varepsilon)>0$ that the following limit expression

$$\varliminf_{T\to\infty}\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\mathbf{I}_{\{|F(X_t;\theta^0)-F(X_t;\theta)|\ge\gamma\}}=b$$

holds for any $\theta\in\Theta$, $|\theta-\theta^0|\ge\varepsilon$, where $0<b=b(\theta,\theta^0,\gamma,F(\cdot))\le1$, $\mathbf{I}_{\{A\}}$ is the identifier of event A.

Then MLE $\hat{\delta}^T$ is strongly consistent:

$$\hat{\delta}^T \xrightarrow[T\to\infty]{\mathbf{P}=1} \delta^0.$$

Define Fisher information matrix:

$$\Gamma_T(\delta) = \sum_{t=1}^{T} \mathbf{E}_{X_t,\delta^0}\{(\nabla_\delta \ln P(\nu_t;\delta,X_t))(\nabla_\delta \ln P(\nu_t;\delta,X_t)')\}.$$

Theorem 2. *Let the following conditions hold:*

A1. MLE $\hat{\delta}^T$ is a consistent estimator of the parameter vector δ^0.

A2. For any fixed $\delta \in \Xi$ functions $F(X;\theta)$, $\frac{\partial F(X;\theta)}{\partial \theta_i}$, $\frac{\partial^2 F(X;\theta)}{\partial \theta_i \partial \theta_j}$, $\frac{\partial^3 F(X;\theta)}{\partial \theta_i \partial \theta_j \partial \theta_s}$, $i,j,s = 1,\ldots,m$, are bounded on $\mathbf{X}$;

A3. $\bar{\Gamma}_T(\delta^0) = \frac{1}{T}\Gamma_T(\delta)$ is a positive definite matrix: $\bar{\Gamma}_T(\delta^0) \succ 0$.

A4. $\varliminf\limits_{T\to\infty} \left|\bar{\Gamma}_T(\delta^0)\right| = b > 0$.

Then MLE $\hat{\delta}^T$ is asymptotically normal distributed:

$$\mathcal{L}\left\{T^{\frac{1}{2}}(\bar{\Gamma}_T(\delta^0)^{\frac{1}{2}})(\hat{\delta}^T - \delta^0)\right\} \xrightarrow[T\to\infty]{} \mathcal{N}_{m+1}(0_{m+1}, \mathbf{I_{m+1}}).$$

4 Plug-in predictive statistic

Under model assumptions (1), (2), (3) plug-in forecasting statistic is

$$\hat{Y}_{T+1} = F(X_{T+1};\hat{\theta}^T). \tag{6}$$

Let us present Fisher information matrix $\Gamma_T(\delta^0)^{-1}$ in a block form:

$$\Gamma_T(\delta^0)^{-1} = \begin{bmatrix} (\Gamma_T(\delta^0)^{-1})^{(1,1)} & (\Gamma_T(\delta^0)^{-1})^{(1,2)} \\ (\Gamma_T(\delta^0)^{-1})^{(2,1)} & (\Gamma_T(\delta^0)^{-1})^{(2,2)} \end{bmatrix},$$

where dimensions of matrices $(\Gamma_T(\delta^0)^{-1})^{(1,1)}$, $(\Gamma_T(\delta^0)^{-1})^{(1,2)}$, $(\Gamma_T(\delta^0)^{-1})^{(2,1)}$, $(\Gamma_T(\delta^0)^{-1})^{(2,2)}$ are $m\times m$, $m\times 1$, $1\times m$, 1×1 correspondingly.

Theorem 3. *Let MLE $\hat{\delta}^T$ be strongly consistent and asymptotically normal distributed estimation of δ^0 and function $F(X;\theta)$ be twice continuously differentiable with regard to θ. Then forecast (6) is asymptotically unbiased:*

$$\mathbf{E}_{X_{T+1},\delta^0}\{\hat{Y}_{T+1} - Y_{T+1}\} \xrightarrow[T\to\infty]{} 0,$$

and its mean squared risk is

$$R = \mathbf{E}_{X_{T+1},\delta^0}\{(\hat{Y}_{T+1} - Y_{T+1})^2\} \xrightarrow[T\to\infty]{}$$
$$\xrightarrow[T\to\infty]{} (\sigma^0)^2 + (\nabla_\delta F(X_{T+1};\theta^0))'(\Gamma_T(\delta^0)^{-1})^{(1,1)}(\nabla_\delta F(X_{T+1};\theta^0)).$$

5 Computer simulations

Consider regression time series:

$$Y_t = F(X_t;\theta^0) + \xi_t = \theta_1^0 X_{t,1}^{\theta_2^0} X_{t,2}^{\theta_3^0} + \xi_t,\ t = 1,\dots,T.$$

where $\theta^0 = (2.248, 0.404, 0.803)'$, $(\sigma^0)^2 = 1$. Let $K = 3$, $a_0 = -\infty$, $a_1 = 12$, $a_2 = 24$, $a_K = +\infty$ and $\{X_{t,1}, X_{t,2}\}_{t=1}^T$ be an analytical grid on $[0,10] \times [0,10]$. For each T we run $Q = 100$ Monte-Carlo simulations and find forecasts $\hat{Y}_{T+1}^q$, $q = 1, ...Q$, for $X_{T+1} = (11, 11)'$. We estimate mean squared risk using $\hat{R}_1 = \frac{1}{Q}\sum_{q=1}^Q \left(\hat{Y}_{n+1}^q - Y_{n+1}^q\right)^2$ and $\hat{R}_2 = \frac{1}{Q}\sum_{q=1}^Q \Big((\hat{\sigma}^{T,q})^2 + (\nabla_\delta F(X_{T+1};\hat{\theta}^{T,q}))'(\Gamma_T(\hat{\delta}^{T,q})^{-1})^{(1,1)}(\nabla_\delta F(X_{T+1};\hat{\theta}^{T,q}))\Big)$. Simulation results are presented in Figure 1. From the figure we see that mean squared risk converges to $(\sigma^0)^2 = 1$.

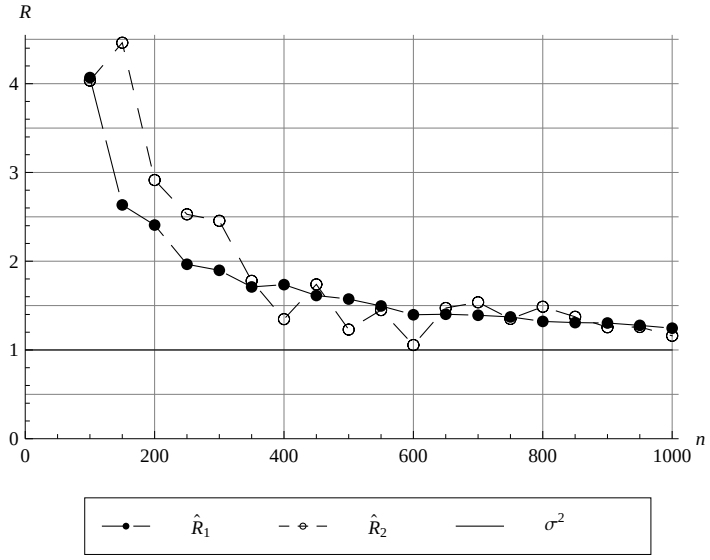


Figure 1: Estimations of squared prediction risks

References

[1] Ageeva H., Kharin Yu. (2015). ML estimation of multiple regression parameters under classification of the dependent variable. *Lithuanian Mathematical Journal*. Vol. **55**(1), pp. 48-60.

[2] Heitjan D.F. (1989). Inference from Grouped Continuous Data: A Review. *Statistical Science*. Vol. **4**(2), pp. 164-183.

[3] McLachlan G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley.

ON PARAMETER ESTIMATION OF STATIONARY GAUSSIAN TIME SERIES OBSERVED UNDER RIGHT CENSORING

I. A. Badziahin
Belarusian State University
Minsk, BELARUS
e-mail: bodiagin@bsu.by

Abstract

Stationary Gaussian time series observed under right censoring are considered. Statistical estimators of the model parameters are constructed by using the method of moments for special auxiliary time series. Consistency of constructed estimators is proved under some additional general conditions.

Consider Gaussian time series x_t observed under right censoring. It means that instead of the exact values $x_1, \dots, x_T$ at the time moments $T_c = \{t : x_t \geq c\}$ only random events are observed [2, 3]:

$$A_t^* = \{x_t \in [c, +\infty)\},\ t \in T_c,$$

where $c \in \mathbb{R}$ is the censoring level, $T \in \mathbb{N}$ is the length of the observation process.

Let $X = (x_1, \dots, x_T)' \in \mathbb{R}^T$ be the vector of the exact observations. Then for Gaussian time series the vector X has a normal distribution $\mathcal{L}(X) = \mathcal{N}(\mu, \Sigma)$, where the mathematical mean μ and the covariance matrix Σ depend on some unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^m$ of the time series model (e.g. for AR(p) model $\theta = (\varphi_1, \dots, \varphi_p, \sigma^2)$, where $\varphi_1, \dots, \varphi_p$ are the autoregression coefficients and σ^2 is the variance of the Gaussian innovation process [1]).

Define the auxiliary time series y_t for the right censored time series x_t [3]:

$$y_t = f_c(x_t) = \begin{cases} x_t, & t \in \{1, \dots, T\} \backslash T_c \\ c, & t \in T_c \end{cases} = \min\{x_t, c\}.$$

Using the method of moments for auxiliary time series y_t, the m values of the second moments $\sigma_\tau = \mathbb{E}\{x_t x_{t+\tau}\}$ for the initial time series x_t can be estimated, i.e. estimators $\hat{\sigma}_\tau$, $0 \leq \tau < m$, can be found. These $\hat{\sigma}_\tau$ with help of the method of moments for initial time series x_t allow to obtain estimators of the model parameters $\hat{\theta}$. The example of this estimation procedure is proposed for the AR(p) model.

The consistency of the constructed estimators $\hat{\theta}$ is proved.

References

[1] Anderson T.W. (1971). *The Statistical Analysis of Time Series*. Wiley, NY.

[2] Kharin Yu.S. (2013). *Robustness in Statistical Forecasting*. Springer, NY.

[3] Park J.W., Genton M.G., Ghosh S.K. (2007). Censored time series analysis with autoregressive moving average models. *Canadian J. Stat.* Vol. **35**(1), pp. 151-168.

ANALYSIS OF SELF-SIMILARITY PROPERTY OF α-STABLE PROCESSES

A. G. Baranovskiy[1], N. N. Troush[2]
Belarusian State University
Minsk, BELARUS
Siedlce University of Natural Sciences and Humanities
Siedlce, POLAND
e-mail: [1]artyom.baranovskiy@gmail.com, [2]TroushNN@bsu.by

Abstract

Self-similarity of network traffic has a strong impact on performance of the network and is a common property of modern telecommunication networks, which makes the study relevant to the industry needs [1].

One of the first steps in modelling network traffic with α-stable processes is research of self-similarity property to find out if the considered process has the long range dependency property. One of the most important parameters related to the self-similarity property is Hurst exponent. There are multiple methods of estimation of the Hurst exponent from the existing data. This work is aimed to compare statistical properties of the most popular estimation techniques applied to α-stable processes.

The aim of the work is to estimate Hurst exponent H from samples of α-stable processes and to perform comparative analysis of statistical properties of the estimates, retrieved using different methods. The following methods are covered in this paper: R/S analysis, variance-time analysis, wavelet analysis and detrended fluctuation analysis. Stable process is considered as a model random process with fractal properties.

The results of the modelling α-stable processes with Hurst exponent H are presented in the work, varying $0.5 < H < 1$. For the retrieved sample the H is estimated using considered methods. The theoretical and empirical research of the statistical properties of the estimates is performed, in particular bias, standard deviation and consistency of the estimates [2].

1 Introduction

Self-similar processes play an important role in probability because of their connection to limit theorems and they are widely used to model natural phenomena. For instance, persistent phenomena in internet traffic, hydrology, geophysics or financial markets are known to be self-similar. Stable processes have attracted growing interest in recent years: data with "heavy tails" have been collected in fields as diverse as economics, telecommunications, hydrology and physics of condensed matter, which suggests using non-Gaussian stable processes as possible models. Self-similar α-stable processes have been proposed to model some natural phenomena with heavy tails.

The stochastic process $X(t)$ is statistically self-similar if $x(at) \stackrel{d}{=} a^H X(at)$, where $a > 0$. Long-range dependence means slow (hyperbolic) decay in the time of the

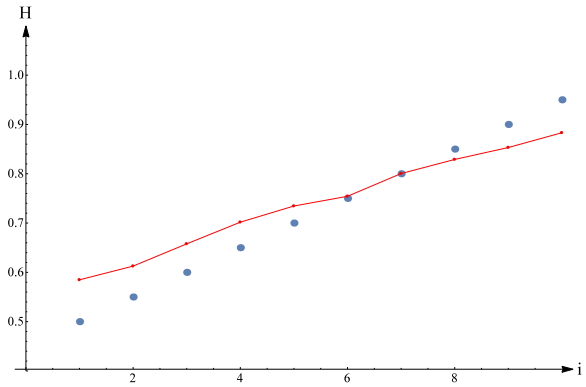


Figure 1: Hurst exponent estimates using R/S analysis

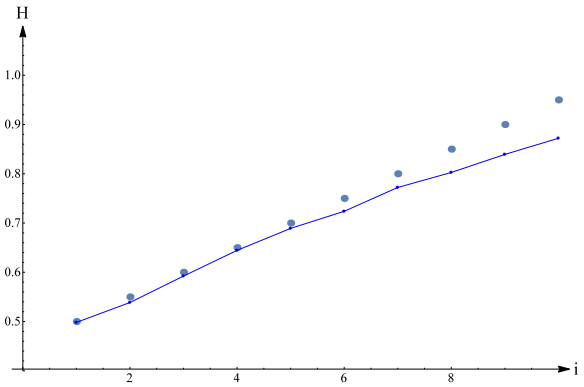


Figure 2: Hurst exponent estimates using variance-time analysis

autocorrelation function of a process. The parameter H (in general $0 < H < 1$) is called the Hurst exponent and is a measure of self-similarity or a measure of duration of long-range dependence of a stochastic process.

Stable distributions are a class of probability laws, and they have intriguing theoretical and practical features. The α-stable distributions are quite effective in the analysis of the financial time series because they can generalize the normal distribution and allow heavy tails and skewness. Despite the fact that the student-t, hyperbolic and normal inverse Gaussian distributions have heavy tail features, the most important reason for preferring the α-stable distributions is that they are supported by the generalized Central Limit Theorem. There is no a close form of α-stable distribution except for Normal, Cauchy and Levy distributions. However, one dimensional stable distribution can be described by the following characteristic function of X $S_\alpha(\beta, \gamma, \sigma)$:

$$\phi(t) = \begin{cases} \exp\left\{-\sigma^\alpha |t|^\alpha \left[1 - i\beta \mathrm{sgn}(t) \tan(\frac{\pi\alpha}{2})\right] + i\mu t\right\} & \text{if } \alpha \neq 1 \\ \exp\left\{-\sigma|t|\left[1 + i\beta \mathrm{sgn}(t)\left(\frac{2}{\pi}\right) \log|t|\right] + i\mu t\right\} & \text{if } \alpha = 1 \end{cases}$$

where $0 < \alpha \leq 2$, $-1 \geq \beta \leq 1$, $\mu, \sigma \in R$, $\sigma > 0$.

In accordance with the fractional Brownian motion, the non-integer alphas in the range $1 < \alpha \leq 2$ are described via long memory and statistical self-similarity properties; these are fractals. Additionally, α is the fractal dimension of the probability space of the time series and can be shown as $\alpha = \frac{1}{H}$, where H is the Hurst exponent and measures the statistical self-similarity.

2 Estimation of Hurst exponent

The estimation is performed using the following approach: at first 100 samples from stable process of size 1024 is generated for each of the considered H values. After that the estimation of Hurst exponent $\widehat{H}$ is performed for the samples using each method, the mean and standard deviation of estimates are calculated.

R/S analysis. This empirical method suggested by G. Hurst is still one of the most popular methods of research of fractal series of different nature.

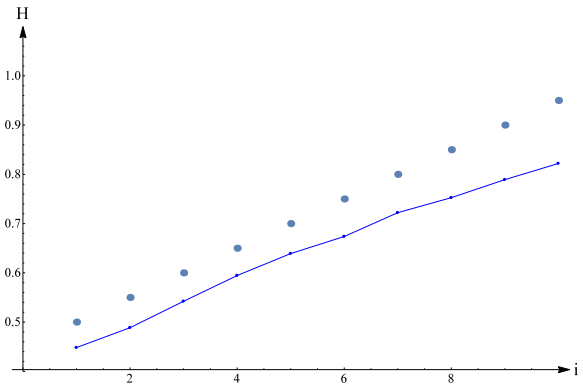


Figure 3: Hurst exponent estimates using detrended fluctuation analysis

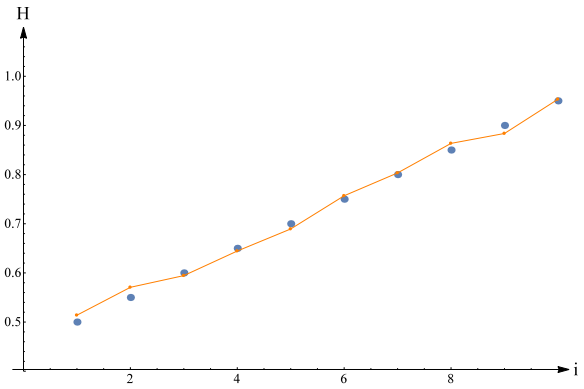


Figure 4: Hurst exponent estimates using wavelet analysis

Table 1: Standard deviation of estimates of Hurst exponent

Estimation method	Standard deviation $S_{\widehat{H}}$	Depends on real H
R/S analysis	$0.029 \leq S_{\widehat{H}} \leq 0.044$	Increases with H
Variance-time analysis	$S_{\widehat{H}} \approx 0.03$	No
Detrended fluctuation analysis	$0.028 \leq S_{\widehat{H}} \leq 0.041$	Increases with H
Wavelet analysis	$S_{\widehat{H}} \approx 0.04$	No

The estimates for Hurst exponent using R/S analysis are biased (see Figure 1) and their standard deviation increases with real H values (see Table 1).

Variance-time analysis is most often used to processes researches in telecommunication networks.

The estimates for Hurst exponent using variance-time analysis are biased (see Figure 2) and their standard deviation does not depend on real H values (see Table 1).

Detrended fluctuation analysis (DFA). DFA is the main method of determining self-similarity for nonstationary time series nowadays. This method is based on the ideology of onedimensional random walks.

The estimates for Hurst exponent using detrended fluctuation analysis are biased (see Figure 3) and their standard deviation increases with real H values (see Table 1).

Wavelet-based estimation. Of recent, the effective tool for a time series analysis is the multiresolution wavelet analysis, whichs main idea consists in the expansion of a time series on an orthogonal base, formed by shifts and the multiresolution copies of the wavelet function.

The estimates for Hurst exponent using wavelet analysis are unbiased (see Figure 4) and their standard deviation does not depend on real H values (see Table 1).

3 Conclusion

The best estimates of Hurst exponent for samples from alpha-stable processes can be retrieved using wavelet analysis, which is confirmed by their theoretical statistical properties [3]. The estimates retrived using this method are unbiased and their standard deviation does not depend on real H values. Estimates retrived using the other methods are biased, which is confirmed empirically. Additionally, the standard deviation of estimates decreases with the increase of sample size for all methods. In case of small sample size to achieve better estimation accuracy the one may want to use the average of fixed unbiased estimates retrieved using multiple methods.

References

[1] Baranovskiy A.G., Troush N.N. (2015) *Modelling network traffic using stable processes*. Collection of the international schientific conference "Theory of probabilities, stochastic processes, mathematical statistics and applications", RIVSH, Minsk (in Russian).

[2] Troush N.N. (1999) *Asymptotic methods of statististical analysis of time series*. BSU, Minsk (in Russian).

[3] Abry P., Delbeke L., Flandrin P. (1999). Wavelet based estimator for the self-similarity parameter of α-stable processes. *Acoustics, Speech, and Signal Processing, 1999. Proceedings.*. Vol. **3**, pp. 1729-1732.

ASYMPTOTIC OPTIMALITY OF THE CHI-SQUARE TEST IN THE CLASS OF PERMUTATION-INVARIANT TESTS

D. M. Chibisov
Steklov Mathematical Institute, Russian Academy of Sciences
Moscow, RUSSIA
e-mail: chibisov@mi.ras.ru

Abstract

Initially, our aim was to prove that in testing the hypothesis H_0 about uniform distribution on $[0,1]$ based on the frequencies $\nu_1,\dots,\nu_N$ of falling the n observations into N equiprobable (under H_0) subintervals of $[0,1]$ the chi-squared test based on $\sum \nu_i^2$ is asymptotically optimal (as $N,n\to\infty$) in the class of symmetric (permutation-invariant) tests. This is the most natural class of tests in the absence of specific alternatives to H_0.

We succeeded in solving the simpler problem that obtains by first taking the limit as $n\to\infty$. Then the frequencies, properly centered and normalized, turn into normally distributed r.v.'s $x_{Ni}\sim\mathcal{N}(\mu_{Ni},1)$, $i=1,\dots,N$, and the hypothesis becomes $H_0:\mu_{Ni}=0$, $i=1,\dots,N$. The r.v.'s x_{Ni} are assumed independent. The constraint $\sum\nu_i=n$ implies the condition $\sum x_{Ni}=0$, which is, however, taken into account by the resulting statistic Z_N^2.

It was proved that for any sequence of alternatives $\boldsymbol{\mu}_N=(\mu_{Ni})_{i=1}^N$, such that $\sum\mu_{Ni}=0$, $\|\boldsymbol{\mu}_N\|=(\sum\mu_{Ni}^2)^{1/2}=O(N^{1/4})$ and $(\mu_{Ni})_{i=1}^N$ satisfy a certain uniform negligibility condition, the sequence of tests based on the statistics $Z_N^2=\sum_{i=1}^N(x_{Ni}-\bar{x}_N)^2$ is asymptotically most powerful within the class of tests symmetric w.r.t. the ordering of components $(x_{Ni})_{i=1}^N$.

The alternatives of order $\|\boldsymbol{\mu}_N\|\asymp N^{1/4}$ are those for which the test achieves a nontrivial power.

EXPECTED ERROR RATES IN CLASSIFICATION OF GAUSSIAN CAR OBSERVATIONS

K. Ducinskas[1], L. Dreiziene[2]
[1,2]*Klaipeda University*
[2] *Vilnius University*
[1,2]*Klaipeda and* [2] *Vilnius, LITHUANIA*
e-mail: [1]kestutis.ducinskas@ku.lt, [2]l.dreiziene@gmail.com

Abstract

Given the neighbourhood structure, the problem of classifying a scalar Gaussian CAR observation into one of two populations specified by different parametric drifts is considered. This paper concerns with classification procedures associated with a parametric plug-in conditional Bayes rule (PBR) obtained by substituting the unknown parameters by their maximum likelihood (ML) estimators in the Bayes rule. For the particular prior distributions of unknown parameters the Bayesian estimators are used. The closed-form expression for the actual error rate associated with aforementioned classification rule and the approximation of the expected error rate (AER) associated with aforementioned PBR is derived. This is the extension of the previous one to the case of complete parametric uncertainty, i.e. when all drift and covariance function parameters are unknown. CAR observations sampled on regular 2-dimensional lattice with respect to the neighbourhood structure based on Euclidean distance between sites is used for simulation experiment.

1 Introduction

Suppose that model of observation $Z(s)$ in population Ω_j is

$$Z(s) = x'(s)\beta_j + \varepsilon(s), \tag{1}$$

where $x(s)$ is a $q \times 1$ vector of non random regressors and β_j is the $q \times 1$ vector of parameters, $j = 1, 2$. The error term $\varepsilon(s)$ is generated by zero-mean CAR $\{\varepsilon(s) : s \in D\}$ with respect to the undirected graph (nodes with neighbourhood system) that will be described later. For given training sample, consider the problem of classification of the $Z_0 = Z(s_0)$ into one of two populations when $x'(s_0)\beta_1 \neq x'(s_0)\beta_2, s_0 \in D$.

Suppose that the set of spatial locations $\{s_i \in D;\ i = 1, ..., n\}$ forming regular or irregular lattice where training sample $T' = (Z(s_1), ..., Z(s_n))$ is taken, and call it the set of training locations. Indexing spatial locations by integers $0, 1, .., n$, denote lattices by $S_n = 1, ..., n$ and $S_n^0 = S_n \cup \{0\}$. Let $Z(s_i) = Z_i, i = 0, ..., n$ then training sample is defined by $T = (Z_1, ..., Z_n)'$ and $T_0 = (Z_0, Z_1, ..., Z_n)'$. Assume that S_n is partitioned into union of two disjoint subsets, i.e. $S_n = S^{(1)} \cup S^{(2)}$, where $S^{(j)}$ is the subset of S_n that contains n_j locations of feature observations from $\Omega_j, j = 1, 2$.

Assume that lattice S_n^0 is endowed with a neighbourhood system $N_0 = \{N_k : k = 0, 1, ..., n\}$ and lattice S_n is endowed with a neighbourhood system $N = \{N_k : k =$

$1, ..., n\}$ where N_k denotes the collection of sites that are neighbours of site, s_k. So we formed undirected graph G^0 specified by lattice S_n^0 and neigbourhood system N^0 and its subgraph G specified by lattice S_n and neigbourhood system N. Define spatial weight $w_{kl} > 0$ as a measure of similarity between sites k and l, and put $w_{kl} = w_{lk}$ and $h_k = \sum_{l \in N_k} w_{kl}$.

The $n \times 2q$ design matrix of training sample T is denoted by X. Then training sample T would be modeled by the joint distribution (Oliveira and Ferreira, 2011)

$$T \sim N_n(X\beta, \sigma^2 V(\alpha)) \tag{2}$$

where

$$V(\alpha) = (I_n + \alpha H)^{-1} \tag{3}$$

and $\sigma > 0$ is a scale parameter and $\alpha \geq 0$ is a spatial dependence parameter and $n \times n$ matrix $H = (h_{kl} : k, l = 1, ..., n)$ is given by

$$h_{kl} = \begin{cases} h_k \;\; if \;\; k = l \\ -w_{kl} \;\; if \;\; k \in N_l. \\ 0 \;\; otherwise \end{cases}$$

In the following set $\Sigma = \sigma^2 V(\alpha)$. Then the variance-covariance matrix of vector T_0 is $var(T_0) = \sigma^2(I_{n+1} + \alpha H^0)^{-1}$ where $H^0 = (h_{kl}, k, l = 0, 1, ..., n)$.

This is the case, when spatial classified training data are collected at fixed locations. Let t denote the realization of T. Set $k = 1 + \alpha h_0$.

Since Z_0 follows model specified in (1)-(3), the conditional distribution of Z_0 given $T = t, \Omega_j$ is Gaussian with mean

$$\mu_{jt}^0 = E(Z_0|T = t; \Omega_j) = x_0'\beta_j + \alpha_0'(t - X\beta), j = 1, 2 \tag{4}$$

and variance

$$\sigma_0^2 = var(Z_0|T = t; \Omega_j) = \sigma^2/k, \tag{5}$$

where $\alpha_0' = \alpha w_0'/k$ and $w_0' = (w_{01}, ..., w_{0n})$.

Under the assumption of complete parametric certainty of populations, the Bayes discriminant fuction (BDF) minimizing the overall misclassification probability (OMP) is specified by (McLachlan, 2004)

$$W_t(Z_0, \Psi) = \Big(Z_0 - \frac{1}{2}(\mu_{1t}^0 + \mu_{2t}^0)\Big)(\mu_{1t}^0 - \mu_{2t}^0)/\sigma_0^2 + \gamma, \tag{6}$$

where $\gamma = ln(\pi_1/\pi_2)$ and $\Psi = (\beta', \theta')'$, $\theta' = (\alpha, \sigma^2)$. Here $\pi_1, \pi_2 (\pi_1 + \pi_2 = 1)$ are respectively prior probabilities of the populations Ω_1 and Ω_2, for observation at location s_0.

The squared Mahalanobis distance between conditional distributions of Z_0 given $T = t$ is specified by

$$\Delta_0^2 = (\mu_{1t}^0 - \mu_{2t}^0)^2/\sigma_0^2 = x_0'(\beta_1 - \beta_2)k/\sigma^2.$$

Denote by $P(\Psi)$ the OMP for BDF defined in (6). Then using the properties of normal distribution we obtain

$$P(\Psi) = \sum_{j=1}^{2} \Big(\pi_j \Phi(-\Delta_0/2 + (-1)^j \gamma/\Delta_0)\Big), \tag{7}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

2 Error rates for plug-in BDF

As it follows we shall write hat above parameters for their estimators based on realization of training sample $T = t$. Put $\hat{\Psi} = (\hat{\beta}', \hat{\theta}')'$ and $\hat{\theta} = (\hat{\alpha}, \hat{\sigma}^2)$, $\hat{\alpha}_0 = \hat{\alpha} w_0/(1 + \hat{\alpha} h_0)$. Then by using (4), (5) we get the estimators of conditional mean and conditional variance

$$\hat{\mu}_{jt}^0 = x_0' \hat{\beta}_j + \hat{\alpha}_0'(t - X\beta), j = 1, 2$$
$$\hat{\sigma}_0^2 = \hat{\sigma}^2/(1 + \hat{\alpha} h_0).$$

Then replacing parameters with their estimators in (6) we form the plug-in BDF

$$W_t(Z_0; \hat{\Psi}) = \Big(Z_0 - \hat{\alpha}_0'(t - X\hat{\beta}) - \frac{1}{2} x_0' I^+ \hat{\beta}\Big)(x_0' I^- \hat{\beta})/\hat{\sigma}_2 + \gamma \tag{8}$$

with $I^+ = (I_q, I_q)$ and $I^- = (I_q, -I_q)$, where I_q denotes the identity matrix of order q.

Lemma 1. *The actual error rate for $W_t(Z_0; \hat{\Psi})$ specified in (8) is*

$$P(\hat{\Psi}) = \sum_{j=1}^{2} \Big(\pi_j \Phi(\hat{Q}_j)\Big). \tag{9}$$

Here

$$\hat{Q}_j = (-1)^j \Big((a_j - \hat{b}) sgn(x_0' I^- \hat{\beta})/\sigma_0 + \hat{\sigma}_0^2 \gamma/(\sigma_0 |x_0' I^- \hat{\beta}|)\Big),$$

where for $j = 1, 2$

$$a_j = x_0' \beta_j + \alpha_0'(t - X\beta) \quad and \quad \hat{b} = \hat{\alpha}_0'(t - X\hat{\beta}) + x_0' I^+ \hat{\beta}/2.$$

Definition 1. The expectation of the actual risk with respect to the distribution of T is called the expected error rate (EER) and is designated as $E_T\Big(P(\hat{\Psi})\Big)$.

We will use the ML estimators of parameters based on training sample. The asymptotic properties of ML estimators established by Mardia and Marshall (1984) under increasing domain asymptotic framework and subject to some regularity conditions are essentially exploited. Hence, the ML estimator $\hat{\Psi}$ is weakly consistent and asymptotically Gaussian, i.e.

$$\hat{\Psi} \sim AN(\Psi, J^{-1}),$$

here the expected information matrix is given by $J = J_\beta \oplus J_\theta$, where $J_\beta = X'\Sigma^{-1}X$ and $(i,j) - th$ element of J_θ is $tr(\Sigma^{-1}\Sigma_i\Sigma^{-1}\Sigma_j)/2$.

Henceforth, denote by (MM) conditions the regularity conditions of Theorem 1 from Mardia and Marshall (1984) and make the following assumption:
(A1) training sample T and estimator $\hat{\theta}$ are statistically independent.

Theorem 1. *Suppose that observation Z_0 to be classified by plug-in PDF and let conditions (MM) and assumption (A1) hold. Then the approximation of EER is*

$$AER = R(\Psi) + \pi_1^* \varphi(-\Delta_0/2 - \gamma/\Delta_0)\Delta_0(K_\beta + K_\alpha + \gamma^2 K_\theta/\Delta_0^2)/2. \tag{10}$$

Here

$$K_\beta = \Lambda' V_\beta \Lambda k,$$
$$\Lambda' = \alpha w_0' X/k - x_0'(I^+/2 + \gamma I^-/\Delta_0^2),$$
$$V_\beta = (X'(I + \alpha H)X)^{-1},$$
$$K_\alpha = w_0'(I + \alpha H)^{-1} w_0 J_{11}^{-1}/k^3,$$
$$K_\theta = \nu' J_\theta^{-1}\nu/k2 \quad where \quad \nu' = (h_0, -1/\sigma_0^2).$$

3 Numerical experiment

In order to demonstrate the results of Theorem 1 simulation experiment was carried out. CAR observations were sampled on regular 2-dimensional lattice with respect to the neighbourhood structure based on Euclidean distance between sites. AER and $P(\hat{\Psi})$ were calculated for different parametric structures. The results of the numerical analysis show that proposed error rates and its approximation formulas could be used as performance evaluation of classification procedures.

References

[1] Ducinskas K., Borisenko I., Simkiene I. (2013). Statistical classification of Gaussian spatial data generated by conditional autoregressive model. *Computational Science and Techniques*. Vol. **1(2)**, pp. 69-79.

[2] Mardia K.V., Marshall R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*. Vol. **71**, pp. 135-146.

[3] McLachlan G.J. (2004). *Discriminant analysis and statistical pattern recognition*. Wiley, New York.

[4] de Oliveira V., Ferreira M.A.R. (2011). Maximum likelihood and restricted maximum likelihood estimation for class of Gaussian Markov random fields. *Metrika*. Vol. **74(2)**, pp. 167-183.

EVALUATION OF EXPECTATION OF A CLASS OF POISSON FUNCTIONALS[1]

A. D. Egorov
Institute of Mathematics, National Academy of Sciences of Belarus
Minsk, BELARUS
e-mail: egorov@im.bas-net.by

Abstract

An approach to evaluation of linear random functionals of stochastic process defined on the probability space generated by Poisson process is suggested.

1 Introduction

In [1] an approach to approximate evaluation of mathematical expectation of a class of nonlinear random functionals based on using the chaos Wiener expansion was suggested. In this report we expand the approach to functionals defined on the probability space generated by Poisson process. Let $P(t) = P(t,\omega)$, $t \in [0,T]$, where it is possible $T = \infty$, be centered Poisson process defined by its characteristic functional

$$\chi_P(\xi) = E[\exp\{i\langle \xi, P\rangle\}] = \exp\Big\{\lambda \int_0^T \Lambda(i\xi(t))dt\Big\},\ \Lambda(x) ::= e^x - x - 1,$$

where the integral $\langle \xi, P\rangle = \int_0^T \xi(t)\, dP(t)$ is defined for $\xi(t) \in L_2[0,T] \cap L_1[0,T]$ as stochastic integral. Let us denote : $e^{\langle \eta, P\rangle} := e^{\langle \eta, P\rangle}\chi^{-1}(-i\eta)$, $G(P;\eta) =: e^{\langle \ln(1+\eta), P\rangle}$: . The functional $G(P;\eta)$ is generating functional of Charlier polynomials [2]- [4]:

$$C_n(P;\eta_1,\dots,\eta_n) = \frac{\partial^n}{\partial\lambda_1\cdots\partial\lambda_n} G\Big(P;\sum_{j=1}^n \lambda_j\eta_j\Big)\Big|_{\lambda_1=\dots=\lambda_n=0};$$

$$E[C_n(P;\varsigma_1,\dots,\varsigma_n)C_n(P;\xi_1,\dots,\xi_n)] = \sum_{(k_1,\dots,k_n)} \prod_{j=1}^n \langle \varsigma_j, \xi_{k_j}\rangle,$$

where $(k_1,\dots,k_n)$ runs through the set of all permutations of $\{1,\dots,n\}$; the set $\{C_\alpha(P_\alpha) \equiv (n_1!\cdots n_k!)^{-1/2} C_{n_1,\dots,n_k}(P;\eta_1,\dots,\eta_k),\ n_1+\cdots+n_k = n,\ n, n_j \in \mathbb{N}\}$ forms full orthonormal system in $L_2(\Omega, P)$,

$$C_{n_1,\dots,n_k}(P;\eta_1,\dots,\eta_k)\} = C_n\Big(P;\underbrace{\eta_1,\dots,\eta_1}_{n_1},\dots,\underbrace{\eta_k,\dots,\eta_k}_{n_k}\Big);$$

$\eta_1,\dots,\eta_k$ are the elements of full orthonormal system in $L_2[0,T]$; $\alpha \in J$, J is the set of multi-index defined by right part of the identity (see [2, 3]).

[1]Supported by Belarusian Republican Foundation for Fundamental Research (project F14D-002).

2 The results

Let us consider linear functionals with random coefficients defined on Poisson process: $F\left(\omega, X_{(\cdot)}\right) = \int\limits_0^T a(s, P_{(\cdot)}) X_s(P_{(\cdot)}) ds$, where the functionals $a(s, P_{(\cdot)})$ and $X_s(P_{(\cdot)})$ admit chaotic developments with respect to Charlier polynomials:

$$a(s, P_{(\cdot)}) = \sum_{\alpha \in J} c_\alpha(s) C_\alpha(P_{(\cdot)}),\ X(P_{(\cdot)}) = \sum_{\alpha \in J} A_\alpha(s) C_\alpha(P_{(\cdot)}),$$
$$c_\alpha(s) = E[a(s, P_{(\cdot)}) C_\alpha(P_{(\cdot)})],\ A_\alpha(s) = E[X_s(P_{(\cdot)}) C_\alpha(P_{(\cdot)})].$$

In this case

$$E[F(\omega, X_{(\cdot)})] = \int\limits_0^T E[a(s, P_{(\cdot)}) X_s(P_{(\cdot)})] ds = \sum_{\alpha \in J} \int\limits_0^T c_\alpha(s) A_\alpha(s) ds,$$

and evaluation of mathematical expectation of given functional reduces to evaluation of coefficients $c_\alpha(s)$, $A_\alpha(s)$ and usual integrals. The approach was used in [1] for the case of functionals defined on probability space generated by Wiener process. In important case when X_s is the solution of stochastic differential equation the coefficients $A_\alpha(s)$ can be evaluated by solution of deterministic equations which one gets after applying the Galerkin method to stochastic differential equations using the Poisson chaos development or can be evaluated exactly. In the last case one can use approximate formulas for evaluation $A_\alpha(s)$ [5]. So it is important in some cases calculate $c_\alpha(s)$ exactly. In this report two cases are considered when $a(s, P_{(\cdot)})$ is Fourier transformation of centered Poisson process and homogeneous polynomial of arbitrary degree from linear functional. We will consider the case $n_1 = \ldots = n_k = 1$ for the simplicity.

1. Let $a(s, P_{(\cdot)})$ be given by its Fourier transform $a(s, P_s) = \int\limits_R \hat{a}(s, u) \exp\{iuP_s\}$, then $c_{\eta_1, \ldots, \eta_n}(s) = \int\limits_R \hat{a}(s, u) E[\exp\{iuP_s\} C_n(P; \eta_1, \ldots, \eta_n)] du$. First let us evaluate

$$I(\lambda_1, \ldots, \lambda_n) \equiv E\Big[\exp\{iuP_s\} : \exp\Big\{ \Big\langle \ln(1 + \sum_{j=1}^n \lambda_j \eta_j), P \Big\rangle \Big\} : \Big] =$$

$$\exp\Big\{ \lambda \int\limits_0^T \Big(\ln\Big(1 + \sum_{j=1}^n \lambda_j \eta_j(t)\Big) - \sum_{j=1}^n \lambda_j \eta_j(t) \Big) dt \Big\} \times$$

$$E\Big[\exp\Big\{ \Big\langle iu 1_{[0,s)} + \ln\Big(1 + \sum_{j=1}^n \lambda_j \eta_j\Big), P \Big\rangle \Big\} \Big] =$$

$$\exp\Big\{ \lambda \int\limits_0^T \Lambda(iu 1_{[0,s]}(t)) dt \Big\} \exp\Big\{ \lambda \int\limits_0^T \Big(e^{iu 1_{[0,s]}(t)} - 1 \Big) \Big(\sum_{j=1}^n \lambda_j \eta_j(t) \Big) dt \Big\}.$$

Then using

$$E[\exp\{iuP_s\} C_n(P; \eta_1, \ldots, \eta_n)] = \frac{\partial^n}{\partial \lambda_1 \cdots \partial \lambda_n} I(\lambda_1, \ldots, \lambda_n)|_{\lambda_1, \ldots, \lambda_n = 0},$$

we get

$$c_{\eta_1,\dots,\eta_n}(s)=\lambda^n\int\limits_R \hat a(s,u)\exp\Big\{\lambda\int\limits_0^T\Lambda(iu1_{[0,s]}(t))dt\Big\}\times$$
$$\prod_{j=1}^n\exp\Big\{\lambda\int\limits_0^T\Big(e^{iu1_{[0,s]}(t)}-1\Big)\eta_j(t)dt\Big\}du.$$

2. Now let $a(t,P_{(\cdot)})=\prod\limits_{m=1}^k\int\limits_0^t f_m(s)dP_s$, $f_m(s)\in L_2[0,T]\cap L_1[0,T]$, $m=1,\dots,k$. As in previous case we will use differentiation of generating functional:

$$E\Big[\prod_{m=1}^k\Big(\int\limits_0^t f_m(\tau)dP_\tau\Big)C_n(P;\eta_1,\dots,\eta_n)\Big]=$$
$$\frac{\partial^{k+n}}{\partial\mu_1\cdots\partial\mu_k\partial\lambda_1\cdots\partial\lambda_n}E\Big[\exp\Big\{\Big\langle\sum_{m=1}^k\mu_m1_{[0,t]}f_m,P\Big\rangle\Big\}\times$$
$$:\exp\Big\{\Big\langle\ln\Big(1+\sum_{j=1}^n\lambda_j\eta_j\Big)\Big\}:\Big]\Big|_{\mu_m=0,\lambda_j=0,\forall m,n}=$$
$$\frac{\partial^{k+n}}{\partial\mu_1\cdots\partial\mu_k\partial\lambda_1\cdots\partial\lambda_n}\exp\Big\{\lambda\int\limits_0^T\Lambda\Big(\sum_{m=1}^k\mu_m1_{[0,t]}(\tau)f_m(\tau)\Big)d\tau\Big\}\times$$
$$E\Big[:\exp\Big\{\Big\langle\sum_{m=1}^k\mu_m1_{[0,t]}f_m,P\Big\rangle\Big\}::\exp\Big\{\Big\langle\ln\Big(1+\sum_{j=1}^n\lambda_j\eta_j\Big)\Big\}:\Big]\Big|_{\mu_m=0,\lambda_j=0,\forall m,n}\equiv(A).$$

Next using the obvious equality

$$\Big\langle\sum_{m=1}^k\mu_m1_{[0,t]}f_m,P\Big\rangle=\Big\langle\ln\Big(\Big[\Big\{\sum_{m=1}^k\mu_m1_{[0,t]}f_m\Big\}-1\Big]+1\Big),P\Big\rangle,$$

we get

$$(A)=\frac{\partial^k}{\partial\mu_1\cdots\partial\mu_k}\exp\Big\{\lambda\int\limits_0^T\Lambda\Big(\sum_{m=1}^k\mu_m1_{[0,t]}(\tau)f_m(\tau)\Big)d\tau\Big\}\times$$
$$\prod_{j=1}^n\lambda\int\limits_0^T\eta_j(\tau)\Big(\exp\Big\{\sum_{m=1}^k\mu_m1_{[0,t]}(\tau)f_m(\tau)\Big\}-1\Big)d\tau\Big|_{\mu_m=0,m=\overline{1,k}}\equiv(B).$$

Note that in the expression $k \geq n$, because $(A) = 0$ in opposed case. Then,

$$(B) = \frac{\partial^k}{\partial\mu_1 \cdots \partial\mu_k}\Big(\chi_P\Big(-i\sum_{m=1}^{k}\mu_m 1_{[0,t]}(\cdot)f_m(\cdot)\Big)F\Big(\sum_{m=1}^{k}\mu_m 1_{[0,t]}(\cdot)f_m(\cdot)\Big)\Big)\Big|_{\mu_m=0,m=\overline{1,k}} =$$

$$\sum_{m=0}^{k}\sum_{(i_1,\ldots,i_m)}\frac{\partial^m}{\partial\mu_{i_1}\cdots\partial\mu_{i_m}}\chi_P\Big(-i\sum_{m=1}^{k}\mu_m 1_{[0,t]}(\cdot)f_m(\cdot)\Big)\times$$

$$\frac{\partial^{k-m}}{\partial\mu_{i_{m+1}}\cdots\partial\mu_{i_k}}F\Big(\sum_{m=1}^{k}\mu_m 1_{[0,t]}(\cdot)f_m(\cdot)\Big)\Big|_{\mu_m=0,m=\overline{1,k}},$$

where $F\Big(\sum\limits_{m=1}^{k}\mu_m 1_{[0,t]}(\cdot)f_m(\cdot)\Big) = \prod\limits_{j=1}^{n}\lambda\int\limits_0^T\eta_j(\tau)\Big(\exp\Big\{\sum\limits_{m=1}^{k}\mu_m 1_{[0,t]}(\tau)f_m(\tau)\Big\}-1\Big)d\tau$; the sum $\sum\limits_{(i_1,\ldots,i_m)}$ is over all possible samples of m numbers from the set $\{1,2,\ldots,k\}$, and where we have used the differentiation formulae with respect to parameter: $\frac{\partial}{\partial\mu_j}F(x_{\mu_j}(\cdot)) = \int\limits_0^T\frac{\delta F(x_{\mu_j}(\cdot))}{\delta x_{\mu_j}(\tau)}\frac{\partial x_{\mu_j}(\tau)}{\partial\mu_j}d\tau$. Note that the only terms containing products of derivatives from all n factors of F will be nonzero after we put $\mu_1 = \cdots = \mu_m = 0$. This implies

$$E[a(t,P_{(\cdot)})C_n(P;\eta_1,\ldots,\eta_n)] = (B) =$$

$$\sum_{m=0}^{k}\sum_{(i_1,\ldots,i_m)}E\Big[\prod_{l=1}^{m}\int_0^t f_{i_l}(\tau)dP_\tau\Big]\prod_{q=1}^{k-m}\Big(\lambda\int_0^t\eta_{i_{m+q}}(\tau)f_{i_{m+q}}(\tau)d\tau\Big).$$

The expectations in right part of this expression are the moments of linear functionals of Poisson process, so they can be evaluated explicitly.

References

[1] Egorov A.D. (2014). Evaluation of expectations of random functionals. *Matematicheskoe modelirovanie*. Vol. **26**, no. 11, pp. 29-32.

[2] Ito Y., Kubo I. (1988). Calculus on Gaussian and Poisson white noises. *Nagoya Math. J.* Vol. **111**, pp. 41-84.

[3] Privault N. (1994). Chaotic and variational calculus in discrete and continuous time for the Poisson process. *Stochasics and Stochastics Reports*. Vol. **51**, pp. 83-109.

[4] Surgailis D. (1984). On multiple Poisson stochastic integrals and associated Markov semigroups. *Probability and Mathematical Ststistics*. Vol. **3**, pp. 217-239.

[5] Egorov A.D., Sobolevsky P.I., Yanovich L.A. (1993). *Functional Integrals. Approximate evaluation and Applications*. Kluver Academic Publishers, Dordrecht.

SIGNIFICANCE LEVEL ANALYSIS FOR ADAPTIVE ALGORITHM OF STATIONARY POISSON STREAM PROCESSING

V. I. Nikitsionak, A. M. Bachar
Belarusian State University
Minsk, BELARUS

Depending on the analyzed sample values the stationary Poisson stream (SPS) of events has the Poisson or the exponential law of intervals between the adjacent events. Here we consider the exponential law under the fixed number m of incoming events. Decision-making time T is assumed to be random. Decision-making means testing simple hypothesis H_0: the distribution parameter (or SPS intensity) $\lambda = \lambda_0$, against alternative H_1: $\lambda = \lambda_1 > 0$. In adaptive algorithm for the aim of finding the optimum decision threshold the intensity λ_0 is estimated using the classified training SPS of events (or the training set, TS), corresponding to SPS processing with $\lambda = \lambda_0$.

The significance level of adaptive algorithm is obtained from the one for optimal algorithm by averaging over all values of unknown parameter λ_0. Using the known approximation of probability integral, we get:

$$F_a(m, F; m_0) \cong a\sqrt{J}(F/a)^J,\ J = \left(1 + 2bm_0^{-1}\left(\sqrt{1/b\ln(a/F)} - d - \sqrt{m}\right)^2\right)^{-1}, \quad (1)$$

where F is a significance level, m_0 is the volume of TS, $(a, b, d) = (0.65, 0.443, 0.75)$. Note that the values $\sqrt{m}$ to be used depend on the required quality parameters of adaptive algorithm. Namely, they depend on power and significance level of decision rule, as well as on the ratio $\Lambda = \lambda_1/\lambda_0$, characterizing the "distance" between hypotheses. From the calculations based on (1) it follows: 1) at the small "distance" $\Lambda = 1.1$ the significance level $F = 10^{-4}$ is reached at rather large TS of volume $m_0 = 21000$, about 40 times greater than the one $m_0 = 500$, providing power parity of adaptive and optimum algorithms. The significance levels $10^{-5} \ldots 10^{-6}$ are reached at even larger values of m_0; 2) at the larger "distance" $\Lambda = 2$ the significance levels $10^{-4} \ldots 10^{-6}$ are reached at $m_0 = 500$ with adaptive and optimum algorithms both of power equal 0.9.

Thus, the adaptive algorithm has satisfactory quality for $\Lambda = 2$, unlike the case $\Lambda = 1.1$, when the required volume of TS grows dramatically.

Further analysis shows the following.

At a close hypothesis and alternative, when $\Lambda = 1.1$ is small enough, the distributions of the observations under hypothesis and alternative are rather close. So as the significance level depends on the left quite a gentle "tail" of the hypothetical distribution, there is a need for highly accurate estimate of a decision threshold. This high accuracy in its turn is achievable at a very large volumes of TS.

On the other hand, for the well separated hypothesis and alternative the accuracy of a threshold estimate at $m_0 = 500$ appears suitable for good quality of decision rule.

STATISTICAL ANALYSIS OF MARKOV CHAINS WITH THE PERIODICALLY CHANGED TRANSITION PROBABILITY MATRICES

E. N. Orlova
Belarusian State University
Minsk, BELARUS
e-mail: orlovaen@bsu.by

Abstract

The paper deals with the problem of a statistical analysis of Markov chains with the periodically changed transition probability matrices. Statistical estimators for parameters of the model by observed time series are constructed.

1 Introduction

The knowledge of discrete time series is necessary for many applications. One usually needs to take into consideration dependence on the previous states of the process. Markov chain is a well known mathematical model adequate for these purposes. For instance, Markov chains are used in signal processing [1], genetics [2], economics [3], information security [4] and many other areas. The problem of development and analysis of Markov chains with a small is rather important [5, 6]. A special case of Markov chains – Markov chain with the periodically changed transition probability matrices – is considered.

2 Mathematical model

Consider Markov chain $\xi_t : \mathbb{N}_0 \to \mathbf{A}$ with a finite set $\mathbf{A}$ of $|\mathbf{A}| > 1$ states, initial distribution $\mathbf{P}\{\xi_0 = i \in \mathbf{A}\} = \pi_i$, $\sum_{i\in\mathbf{A}} \pi_i = 1$, and the matrices of transitions probabilities, T-periodically changed after every M observations:

$$\mathbf{P}\{\xi_{t+1} = j | \xi_t = i\} = P_{i,j}^{([t/M] \bmod T)}, \ i, j \in \mathbf{A}.$$

Remark. Further consider $N = LTM$, $L \in \mathbb{N}$.

Theorem 1. *The probability of realization* $\{d_0, d_1, \ldots, d_N\}$ *for the considered Markov chain* ξ *is:*

$$\mathbf{P}\{\xi_0 = d_0, \xi_1 = d_1, \ldots, \xi_N = d_N\} = \pi_{d_0} \prod_{r=0}^{T-1} \prod_{l=0}^{L-1} \prod_{m=0}^{M-1} P^{(r)}_{d_{lTM+(r-1)M+m}, d_{lTM+(r-1)M+m+1}}.$$

Theorem 2. *The state of the considered Markov chain ξ at the moment $t = lTM + (r-1)M + m$ has the following probability distribution:*

$$\pi^t = (\mathbf{P}\{\xi_t = i\})_{i\in\mathbf{A}} = \left(\left(P^{(r-1)}\right)'\right)^m \left(\left(P^{(r-2)}\right)'\right)^M \dots \left(\left(P^{(0)}\right)'\right)^M \times \left[\left(\left(P^{(T-1)}\right)'\right)^M \dots \left(\left(P^{(0)}\right)'\right)^M\right]^l \pi.$$

3 Statistical estimation of parameters

Construct now the maximum likelihood estimators of the matrices $P^{(0)}, \dots, P^{(T-1)}$ using an observed realization $X = \{x_0, x_1, \dots, x_N\}$ of length $N+1$ of Markov chain ξ with the periodically changed transition probability matrices. All other parameters are assumed to be known.

For $i, j \in \mathbf{A}$ and $r = 0, \dots, T-1$, introduce the notations:

$$n_{ij}^{(r)} = \sum_{l=0}^{L-1}\sum_{m=0}^{M-1} \delta_{x_{lTM+(r-1)M+m},i}\delta_{x_{lTM+(r-1)M+m+1},j},\ n_i^{(r)} = \sum_{j\in\mathbf{A}} n_{ij}^{(r)}.$$

Theorem 3. *If true values M, T and π are known, then the maximum likelihood estimators for the one-step transition probabilities $p_{ij}^{(r)}$, $i, j \in \mathbf{A}$, $r = 0, \dots, T-1$, are:*

$$\hat{p}_{ij}^{(r)} = \frac{n_{ij}^{(r)}}{n_i^{(r)}}.$$

References

[1] Li Y. et al. (2010). Spectrum Usage Prediction Based on High-order Markov Model for Cognitive Radio Networks. *10th IEEE Conf. Comp. Inf. Tech.* pp. 2784-2788.

[2] Waterman M.S. (1999). *Mathematical Methods for DNA Sequences*. Chapman and Hall/CRC: Boca Raton.

[3] Ching W.K. (2004). High-order Markov chain models for categorical data sequences. *Inc. Naval Research Logistics*. Vol. **51**(4), pp. 557-574.

[4] Kharin Yu.S., Bernik V.I., Matveev G.V., Agievich S.V. (2003). *Mathematical and computer foundation of cryptology*. Novoe znanie, Minsk (in Russian).

[5] Kharin Yu.S., Petlitskii A.I. (2007). A Markov chain of order s with r partial connections and statistical inference on its parameters. *Discrete Mathematics and Applications*. Vol. **17**(3), pp. 295-317.

[6] Raftery A.E. (1985). A Model for High-Order Markov Chains. *Royal Statistical Society*. Vol. **B-47**(3), pp. 528-539.

CONSISTENT ESTIMATORS OF DRIFT PARAMETER IN STOCHASTIC DIFFERENTIAL EQUATIONS DRIVEN BY FRACTIONAL BROWNIAN MOTION

K. V. Ralchenko
Taras Shevchenko National University of Kyiv
Kyiv, UKRAINE
e-mail: k.ralchenko@gmail.com

Abstract

We study a problem of drift parameter estimation in a stochastic differential equation driven by fractional Brownian motion. The form of the likelihood ratio in this model is in general rather complicated. However, in the simplest case it can be simplified and we can discretize it to establish the a. s. convergence of the discretized version of maximum likelihood estimator to the true value of parameter. We also investigate non-standard estimators of the drift parameter showing further its strong consistency.

1 Introduction and model description

Stochastic differential equations driven by a fractional Brownian motion have been a subject of an active research for the last two decades. Main reason is that such equations seem to be one of the most suitable tools to model long-range dependence in many applied areas, such as physics, finance, biology, network studies etc.

This paper deals with statistical estimation of drift parameter for a stochastic differential equation with fBm by discrete observation of its solution. We propose three new estimators and prove their strong consistency under the so-called "high-frequency data" assumption that the horizon of observations tends to infinity, while the interval between them goes to zero. Moreover, we obtain almost sure upper bounds for the rate of convergence of the estimators. The estimators proposed go far away from being maximum likelihood estimators, and this is their crucial advantage, because they keep strong consistency but are not complicated technically and are convenient for the simulations.

Fractional Brownian motion (fBm) with Hurst parameter $H \in (0,1)$ is a centered Gaussian process $\{B_t^H, t \geq 0\}$ on a complete probability space $(\Omega, \mathcal{F}, \mathsf{P})$ with the covariance

$$\mathsf{E}\left[B_t^H B_s^H \right] = \frac{1}{2}(s^{2H} + t^{2H} - |t-s|^{2H}).$$

It is well known that B^H has a modification with almost surely continuous paths (even Hölder continuous of any order up to H), and further we will assume that it is continuous itself.

In what follows we assume that the Hurst parameter $H \in (1/2, 1)$ is fixed. In this case, the integral with respect to the fBm B^H will be understood in the generalized

Lebesgue–Stieltjes sense. Its construction uses the fractional derivatives, defined for $a < b$ and $\alpha \in (0,1)$ as

$$\big(D^{\alpha}_{a+}f\big)(x) = \frac{1}{\Gamma(1-\alpha)}\bigg(\frac{f(x)}{(x-a)^{\alpha}} + \alpha\int_a^x \frac{f(x)-f(u)}{(x-u)^{1+\alpha}}du\bigg),$$
$$\big(D^{1-\alpha}_{b-}g\big)(x) = \frac{e^{-i\pi\alpha}}{\Gamma(\alpha)}\bigg(\frac{g(x)}{(b-x)^{1-\alpha}} + (1-\alpha)\int_x^b \frac{g(x)-g(u)}{(u-x)^{2-\alpha}}du\bigg).$$

It follows from Hölder continuity of B^H that for $\alpha \in (1-H,1)$ $D^{1-\alpha}_{b-}B^H_{b-} \in L_{\infty}[a,b]$ a.s. Then for a function f with $D^{\alpha}_{a+}f \in L_1[a,b]$ we can define integral with respect to B^H as the generalized Lebesgue-Stieltjes integral:

$$\int_a^b f(x)\,dB^H(x) := e^{i\pi\alpha}\int_a^b (D^{\alpha}_{a+}f)(x)(D^{1-\alpha}_{b-}B^H_{b-})(x)\,dx. \tag{1}$$

Consider a stochastic differential equation

$$X_t = X_0 + \theta\int_0^t a(X_s)ds + \int_0^t b(X_s)dB^H_s, \tag{2}$$

where X_0 is a non-random coefficient. In [2] it is shown that this equation has a unique solution under the following assumptions: there exist constants $\delta \in (1/H-1,1]$, $K>0$, $L>0$ and for every $N>0$ there exists $R_N>0$ such that

(A) $|a(x)| + |b(x)| \le K$ for all $x, y \in \mathbb{R}$,

(B) $|a(x)-a(y)| + |b(x)-b(y)| \le L\,|x-y|$ for all $x, y \in \mathbb{R}$,

(C) $|b'(x)-b'(y)| \le R_N\,|x-y|^{\delta}$ for all $x \in [-N,N], y \in [-N,N]$.

Our main problem is to construct an estimator for θ based on discrete observations of X. Specifically, we will assume that for some $n \ge 1$ we observe values $X_{t^k_n}$ at the following uniform partition of $[0,2^n]$: $t^n_k = k2^{-n}$, $k = 0,1,\dots,2^{2n}$.

In order to construct consistent estimators for θ, we need another technical assumption, in addition to conditions (A)–(C):

(D) $a(x)$ and $b(x)$ are separated from zero.

2 Maximum-likelihood estimator

In [3] the explicit form of the likelihood ratio was established. In the general case that formula is not suitable for applications because it involves a lot of weakly singular kernels and it is quite impossible to get its convergence to the true value of the parameter. But even if we get the convergence, the simulation error will be large enough to annihilate our efforts in discretization.

In order to avoid this technical difficulties, we consider the simplest case. Consider an equation

$$dX_t = \theta b(X_t)dt + b(X_t)dB^H_t.$$

In this case the maximum-likelihood estimator can be written as follows [3]:

$$\hat{\theta}_t^{(1)} = \frac{\int_0^t s^{-\alpha}(t-s)^{-\alpha} b^{-1}(X_s) dX_s}{\mathrm{B}(1-\alpha, 1-\alpha) t^{1-2\alpha}}, \tag{3}$$

where $\alpha = H - \frac{1}{2}$, $\mathrm{B}(x, y)$ is the beta function. Now we consider an estimator

$$\hat{\theta}_n^{(2)} = \frac{\sum_{k=1}^{2^{2n}-1} (t_k^n)^{-\alpha} (2^n - t_k^n)^{-\alpha} b^{-1}\left(X_{t_{k-1}^n}\right)\left(X_{t_k^n} - X_{t_{k-1}^n}\right)}{\mathrm{B}(1-\alpha, 1-\alpha) 2^{n(1-2\alpha)}}.$$

This estimator is a discretized version of the estimator (3).

Let us introduce the following conditions in addition to (A)–(D):

(a) $|b(x) - b(y)| \le C_1 |x - y|$, for all $x, y \in \mathbb{R}$,

(b) $C_4 \le |b(x)| \le C_2(1 + |x|)$, for all $x \in \mathbb{R}$,

(c) $|b'(x) - b'(y)| \le C_3 |x - y|^\rho$, for all $x, y \in \mathbb{R}$,

where $C_1, \ldots, C_4$ are positive constants and $\rho \in (1/H - 1, 1]$.

Theorem 1. *Under conditions* (a)–(c), $\hat{\theta}_n^{(2)}$ *is strongly consistent. Moreover, for any* $\beta \in (1/2, H)$ *and* $\gamma > 1/2$ *there exists a random variable* $\eta = \eta_{\beta,\gamma}$ *with all finite moments such that* $\left|\hat{\theta}_n^{(2)} - \theta\right| \le \eta n^{\kappa+\gamma} 2^{-\tau n}$, *where* $\kappa = \gamma/\beta$, $\tau = (1-H) \wedge (2\beta - 1)$.

3 Consistent estimators for drift parameter

We now define an estimator, which is a discretized version of a maximum likelihood estimator for $F(X)$, where $F(x) = \int_0^x b(y)^{-1} dy$:

$$\hat{\theta}_n^{(3)} = \frac{2^n \sum_{k=1}^{2^{2n}} (t_k^n)^{-\alpha} (2^n - t_k^n)^{-\alpha} b^{-1}\left(X_{t_{k-1}^n}\right)\left(X_{t_k^n} - X_{t_{k-1}^n}\right)}{\sum_{k=1}^{2^{2n}} (t_k^n)^{-\alpha} (2^n - t_k^n)^{-\alpha} b^{-1}\left(X_{t_{k-1}^n}\right) a\left(X_{t_{k-1}^n}\right)}.$$

Theorem 2. *Under conditions* (A)–(D), *Theorem 1 holds for* $\hat{\theta}_n^{(3)}$.

Consider a simpler estimator:

$$\hat{\theta}_n^{(4)} = \frac{2^n \sum_{k=1}^{2^{2n}} b^{-1}\left(X_{t_{k-1}^n}\right)\left(X_{t_k^n} - X_{t_{k-1}^n}\right)}{\sum_{k=1}^{2^{2n}} b^{-1}\left(X_{t_{k-1}^n}\right) a\left(X_{t_{k-1}^n}\right)}.$$

This is a discretized maximum likelihood estimator for θ in equation (2), where B^H is replaced by Wiener process. Nevertheless, this estimator is consistent as well. Namely, we have the following result.

Theorem 3. *Theorem 2 holds for* $\hat{\theta}_n^{(4)}$.

In the paper [1] the following non-standard estimator for θ was considered:

$$\hat{\theta}_t^{(5)} = \frac{\int_0^t a(X_s)b^{-2}(X_s)dX_s}{\int_0^t a^2(X_s)b^{-2}(X_s)ds}.$$

We define a discretized version of $\hat{\theta}_t^{(5)}$. Put

$$\hat{\theta}_n^{(6)} := \frac{2^n \sum_{k=1}^{2^{2n}} a\left(X_{t_{k-1}^n}\right) b^{-2}\left(X_{t_{k-1}^n}\right)\left(X_{t_k^n} - X_{t_{k-1}^n}\right)}{\sum_{k=1}^{2^{2n}} a^2\left(X_{t_{k-1}^n}\right) b^{-2}\left(X_{t_{k-1}^n}\right)}.$$

Let $\varphi(t) = \frac{a(X_t)}{b(X_t)}$,

$$\widehat{\varphi}_n(t) := \sum_{k=0}^{2^{2n}-1} \varphi(t_k^n) I_{[t_k^n, t_{k+1}^n)}(t).$$

Theorem 4. *Under conditions* (a)–(c)*, assume that there exist constants* $\beta > 1 - H$ *and* $p > 1$ *such that*

$$\frac{2^{n(H+\beta)} n^p \int_0^{2^n} \left|\left(D_{0+}^{\beta} \widehat{\varphi}_n\right)(s)\right| ds}{\sum_{k=1}^{2^{2n}} \varphi^2(t_{k-1}^n)} \to 0 \quad a.\,s.\ \text{at}\ n \to \infty.$$

Then $\hat{\theta}_n^{(6)}$ *is strongly consistent.*

References

[1] Kozachenko Y., Melnikov A., Mishura Y. (2015). On drift parameter estimation in models with fractional Brownian motion. *Statistics*. Vol. **49**(1), pp. 35–62.

[2] Lyons T. J. (1998). Differential equations driven by rough signals. *Rev. Mat. Iberoamericana*. Vol. **14**(2), pp. 215–310.

[3] Mishura Y., Ralchenko K. (2014). On drift parameter estimation in models with fractional Brownian motion by discrete observations. *Austrian J. Statist.* Vol. **43**, pp. 217–228.

STATISTICAL INFERENCE FOR RANDOM FIELDS IN THE SPECTRAL DOMAIN BASED ON TAPERED DATA

L. M. Sakhno
Taras Shevchenko National University of Kyiv
Kyiv, UKRAINE
e-mail: lms@univ.kiev.ua

Let $X(t)$, $t \in I$, be a real-valued measurable strictly stationary zero-mean random field, where I is R^d or Z^d endowed with the measure $\nu(\cdot)$ which is the Lebesgue measure or the counting measure ($\nu(\{t\}) = 1$) respectively. Suppose that all order moments exist and the field $X(t)$ has spectral densities of all orders $f_k(\lambda_1, ..., \lambda_{k-1}) \in L_1(S^{k-1})$, $k = 2, 3, ...$, where $S = R^d$ or $(-\pi, \pi]^d$ for the continuous-parameter or discrete-parameter cases respectively.

Let the field $X(t)$ be observed over the domain $D_T = [-T, T]^d \subset I$. Consider the problem of estimation of integrals of cumulant spectra of orders $k = 2, 3, ...$

$$J_k(\varphi_k) = \int_{S^{k-1}} \varphi_k(\lambda) f_k(\lambda) d\lambda \tag{1}$$

for appropriate functions $\varphi_k(\lambda)$ with $\varphi_k(\lambda) f_k(\lambda) \in L_1(S^{k-1})$. The functionals (1) can be used to represent some characteristics of stochastic processes and fields in nonparametric setting and also appear in the parametric estimation in the spectral domain, e.g., when the minimum contrast (or quasi-likelihood) estimators are studied.

We will base our analysis on tapered data $\{h_T(t) X(t), t \in D_T\}$, where $h_T(t) = h(t/T)$, $t \in R^d$, and the taper $h(t)$ satisfies some conditions. The use of tapers leads to the bias reduction of estimates, which is important when dealing with spatial data: tapers can help to fight the so-called "edge effects".

Denote $H_{k,T}(\lambda) = \int h_T(t)^k e^{-i(\lambda,t)} \nu(dt)$ and define the finite Fourier transform of tapered data: $d_T^h(\lambda) = \int h_T(t) X(t) e^{-i(\lambda,t)} \nu(dt)$, $\lambda \in S$, the tapered periodograms of the second and the third orders:

$$I_{2,T}^h(\lambda) = \frac{|d_T^h(\lambda)|^2}{(2\pi)^d H_{2,T}(0)}, \; I_{3,T}^h(\lambda_1, \lambda_2) = \frac{d_T^h(\lambda_1) d_T^h(\lambda_2) d_T^h(-\lambda_1 - \lambda_2)}{(2\pi)^{2d} H_{3,T}(0)}$$

(provided that $H_{2,T}(0) \neq 0$, $H_{3,T}(0) \neq 0$) and the tapered periodogram of k-th order:

$$I_{k,T}^h(\lambda_1, ..., \lambda_{k-1}) = \frac{1}{(2\pi)^{(k-1)d} H_{k,T}(0)} \prod_{i=1}^{k} d_T^h(\lambda_i), \lambda_i \in S,$$

(provided that $H_{k,T}(0) \neq 0$), where $\sum_{i=1}^k \lambda_i = 0$, but no proper subset of λ_i has sum 0.

We consider the empirical spectral functional of k-th order

$$J_{k,T}(\varphi_k) = \int_{S^{k-1}} \varphi_k(\lambda) I_{k,T}^h(\lambda) d\lambda. \tag{2}$$

as an estimate for the spectral functional (1). We discuss the questions: (i) evaluation of bias and (ii) conditions for asymptotic normality of (2). We pay special attention to the case $k = 2$ and present applications for parameter estimation of particular models of random fields.

AN IMPROVED K-NEAREST NEIGHBORS ALGORITHM FOR THE ANALYSIS OF TWO-COLOR DNA MICROARRAY DATA WITH SPOT QUALITY FACTORS

A. Svidrytski[1], M. Yatskou, V. Apanasovich
Belarusian State University
Minsk, BELARUS
e-mail: [1]svidrytski@gmail.com

Abstract

Algorithms for the classification of gene expression data require the continuous improvement of their efficiency. This paper presents a modification of a k-nearest neighbors algorithm which increases an efficiency of classification including quality parameter of each microarray spot. The efficiency of classification is achieved by recalculation of distances between classified and classifying objects. We also introduce an enhanced microarray data simulation model that includes spot quality parameters.

1 Introduction

Microarrays are one of the newest instruments of biology and medicine [1]. Their advantage caused by possibility to conduct an enormous number of specific reactions and interactions of biopolymer molecules simultaneously.

The first step in the entire microarray analysis is DNA microarray image processing, using such tools as GenePix or MAIA (MicroArray Image Analysis) [3]. Every mistake made on this step can further influence the final results significantly.

As a rule, after retrieving data, analysis of microarray gene expression includes step that removes objects with low quality [3]. In this paper there is presented an improved kNN algorithm, using quality parameter as a weight factor, which makes it possible to increase efficiency of a microarray analysis taking into account spots with lower quality. There is also presented algorithm for microarray simulation adapted for inclusion of quality parameter. This microarray model was applied to evaluate efficiency of the modified algorithm in comparison with original one.

2 DNA microarray model with a quality parameter

2.1 MAIA and a quality parameter

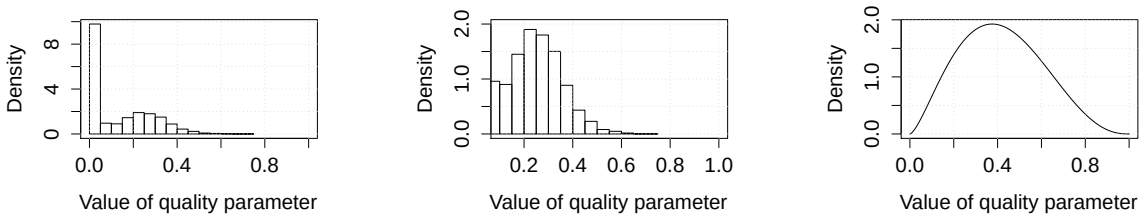
Microarray image analysis software like GenePix or MAIA provides an integrative estimate of spot quality in range from 0 to 1 after image processing [3]. The quality value can be used as a weight factor for classification analysis.

2.2 DNA microarray simulation model with a quality parameter

Simulation model of microarray gene expression values must be as similar to real data as possible. This is achieved by simulation of physical phenomena in the model. Dembele [2] proposed the universal microarray simulation model that is most advanced up to date. Thus we took this model for adding quality factor.

A distribution of the quality parameter was obtained resting on the microarray data in a whole-genome microarray experiment assessing well-characterized transcriptional modifications induced by the transcription regulator SNAI1 [3].

Histogram of the total spot quality parameters is shown in figure 1*a*. Such low-quality spots appear mainly due to poor microarray experiment conduction. Filtering objects with very low quality (from 0 to 0.1) resembles shape of beta-distribution with mode about 0.25 (figure 1*b*). Data of better quality would have mode higher than 0.25. In this work we took mode for the distribution equal 0.375. Empirical choice of the parameters for beta distribution equals to $\alpha_1 = 2.5$ and $\alpha_2 = 3.5$. Plot of density function with these parameters is presented in the figure 1*c*.



a *b* *c*

Figure 1: Histograms and Probability Density Function of the spot quality parameter: *a)* histogram of microarray data taken from [3] *b)* histogram of microarray data after filtering objects with qualities between 0 and 0.1 *c)* curve of beta-distribution with fitted parameters; it approximately resembles the shape of the histogram with shifted mode.

The next step is to bind the value of the quality parameter to value of gene expression that was implemented adding a Gaussian noise. Variance of the distribution is the greater the worse object quality is. The variance for i-th spot was calculated according to expression $\sigma_i = 1 - q_i$.

3 kNN modification using a quality parameter

The idea for the modification of the original k-nearest neighbors method involves adjusting distances between objects of training and test samples. Adjustments weight the between-object distances so that if the lower quality of training object then it contributes less to classifying test objects. The illustration of this approach is shown in

figure 2.

Algorithm:

1. Initialize the number of nearest neighbors k, the volumes of training and test samples.
2. Find distances d_{ij} between all objects of the training and test samples.
3. Modify distances with some function f which depends on quality parameter q_i: $d_{ij}^w = f(d_{ij})$.
4. Identify k nearest neighbors for each test object.
5. Determine the class label for each test object.

A function f was chosen hyperbolic: $d_{ij}^w = f(d_{ij}) = d_{ij}/qi$. This function is chosen because when a quality of spot goes to 0 then d_{ij}^w diverges to infinity. d_{ij} can be also logarithmically transformed: $d_{ij}^w = f(d_{ij}) = d_{ij} \cdot (-\log_a q_i + 1)$.

4 Results and discussion

4.1 Description of numerical experiments

Numerical experiment included:

1. Generation of test sample consisted of 10000 spots, where 100 spots are up regulated and 100 spots are down regulated.
2. Generation of training sample consisted of 300 spots where 100 spots are up regulated and 100 spots are down regulated. Expression values of neutral spots were distributed between -.1 and .1.
3. Receiving the number of correctly classified spots with classic and modified kNN methods considering the same test and training samples.

The described experiment was repeated 150 times under certain conditions. The changing conditions involved variation of the lowest value of spot quality for both training and test samples what reflected general microarray quality.

4.2 Results

The following formula was used to compare weighted and basic algorithms:

$$efficiency = 100\% \cdot \frac{T^{weighted}}{T^{basic}},$$

where $T^{weighted}$ and T^{basic} are the numbers of correctly classified objects by improved and basic algorithms respectively. Figure 3 represents heatmap, where the more green color reflects the more efficient improved algorithm is. Each cell contains average value of 150 efficiency values obtained for the certain lowest value of quality parameter.

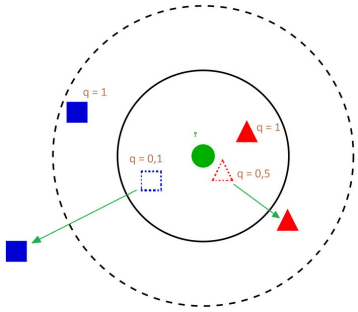


Figure 2: This picture illustrates how the method of distances modification works. A triangle with quality 0.5 is moved to a certain distance from the object for classification. A square which has lower quality 0.1 is moved even further. Squares and triangles that have quality 1 stay at the same place.

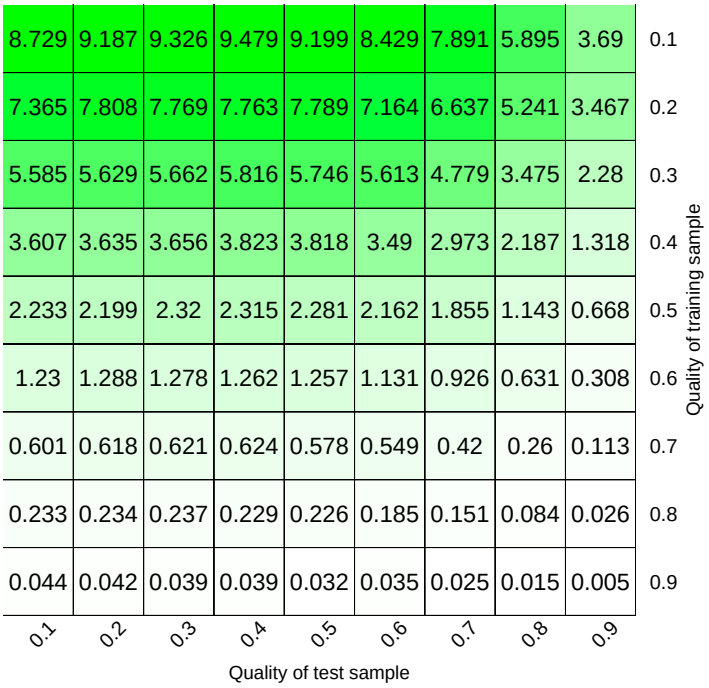


Figure 3: Heatmap of percent of efficiency of improved method in comparison with basic algorithm of kNN

5 Conclusion

This work presents a modified kNN algorithm that takes into account a spot quality parameter of a microarray. The algorithm works more efficiently than the classical one, when quality of training and test samples is worse. The idea of using the spot quality parameter can be embedded into more advanced algorithms of classification and cluster analysis.

References

[1] Zhang L. et al. (2015). Whole transcriptome microarrays identify long non-coding RNAs associated with cardiac hypertrophy. *Genom Data*. Vol. **5**, pp. 68–71.

[2] Dembele D.A. (2013). A Flexible Microarray Data Simulation Model. *Microarrays*. Vol. **2**, pp. 115–130.

[3] Yatskou M. et al. (2008). Advanced spot quality analysis in two-colour microarray experiments. *BMC Research Notes*. Vol. **1**(80).

GUARANTEED CHANGE POINT DETECTION OF LINEAR AUTOREGRESSIVE PROCESSES WITH UNKNOWN NOISE VARIANCE[1]

S. E. Vorobeychikov[2], Y. B. Burkatovskaya[3]
[2] *Tomsk State University*
[3] *Tomsk Polytechnic University*
Tomsk, RUSSIA
e-mail: [2]sev@mail.tsu.ru, [3]tracey@tpu.ru

Abstract

The problem of change point detection of autoregressive processes with unknown parameters is considered. A sequential procedure with guaranteed quality is proposed and both asymptotic and non-asymptotic properties of the algorithm are studied.

1 Introduction and problem statement

The problem of sequential change point detection for autoregressive processes often arises in different applications connected with time series analysis. The most difficult case is the case when all the process parameters are unknown. Theoretical properties of the procedures are commonly studied asymptotically when the number of observations before a change point tends to infinity. For small samples as a rule simulation study is conducted.

This paper develops an alternative approach in the frame of guaranteed sequential methods. It is based on the method of change point detection for AR(p) process proposed in [1]. In this study such an approach is applied to a general autoregressive model with unknown parameters. Using a special stopping rule we construct statistics which variances are bounded from above by a known constant. Hence, we can estimate the probabilities of false alarm and delay non-asymptotically, but asymptotic properties of the statistics are also investigated and more precise results are obtained.

We consider the scalar autoregressive process to be specified by the equation

$$x_{k+1} = A_k\lambda + B_k\xi_{k+1}, \tag{1}$$

where $\{\xi_k\}_{k\geq 0}$ is a sequence of independent identically distributed random variables with zero mean and unit variance. The density distribution function $f_\xi(x)$ of $\{\xi_k\}_{k\geq 0}$ is strictly positive for any x. The value $m > 1$ defines the order of the process; $\lambda = [\lambda_1, \ldots, \lambda_m]$ is the parameter vector of dimension $m \times 1$; A_k is the known $1 \times m$ matrix, the unknown noise variance B_k is bounded from above, i.e., $B_k^2 \leq D^2 < \infty$, $\mathcal{F}_k = \sigma\{\xi_1, ..., \xi_k\}$ is the σ-algebra generated by variables $\{\xi_1, ..., \xi_k\}$, A_k and B_k are $\mathcal{F}_k$-measurable. The value of the parameter vector λ changes at the change point θ:

$$\lambda = \lambda(k) = \begin{cases} \mu_0, & \text{if } k < \theta; \\ \mu_1, & \text{if } k \geq \theta. \end{cases}$$

[1]This paper is supported by The National Research Tomsk State University Academic D.I. Mendeleev Fund Program (NU 8.1.55.2015 L) in 2014–2015 and by Russian Foundation for Basic Research Grant 16-01-00121 A.

Values of the parameters before and after θ are supposed to be unknown. The difference between μ_0 and μ_1, for some known Δ, satisfies the condition

$$(\mu_0 - \mu_1)'(\mu_0 - \mu_1) \geq \Delta. \tag{2}$$

The problem is to detect the change point θ from observations x_k.

2 Guaranteed parameter estimator

Let $N_1 \geq m$ be the instant of the estimating procedure start, $n > 0$ is the volume of the initial sample used to estimate the noise variance. The estimator is constructed in the form

$$\begin{gathered} \tilde{\lambda}^*(H) = C^{-1}(N_1+n,\tau) \sum_{k=N_1+n}^{\tau} v_k A_k' x_{k+1}; \\ C(N_1+n,\tau) = \sum_{k=N_1+n}^{\tau} v_k A_k' A_k. \end{gathered} \tag{3}$$

We choose the value $\Gamma(N_1, n)$ from the following condition:

$$E\left(D^2/\Gamma(N_1,n)\right) \leq 1. \tag{4}$$

The weights on the interval $[N_1+n, N_1+n+\sigma]$ are taken in the form

$$v_k = \begin{cases} (\Gamma(N_1,n) A_k A_k')^{-1/2}, & \text{if } A_{N_1}, \ldots, A_k \text{ are linearly independent;} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The weights v_k on the interval $[N_1+n+\sigma+1, \tau-1]$ are found from the following condition:

$$\nu_{\min}(N_1+n,k)/\Gamma(N_1,n) = \sum_{l=N_1+n+\sigma}^{k} v_l^2 A_l A_l', \tag{6}$$

where $\nu_{\min}(N_1+n,k)$ is the minimal eigenvalue of the matrix $C(N_1+n,k)$.

Choosing a positive parameter H, we define the stopping time $\tau = \tau(H)$ as

$$\tau = \inf\left(N > N_1+n : \ \nu_{\min}(N_1+n,N) \geq H\right). \tag{7}$$

At the instant τ, the weight is found from the condition:

$$\nu_{\min}(N_1,\tau)/\Gamma(N_1,n) \geq \sum_{l=N_1+n+\sigma}^{\tau} v_l^2 A_l A_l', \quad \nu_{\min}(N_1,\tau) = H. \tag{8}$$

The parameter H defines the accuracy of the estimator. The choice of the weights v_k allows us to establish a non-asymptotic upper bound for the accuracy.

Theorem 1. *Let the parameter λ in (1) be constant, the compensating factor $\Gamma(N_1,n)$ satisfy condition (4) and the weights v_k determined in (5–6) be such that*

$$\sum_{k=0}^{\infty} v_k^2 A_k A_k' = \infty \ \ a.s. \tag{9}$$

Then the stopping time τ (7) is finite with probability one and the mean square accuracy of estimator (3) is bounded from above

$$E||\lambda^*(H) - \lambda||^2 \leq P(H)/H^2, \quad P(H) = H + m - 1. \tag{10}$$

Condition (9) hold true for the process AR(p).

Example. Let the observed process AR(p) be described by equation

$$x_{k+1} = \lambda_1 x_k + \ldots + \lambda_m x_{k-m+1} + B\xi_{k+1}. \tag{11}$$

Then the compensating factor can be chosen in the form proposed in [2]

$$\Gamma(N_1, n) = D(N_1, n) \sum_{l=N_1}^{N_1+n-1} x_l^2. \tag{12}$$

If the noises $\{\varepsilon_k\}_{k\geq 1}$ in (11) are normally distributed with zero mean and unit variance then the multiplier $D(N_1, n) = (n-2)^{-1}$.

3 Change point detection

Consider now the change point detection problem for process (1). We construct a set of sequential estimation plans

$$(\tau_i, \lambda_i^*) = (\tau_i(H), \lambda_i^*(H)),\ i \geq 1,$$

where $\{\tau_i\},\ i \geq 0$ is the increasing sequence of the stopping instances ($\tau_0 = -1$), and λ_i^* is the guaranteed parameter estimator on the interval $[\tau_{i-1}+1, \tau_i]$. The following condition holds true for the estimator

$$E\,||\lambda_i^*(H) - \lambda||^2 \leq P(H)/H^2. \tag{13}$$

Then we choose an integer $l > 1$. We associate the statistic J_i with the $i-$th interval $[\tau_{i-1}+1, \tau_i]$ for all $i > l$

$$J_i = \left(\lambda_i^* - \lambda_{i-l}^*\right)' \left(\lambda_i^* - \lambda_{i-l}^*\right). \tag{14}$$

This statistic is the squared deviation of the estimators with numbers i and $i-l$.

Theorem 2. *The expectation of the statistics J_i (14) satisfies the following inequalities:*

$$E\left[J_i |\, \tau_i < \theta\right] \leq 4P(H)/H^2, \quad E\left[J_i |\, \tau_{i-l} < \theta \leq \tau_{i-1}\right] \geq \Delta - 4\sqrt{\Delta P(H)/H^2}. \tag{15}$$

Hence, the change of the expectation of the statistic J_i allows us to construct the following change point detection algorithm. We choose the values of the parameter H and of the parameter $\delta > 0$ satisfying the following condition

$$4P(H)/H^2 < \delta < \Delta - 4\sqrt{\Delta P(H)/H^2}. \tag{16}$$

The J_i values are compared with the threshold δ. The change point is considered to be detected when the statistic exceeds δ. Due to the application of the guaranteed parameter estimators in the statistics, we can bound the probabilities of false alarm and delay from above.

Theorem 3. *The probability of false alarm P_0^i and the probability of delay P_1^i in any observation cycle $[\tau_{i-1}+1, \tau_i]$ are bounded from above*

$$P_0^i \leq 4P(H)/\delta H^2, \quad P_1^i \leq 4P(H)/\left((\sqrt{\Delta} - \sqrt{\delta})^2 H^2\right). \tag{17}$$

4 Asymptotic properties of the statistics

In the following theorem an asymptotic upper bound for the probability of large values of the standard deviation for the estimator (3) is obtained.

Theorem 4. *If for process (1)* $B_k^2 \le D^2 < \infty$, *and*

$$\max_{1\le k\le \tau(H)} \frac{v_k^2 D^2 ||A_k||^2}{\Gamma(N_1,N)H} \to^{\mathcal{P}} 0, \ as\ H\to\infty;$$

and the compensating factor $\Gamma(N_1,N)$ *satisfies the following conditions*

$$N\to\infty, \quad N/H\to 0 \quad as\ H\to\infty, \quad \Gamma(N_1,N)\to^{\mathcal{P}} const \quad as\ N\to\infty$$

then for sufficiently large H *in the conditions of Theorem 1*

$$\mathcal{P}\left\{||\lambda^*-\lambda||^2 > x\right\} \le 2\left(1-\Phi\left(\sqrt{\tfrac{xH^2}{H+m-1}}\right)\right), \tag{18}$$

where $\Phi(\cdot)$ *is the standard normal distribution function.*

The following theorem provides the asymptotic inequalities for the probabilities of false alarm and delay for the change point detection procedure.

Theorem 5. *For process (1) in the conditions of Theorem 4 for sufficiently large* H *the probabilities of false alarm* P_0^i *and delay* P_1^i *in any observation cycle* $[\tau_{i-1}+1,\tau_i]$ *are bounded from above*

$$\begin{gathered} P_0^i = \mathcal{P}\left\{||\zeta_i-\zeta_{i-l}||^2>\delta\right\} \le 4\left(1-\Phi\left(\sqrt{\tfrac{\delta H^2}{4(H+m-1)}}\right)\right); \\ P_1^i \le \mathcal{P}\left\{||\zeta_i-\zeta_{i-l}||^2 > \left(\sqrt{\Delta}-\sqrt{\delta}\right)^2\right\} \le 4\left(1-\Phi\left(\sqrt{\tfrac{\left(\sqrt{\Delta}-\sqrt{\delta}\right)H^2}{4(H+m-1)}}\right)\right), \end{gathered} \tag{19}$$

where $\Phi(\cdot)$ *is the standard normal distribution function.*

The conditions of the theorems hold true for the stable AR(p) process.

References

[1] Burkatovskaya Y.B., Vorobeychikov S.E. (2011). Change point detection of autoregressive process with unknown parameters. *Preprints of the 18th IFAC World Congress.* pp. 13215–13220.

[2] Konev V.V., Dmitrienko A.A. (1994). On guaranteed estimation of autoregression parameters when the noise variance is unknown. *Automatics and Remote Control.* Vol. **2**, pp. 87–99.

Section 3

PROBABILISTIC AND STATISTICAL ANALYSIS OF DISCRETE DATA

ON THE LIMIT DISTRIBUTION OF THE MAXIMUM VERTEX DEGREE IN A CONDITIONAL CONFIGURATION GRAPH

I. A. Cheplyukova
Institute for Applied Mathematical Research, Karelian Research Centre of RAS
Petrozavodsk, RUSSIA
e-mail: chia@krc.karelia.ru

Abstract

Configuration graphs where vertex degrees are independent identically distributed random variables are often used for modeling complex networks, such as the Internet, social media and others. We consider a random graph consisting of N vertices. The random variables $\eta_1, \dots, \eta_N$ are equal to the degrees of vertices with the numbers $1, \dots, N$. The probability $\mathbf{P}\{\eta_i = k\}$, $i = 1, \dots, N$, is equivalent to $h(k)/k^\tau$ as $k \to \infty$, where $h(x)$ is a slowly varying function integrable in any finite interval, $\tau > 1$. We obtain the limit distribution of the maximum vertex degree under the condition that the sum of degrees is equal to n and $N, n \to \infty$.

1 Introduction

Much attention has been paid to studying the asymptotic behaviour and the structure of random graphs which simulate various complex networks, such as the Internet or telecommunication networks (see e.g. [4], [5]). One of the most commonly used random graphs is the configuration model with the degree of vertices distributed identically and independently. The notion of the configuration graph was introduced in [1] for the first time. The process of graph construction consists of two stages. First, each numbered vertex of such a graph is assigned a certain degree in accordance with a given distribution. The vertex degree is the number of stubs that are numbered in an arbitrary order. Stubs are vertex edges for which adjacent nodes are not yet determined (semiedges). The graph is constructed at the second stage by joining each stub to another equiprobably to form edges. It is clear that we need to use the auxiliary vertex for the sum of degrees to be even. This vertex has the degree 0 if the sum of all other vertices is even, else the degree is 1.

A fundamental trait of many real networks is that the number of nodes with the degree k is near proportional to $k^{-\tau}$, $k \to \infty$, $\tau > 1$. There are many papers where the results describing the limit behaviour of different random graph characteristics were obtained. In [10] the configuration graph was considered where vertex degrees η have the distribution

$$\mathbf{P}\{\eta \geq k\} = h(k)k^{-\tau+1}, \quad k = 1, 2, \dots, \tag{1}$$

where $h(k)$ is a slowly varying function. The authors of this paper are convinced (without proof) that the function $h(k)$ does not influence limit results and that to study the configuration graph one can replace $h(k)$ with the constant 1. Various characteristics

of such graphs were studied, for example in [7] the limit theorem for the sum of vertex degrees was obtained. In our work we will show that the role of the slowly varying function $h(k)$ is more complicated.

We consider a random graph where random variables $\eta_1, \dots, \eta_N$ equal to the degrees of vertices with the numbers $1, \dots, N$ have the distributions

$$p_k = \mathbf{P}\{\eta_1 = k\} = \frac{h(k)}{k^\tau \Sigma(1,\tau)}, \tag{2}$$

where $k = 1, 2, \dots$, $\tau > 1$, $h(x)$ is a slowly varying function integrable in any finite interval and

$$\Sigma(x,y) = \sum_{k=1}^{\infty} x^k \frac{h(k)}{k^y}. \tag{3}$$

Further we consider the subset of random graphs under the condition that $\eta_1 + \dots + \eta_N = n$. Analysis of conditional random graphs was first carried out in [9]. It is not difficult to see that the addition of function $h(x)$ to the distribution (2) allows to consider this model as a generalization of the random graphs considered in [7]- [9]. For such random graphs, in [2], [3] the limit distributions of the maximum vertex degree were obtained as $n, N \to \infty$ and $1 < n/N \le C < \Sigma(1,\tau-1)/\Sigma(1,\tau)$, where C is a positive constant and $\Sigma(x,y)$ is determined by the relation (3). Now we obtain the limit distributions of the maximum vertex degree as $n, N \to \infty$ and $n/N \nearrow \Sigma(1,\tau-1)/\Sigma(1,\tau)$. Note that if $\tau < 2$, then $n/N \to \infty$.

2 The main result

We denote by $\xi_1, \dots, \xi_N$ auxiliary independent identically distributed random variables such that

$$p_k(\lambda) = \mathbf{P}\{\xi_i = k\} = \frac{\lambda^k p_k \Sigma(1,\tau))}{\Sigma(\lambda,\tau)}, \quad i = 1, 2, \dots, N, \quad k = 1, 2, \dots, \quad 0 < \lambda < 1.$$

From this we obtain

$$m = \mathbf{E}\xi_1 = \frac{\Sigma(\lambda,\tau-1)}{\Sigma(\lambda,\tau)}.$$

Let $\lambda = \lambda(N,n)$ be determined by the relation

$$\frac{\Sigma(\lambda,\tau-1)}{\Sigma(\lambda,\tau)} = \frac{n}{N}.$$

We introduce the conditions:

(A1) $\tau > 4$;
(A2) $3 < \tau \le 4$, $(1-\lambda)^{\tau-4-\epsilon}/\sqrt{\mathrm{N}} \to 0$;
(A3) $5/2 < \tau \le 3$, $\mathrm{N}(1-\lambda)^{11-3\tau+\epsilon} \ge \mathrm{C}_3 > 0$;
(A4) $\tau = 5/2$, $\mathrm{N}(-\ln(1-\lambda))^2(1-\lambda)^{7/2+\epsilon} \ge \mathrm{C}_4 > 0$;
(A5) $1 < \tau < 5/2$, $\mathrm{N}(1-\lambda)^{6-\tau+\epsilon} \ge \mathrm{C}_5 > 0$,

where ϵ is some sufficiently small positive constant.

We denote by $\eta_{(N)}$ the maximum vertex degree.

Theorem. *Let $N, n \to \infty$, $n/N \nearrow \Sigma(1, \tau - 1)/\Sigma(1, \tau)$, parameters τ, N, n are determined by one of the conditions (A1)–(A5), and $r = r(N, n)$ satisfies*

$$\frac{N\lambda^{r+1}h(r+1)}{(r+1)^{\tau}\Sigma(\lambda,\tau)(1-\lambda)} \to \gamma,$$

where γ is a positive constant. Then for any fixed $k = 0, \pm 1, \ldots$

$$\mathbf{P}\{\eta_{(N)} \le r\} = e^{-\gamma}(1 + o(1)).$$

3 Proof of the theorem

The technique for obtaining these theorems is based on the generalized allocation scheme suggested by V.F.Kolchin [6]. It is readily seen that for our subset of graphs

$$\mathbf{P}\{\eta_1 = k_1, \ldots, \eta_N = k_N\} = \mathbf{P}\{\xi_1 = k_1, \ldots, \xi_N = k_N | \xi_1 + \ldots + \xi_N = n\}.$$

Therefore the conditions of the generalized allocation scheme are valid.

Let $\xi_1^{(r)}, \ldots, \xi_N^{(r)}$ and $\tilde{\xi}_1^{(r)}, \ldots, \tilde{\xi}_N^{(r)}$ be two sets of independent identically distributed random variables such that

$$\mathbf{P}\{\xi_1^{(r)} = k\} = \mathbf{P}\{\xi_1 = k | \xi_1 \le r\}.$$

We also put $\zeta_N = \xi_1 + \ldots + \xi_N$, $\zeta_N^{(r)} = \xi_1^{(r)} + \ldots + \xi_N^{(r)}$, $P_r = \mathbf{P}\{\xi_1 > r\}$. It is shown in [6] that

$$\mathbf{P}\{\eta_{(N)} \le r\} = (1 - P_r)^N \frac{\mathbf{P}\{\zeta_N^{(r)} = n\}}{\mathbf{P}\{\zeta_N = n\}}. \tag{4}$$

From (4) we see that to obtain the limit distributions of $\eta_{(N)}$ it suffies to consider the asymptotic behaviour of the sums of auxiliary independent identically distributed random variables ζ_N, $\zeta_N^{(r)}$. To solve these problems one has to find both integral and local convergence of the distributions of these sums to limit laws under the conditions of array schemes, which is the main difficulty.

The study was supported by the Russian Foundation for Basic Research, grant 16-01-00005.

References

[1] Bollobas B. (1980). A probabilistic proof of an asymptotic formulae for the number of labelled regular graphs. *Eur.J.Comb.*. Vol. **1**, pp. 311-316.

[2] Cheplyukova I.A., Pavlov Yu.L. (2015). Limit distributions of vertex degrees in a conditional configuration graph. Book of abstracts Eighth international workshop on simulation, pp. 49-50.

[3] Cheplyukova I.A., Pavlov Yu.L. (2015). On vertex degrees in a conditional configuration graph. Proceedings of the IX International Workshop Applied Problems in Theory of Probabilities and Mathematical Statistics related to modeling of information systems, pp. 27-30.

[4] Faloutsos C., Faloutsos P., Faloutsos M. (1999). On power–law relationships of the Internet topology. *Computer Communications Rev.*. Vol. **29**, pp. 251-262.

[5] Hofstad R. Random graphs and complex networks. (2011). *Eindhoven university of technology*.

[6] Kolchin V.F. (2010). *Random Graphs*. Cambridge Univ.Press, Cambridge.

[7] Pavlov Yu.L. (2007). The limit distribution of the size of a giant component in an Internet-type random graph. *Discrete mathematics and applications*. Vol. **19**, issue 3, pp. 22-34.

[8] Pavlov Yu.L. (2010). On conditional Internet graphs whose vertex degrees have no mathematical expectation. *Discrete mathematics and applications*. Vol. **33**, issue 3, pp. 20-33.

[9] Pavlov Yu.L., Cheplyukova I.A. (2008). Random graps of Internet type and the generalized allocation scheme. *Discrete mathematics and applications*. Vol. **18**, issue 5, pp. 447-464.

[10] Reittu H., Norros I. (2004). On the power-law random graph model of massive data networks. *Performance Evaluation*. Vol. **55**, pp. 3-23.

SOME REMARKS ON THE NONCENTRAL PEARSON STATISTICS DISTRIBUTIONS

M. V. Filina[1], A. M. Zubkov[2]
Steklov Mathematical Institute, Russian Academy of Sciences
Moscow, RUSSIA
e-mail: [1]mfilina@mi.ras.ru, [2]zubkov@mi.ras.ru

Abstract

By means of numerical algorithms we investigate the exact distributions of the Pearson statistics under alternatives and possibility to use the noncentral chi-square or normal distributions as approximations.

1 Introduction

Let $\nu_1, \ldots, \nu_N$ be frequencies of N outcomes of a multinomial scheme in a sample of size T. A most popular goodness-of-fit test for the hypothesis H_p: "probabilities of outcomes are positive and equal to $p_1, \ldots, p_N$" is based on the Pearson statistics

$$X^2_{N,T} = \sum_{i=1}^{N} \frac{(\nu_i - Tp_i)^2}{Tp_i}\,. \tag{1}$$

If the hypothesis H_p is valid, then the distribution of $X^2_{N,T}$ converges (as $T \to \infty$) to the chi-square distribution with $N-1$ degrees of freedom having mean $N-1$ and variance $2(N-1)$. It is well-known that if the hypothesis is not valid, then in the triangular scheme with $T \to \infty$ and true probabilities of outcomes having the form $\pi_1 = p_1 + \frac{a_1}{\sqrt{T}}, \ldots, \pi_N = p_n + \frac{a_N}{\sqrt{T}}$ ($a_1, \ldots, a_N$ are fixed and $a_1 + \ldots + a_N = 0$) the distribution of the Pearson statistics $X^2_{N,T}$ converges to the noncentral chi-square distribution with noncentrality parameter $\lambda = \sum_{i=1}^{N} \frac{a_k^2}{p_k} = \mathbf{E}X^2_{N,T} - (N-1)$. If $T \to \infty$ and true probabilities of outcomes $\pi_1, \ldots, \pi_N$ are fixed, $\sum_{i=1}^{N}(\pi_i - p_i)^2 > 0$, then the Pearson statis1tics $X^2_{N,T}$ is asymptotically normal (see [1]) with mean

$$\mathbf{E}X^2_{N,T} = N - 1 + (T-1)\sum_{i=1}^{N} \frac{(\pi_i - p_i)^2}{p_i} + \sum_{i=1}^{N} \frac{\pi_i - p_i}{p_i}$$

and variance ([2, 3])

$$\mathbf{D}X^2_{N,T} = \frac{1}{T}\left((T-1)(6-4T)\left[\sum\nolimits_{i=1}^{N} \frac{\pi_i^2}{p_i}\right]^2 + 4(T-1)(T-2)\sum\nolimits_{i=1}^{N} \frac{\pi_i^3}{p_i^2} - \right. \tag{2}$$

$$\left. -4(T-1)\sum\nolimits_{i=1}^{N} \frac{\pi_i^2}{p_i}\sum\nolimits_{i=1}^{N} \frac{\pi_i}{p_i} + 6(T-1)\sum\nolimits_{i=1}^{N} \frac{\pi_i^2}{p_i^2} - \left[\sum\nolimits_{i=1}^{N} \frac{\pi_i}{p_i}\right]^2 + \sum\nolimits_{i=1}^{N} \frac{\pi_i}{p_i^2}\right);$$

in this case the "noncentrality parameter" $\mathbf{E}X^2_{N,T} - (N-1)$ tends to infinity as a linear function of T. So, there are a vast space between the conditions of these two theorems. Moreover, in the case of convergence to the non-central chi-square distribution the latter depends on the noncentrality parameter and on N only, whereas in the case of asymptotic normality the asymptotic variance of $X^2_{N,T}$ depends essentially on all probabilities $\pi_1, \ldots, \pi_N$ (and usually in practice these probabilities are unknown).

2 Results

Using the algorithms of exact computation of Pearson statistics distributions (see [4, 5]) we investigate the accuracy of approximations of these distributions by noncentral chi-square and normal distributions.

The character of the dependence of the variance on $\pi_1, \ldots, \pi_N$ may be illustrated by the case $p_1 = \ldots = p_N = \frac{1}{N}$: here

$$\mathbf{E}X^2_{N,T} = N - 1 + (T-1)N \sum\nolimits_{i=1}^{N} \left(\pi_i - \tfrac{1}{N}\right)^2,$$

$$\mathbf{D}X^2_{N,T} = \frac{N^2}{T}\left((T-1)(6-4T)\left[\sum_{i=1}^{N}\pi_i^2\right]^2 + 4(T-1)(T-2)\sum_{i=1}^{N}\pi_i^3 + 2(T-1)\sum_{i=1}^{N}\pi_i^2\right).$$

For fixed values of N, T and of the noncentrality parameter $(T-1)N\sum_{i=1}^{N}\left(\pi_i - \frac{1}{N}\right)^2$ (i. e. fixed value of $\sum_{i=1}^{N}\pi_i^2$) the extremal values of $\sum_{i=1}^{N}\pi_i^3$ (and, consequently, $\mathbf{D}X^2_{N,T}$) are realized on the sets of probabilities of the form $(u_1, \ldots, u_1, u_2, \ldots, u_2, 0, \ldots, 0)$.

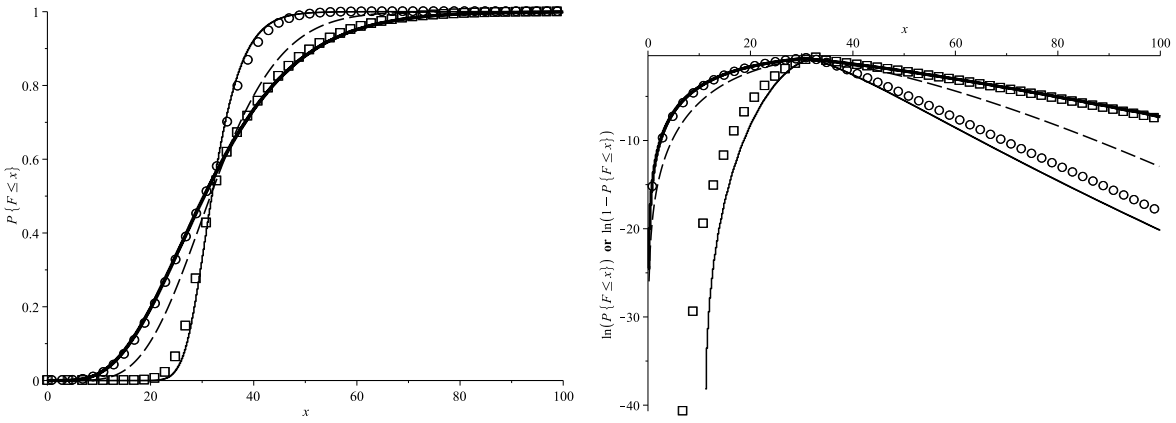


Figure 1: Distribution functions of $X^2_{10,100}$ with extremal values of $\mathbf{D}X^2_{10,100}$ and logarithms of their tails for $\lambda = 23.56$, and of noncentral chi-square with 9 degrees of freedom and noncentrality parameter $\lambda = 23.56$.

On the left part of Fig.1 for the case $N = 10$, $T = 100$, $\lambda = 23.56$ the graph of noncentral chi-square distribution with 9 degrees of freedom and noncentrality parameter λ (dotted line) and the graphs of exact distributions of $X^2_{10,100}$ for sets of probabilities

realizing the minimal and maximal variances are presented; squares and circles correspond to minimal and maximal values of distribution function of $X^2_{10,100}$ observed for the random sample of sets of probabilities $\pi_1, \ldots, \pi_N$ giving $\lambda = 23.56$. On the right part of Fig.1 for the same distribution functions $F(x)$ the graphs of $\ln \min\{F(x), 1 - F(x)\}$ are presented.

For the same parameters $N = 10$, $T = 100$ the differences between distribution functions of $X^2_{10,100}$ with maximal (minimal) variance and of non-central chi-square distribution with 9 degrees of freedom for three values of λ (1.14, 7.40, 23.56) are shown in the upper part of Fig.2. In the lower part of Fig.2 the corresponding differences between logarithms of tails are shown.

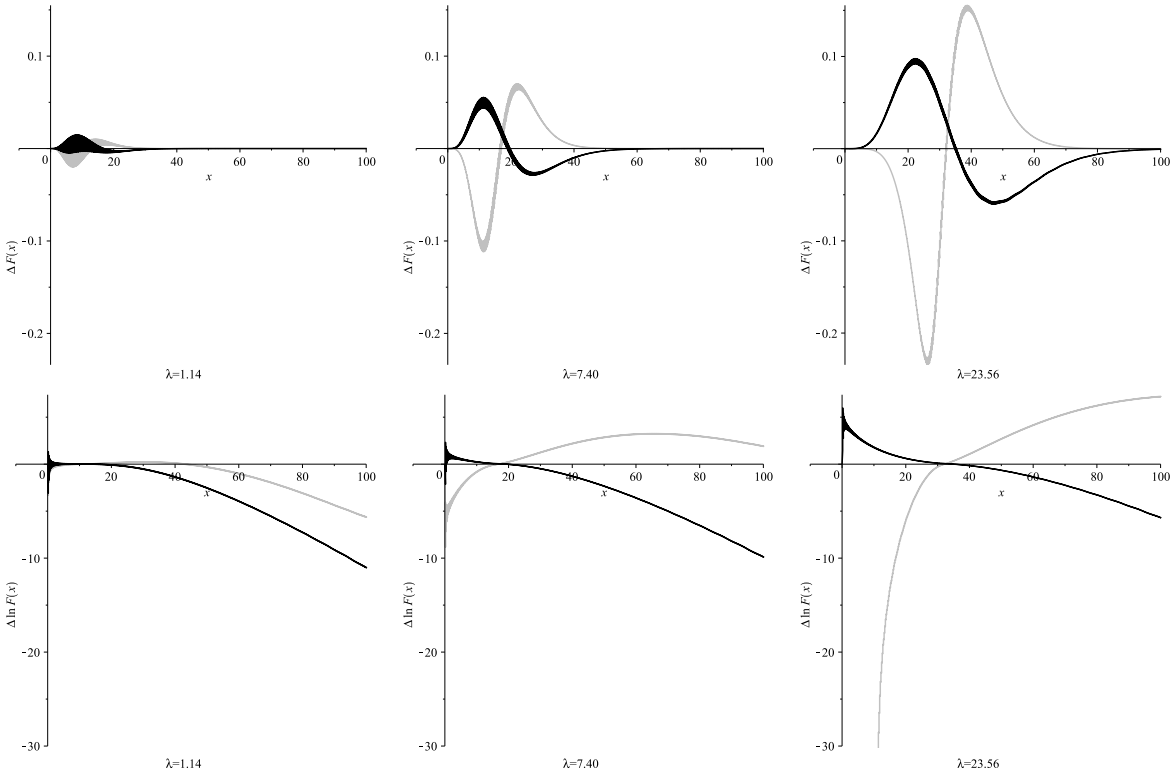


Figure 2: Differences between distribution functions and logarithms of tails for $N = 10, T = 100$.

If outcome probabilities $p_1, \ldots, p_N$ are not equal, then the differences between distributions of the Pearson statistics (1) computed for samples with outcome probabilities $\pi_1, \ldots, \pi_N$ with fixed value of the noncentrality parameter (i. e. the mean) appears to be larger and the sets of such N-dimensional vectors $\pi_1, \ldots, \pi_N$ are asymmetrical. So, the investigation of forms and sizes of accurate confidence sets of probabilities based on the values of Pearson statistics as well as the power of tests appears to be a nontrivial problems.

References

[1] Broffitt J.D., Randles R.H. (1977). A power approximation for the chi-square goodness-of-fit test: Simple hypothesis case. *J. Amer. Stat. Assoc.* Vol. **72**(359), pp. 604-607.

[2] Yarnold J.K. (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution. *J. Amer. Stat. Assoc.* Vol. **65**(330), pp. 864-886.

[3] Yarnold J.K. (1972). Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set. *Ann. Math. Stat.* Vol. **43**(5), pp. 1566-1580.

[4] Zubkov A.M., Filina M.V. (2008). Exact computation of Pearson statistics distribution and some experimental results. *Austrian J. Stat.* Vol. **37**(1), pp. 129-135.

[5] Zubkov A.M., Filina M.V. (2011). Tail properties of Pearson statistics distributions. *Austrian J. Stat.* Vol. **40**(1,2), pp. 47-54.

MODERN EMPIRICAL LIKELIHOOD CONCEPTS

Gregory Gurevich[1], Albert Vexler[2], Yang Zhao[2]
[1]*Dep. Industrial Engineering and Management, Shamoon College of Engineering*
Beer Sheva, ISRAEL
[2]*Dep. Biostatistics, State University of New York*
Buffalo, USA
e-mail: [1]gregoryg@sce.ac.il

Abstract

Statistical strategies to make decisions via formal rules play important roles in statistical and engineering practice. When the forms of data distributions are specified, the likelihood ratio principle is a central doctrine for developing statistical decision-making mechanisms in various experiments. However, it is well known that when key assumptions are not met, parametric likelihood procedures may be suboptimal or biased. One very important issue in statistical and engineering research is to preserve efficiency of the statistical inference through the use of robust likelihood-type techniques. Towards this end, the modern statistical literature has shifted focus towards robust and efficient nonparametric likelihood methods. In this note we present and shortly outline recently developed empirical likelihood (EL) techniques. In particular, we show that since EL techniques and parametric likelihood methods are closely related concepts, one may apply corresponding EL functions to replace their parametric likelihood counterparts in known and well developed parametric procedures, constructing novel nonparametric methods.

1 Introduction

The likelihood principle is one of the most important concepts for inference in parametric models. Neyman and Pearson [4] provided strong arguments that show the likelihood ratio approach can lead to most powerful statistical decision-making rules according to the Neyman-Pearson (NP) lemma. The recently proposed EL methodology employs the likelihood concept in a distribution-free fashion, approximating optimal parametric likelihood-based procedures (e.g., Qin and Lawless [6], Lazar and Mykland [2], Owen [5], Lazar [3], Vexler and Gurevich [7, 8], Vexler et al. [9, 10]). Similarly to the parametric likelihood concept, the EL methodology provides relatively simple strategies to construct powerful statistical tests that can be applied in various complex statistical and engineering studies. In this note we outline the EL methodology and its applications.

2 The EL methodology

The classical EL methodology, which is a *distribution function-based* approach, has been shown to have attractive properties for testing hypotheses regarding parameters

(e.g. moments) of distributions (e.g., Vexler et al. [11]). Let $X_1, ..., X_k$ denote independent and identically distributed (iid) data points from a distribution function F that corresponds to a density function f. The EL function has the form $L_p = \prod_{i=1}^k p_i$, where the estimated probability weights $p_i, i = 1, ..., k$, maximize L_p and satisfy empirical constraints corresponding to hypotheses of interest. For example, if the null hypothesis is $H_0 : E(X_1) = 0$, then the values of p_i's in the H_0-EL L_p should be chosen to maximize L_p given $\sum_{i=1}^k p_i = 1$ and $\sum_{i=1}^k p_i X_i = 0$, where the constraint $\sum_{i=1}^k p_i X_i = 0$ is an empirical version of $E(X_1) = 0$. Computation of $p_i, i = 1, ..., k$, is based on a simple exercise in Lagrange multipliers. This nonparametric approach is a result of consideration of the 'distribution functions'-based likelihood $\prod_{i=1}^k (F(X_i) - F(X_i-))$ over all distribution functions F (see [5] for details).

According to the NP lemma, the most powerful test statistics have structures that are related to density-based (DB) likelihood ratios. Motivated by this fact, alternatively to the 'distribution functions'-based EL methodology, Vexler and Gurevich [7, 8] proposed to use the central idea of the EL technique to develop DB empirical approximations to the likelihood $L_f = \prod_{i=1}^k f(X_i)$. To outline this technique, we represent the likelihood function L_f in the form

$$L_f = \prod_{i=1}^k f(X_i) = \prod_{i=1}^k f(X_{(i)}) = \prod_{i=1}^k f_i, \tag{1}$$

where $f_i = f(X_{(i)})$, and $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(k)}$ are order statistics based on $X_1, ..., X_k$. Following the EL method, one can obtain estimated values of $f_i, i = 1, ..., k$, that maximize L_f and satisfy an empirical version of the constraint $\int f(u)du = 1$. To formalize this constraint, Vexler and Gurevich proposed the following result [7].

Proposition 1. *Assume $X_{(j)} = X_{(1)}$, if $j \leq 1$, and $X_{(j)} = X_{(k)}$, if $j \geq k$. Then for $f(u)du$ and all integer m, we have:*

$$\sum_{j=1}^k \int_{X_{(j-m)}}^{X_{(j+m)}} = 2m \int_{X_{(1)}}^{X_{(k)}} - \sum_{l=1}^{m-1} (m-l) \int_{X_{(k-l)}}^{X_{(k-l+1)}} - \sum_{l=1}^{m-1} (m-l) \int_{X_{(l)}}^{X_{(l+1)}}.$$

Denote $H_m = \frac{1}{2m} \sum_{j=1}^k \int_{X_{(j-m)}}^{X_{(j+m)}} f(x)dx$. Since $\int_{X_{(1)}}^{X_{(k)}} f(x)dx \leq \int_{-\infty}^{+\infty} f(x)dx = 1$, Proposition 1 shows that $H_m \leq 1$, as well as, one can expect that $H_m \approx 1$, when $m/k \to 0$ as $m, k \to \infty$.

Taking into account definitions of hypotheses for which we need to test, one can empirically approximate $\int_{X_{(j-m)}}^{X_{(j+m)}} f(x)dx$, e.g., via $\int_{X_{(j-m)}}^{X_{(j+m)}} f(x)dx \cong (X_{(j+m)} - X_{(j-m)}) f_i$ and then represent the condition $H_m \leq 1$ in an empirical form, for example

$$\tilde{H}_m \leq 1, \tilde{H}_m = \frac{1}{2m} \sum_{j=1}^k (X_{(j+m)} - X_{(j-m)}) f_j. \tag{2}$$

This implies that we can obtain values of $f_i, i = 1, ..., k$, that maximize L_f and satisfy an empirical version of the constraint $H_m \leq 1$. For example, when the constraint (2) is

in effect, the Lagrange technique results in $f_i = 2m(k(X_{(i+m)} - X_{(i-m)}))^{-1}, i = 1, ..., k$ (here $X_{(j)} = X_{(1)}$, if $j \leq 1$ and $X_{(j)} = X_{(k)}$, if $j \geq k$) that gives the DB EL in the simple form

$$\prod_{i=1}^{k} 2m(k(X_{(i+m)} - X_{(i-m)}))^{-1}. \tag{3}$$

3 Applications of the DB EL approach

The Goodness-of-Fit tests: Consider a statement, when using iid observations $X_1, ..., X_k$, we want to test the hypothesis

$$H_0 : X_1, ..., X_k \sim F_0 \text{ versus } H_1 : X_1, ..., X_k \sim F_1, \tag{4}$$

where F_0 and F_1 are some distributions with density functions $f_0(x)$ and $f_1(x)$, respectively. The NP lemma stays that the most powerful test-statistic for (4) is the likelihood ratio

$$\prod_{i=1}^{k} f_1(X_i) \Big[\prod_{i=1}^{k} f_0(X_i) \Big]^{-1}. \tag{5}$$

If the alternative distribution function F_1 is unknown, the likelihood function at the numerator of (5) can be approximated using the DB EL (3). This provides the test-statistic

$$T_{mk} = \prod_{i=1}^{k} \frac{2m}{k(X_{(i+m)} - X_{(i-m)})} \Big[\prod_{i=1}^{k} f_0(X_i) \Big]^{-1}. \tag{6}$$

The power of the tests based on the statistic T_{mk} strongly depends on values of m. Using maximum likelihood type considerations, Vexler and Gurevich proposed the test statistic [7]:

$$T_k^* = \min_{1 \leq m < k^{1-\delta}} \Big(\prod_{i=1}^{k} \frac{2m}{k\big(X_{(i+m)} - X_{(i-m)}\big)} \Big/ \prod_{i=1}^{k} f_0(X_i) \Big), \ 0 < \delta < 1,$$

as an improvement of the test statistic T_{mk}. The authors also constructed DB EL goodness of fit tests for several scenarios when the function $f_0(x)$ is known up to parameters.

The two-sample tests: Let $X_1, ..., X_n$ and $Y_1, ..., Y_n$ be independent samples that consist of iid observations from distribution functions F_X and F_Y with density functions $f_X(x)$ and $f_Y(y)$, respectively. The problem is to test for

$$H_0 : F_Y = F_X = F_Z \text{ versus } H_1 : F_Y \neq F_X \tag{7}$$

where distributions F_Z, F_X and F_Y are unknown. In this case, the likelihood ratio statistic is

$$\prod_{i=1}^{n} f_X(X_i) \prod_{j=1}^{k} f_Y(Y_j) \Big[\prod_{i=1}^{n} f_Z(X_i) \prod_{j=1}^{k} f_Z(Y_j) \Big]^{-1} = \prod_{i=1}^{n} f_{X,i} \prod_{j=1}^{k} f_{Y,j} \Big[\prod_{i=1}^{n} f_{ZX,i} \prod_{j=1}^{k} f_{ZY,j} \Big]^{-1}, \tag{8}$$

where a density function f_Z corresponds to F_Z, $f_{X,i} = f_X(X_{(i)}), f_{Y,j} = f_Y(Y_{(j)})$, and $f_{ZX,i} = f_Z(X_{(i)}), f_{ZY,j} = f_Z(Y_{(j)})$, $i = 1, ..., n, j = 1, ..., k$; $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$, $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(k)}$ are the order statistics based on the observations $X_1, ..., X_n$ and $Y_1, ..., Y_k$, respectively. Gurevich and Vexler [1] applied the method of the DB EL to approximate the ratio (8). For example, the "$H_m \leq 1$"-type constraint mentioned above, with respect to $f_{X,i} = f_X(X_{(i)}), i = 1, ..., n$ can be rewritten using the hypothesis context as

$$H_m \leq 1, H_m = \frac{1}{2m} \sum_{i=1}^{n} \int_{X_{(i-m)}}^{X_{(i+m)}} \frac{f_X(u)}{f_Z(u)} f_Z(u) du.$$

In a similar manner to deriving the DB EL (3) with the constraint (2), by applying the approximate analog to the mean-value integration theorem, Gurevich and Vexler [1] defined the DB EL test-statistic for (7) in the form $V_{nk} = ELR_{X,n} ELR_{Y,k}$, where $ELR_{X,n} = \min\limits_{a_n \leq m \leq b_n} \prod\limits_{i=1}^{n} 2m \Big[n \Big(\big(F_{Z(n+k)}\big(X_{(i+m)}\big) - F_{Z(n+k)}\big(X_{(i-m)}\big) \big) \Big) \Big]^{-1}, X_{(j)} = X_{(1)}$, if $j \leq 1, X_{(j)} = X_{(n)}$, if $j \geq n; ELR_{Y,k} = \min\limits_{a_k \leq r \leq b_k} \prod\limits_{i=1}^{k} 2r \Big[k \Big(F_{Z(n+k)}\big(Y_{(i+r)}\big) - F_{Z(n+k)}\big(Y_{(i-r)}\big) \Big) \Big]^{-1}, Y_{(j)} = Y_{(1)}$, if $j \leq 1; Y_{(j)} = Y_{(k)}$, if $j \geq k; F_{Z(n+k)}(u) = \frac{1}{n+k} \Big(\sum_{i=1}^{n} I(X_i \leq u) + \sum_{j=1}^{k} I(Y_j \leq u) \Big); a_l = l^{0.5+\delta}, b_l = min\big(l^{1-\delta}, \frac{l}{2}\big), \delta \in (0, 0.25), l = n, k$. This EL ratio test-statistic V_{nk} approximates the optimal likelihood ratio (8).

4 Conclusions

Similarly to the parametric likelihood concept, the EL methodology provides relatively simple strategies to construct powerful statistical tests that can be applied in various studies. The extreme generality of EL methods and their wide range of usefulness partly result from the simple derivation of the EL statistics as components of composite parametric likelihood based systems, efficiently attending to any observed data and relevant information. Note that EL based methods are employed in much of modern statistical practice, and we cannot describe all relevant theory and examples. The reader interested in the EL methods will find more details and many pertinent articles in recent statistical journals publications.

References

[1] Gurevich G., Vexler A. (2011). A two-sample empirical likelihood ratio test based on samples entropy. *Statistics and Computing*, Vol. **21**, pp. 657-670.

[2] Lazar N.A., Mykland P.A. (1998). An evaluation of the power and conditionality properties of empirical likelihood. *Biometrika*, Vol. **85**, pp. 523-534.

[3] Lazar N.A. (2003). Bayesian empirical likelihood. *Biometrika*, Vol. **90**, pp. 319-326.

[4] Neyman J., Pearson E.S. (1992). *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, Springer.

[5] Owen A. (2001). *Empirical Likelihood*. Chapman & Hall, CRC, Boca Raton.

[6] Qin J., Lawless J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, Vol. **22**, pp. 300-325.

[7] Vexler A., Gurevich G. (2010). Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Computational Statistics and Data Analysis*, Vol. **54**, pp. 531-545.

[8] Vexler A., Gurevich G. (2010). Density-based empirical likelihood ratio change point detection policies. *Communications in Statistics-Simulation and Computation*, Vol. **39**, pp. 1709-1725.

[9] Vexler A., Gurevich G., Hutson A.D. (2013). An Exact Density-Based Empirical Likelihood Ratio Test for Paired Data. *Journal of Statistical Planning and Inference*, Vol. **143**, pp. 334-345.

[10] Vexler A., Tsai W-M., Gurevich G., Yu J. (2012). Two-sample density-based empirical likelihood ratio tests based on paired data, with application to a treatment study of Attention-Deficit/Hyperactivity Disorder and Severe Mood Dysregulation. *Statistics in Medicine*, Vol. **31**, pp. 1821-1837.

[11] Vexler A., Yu J., Hutson A.D. (2011). Likelihood Testing Populations Modeled by Autoregressive Process Subject to the Limit of Detection in Applications to Longitudinal Biomedical Data. *Journal of Applied Statistics*, Vol. **38**, pp. 1333-1346.

MODELING UNBIASED ESTIMATORS WITH GOOD ASYMPTOTIC PROPERTIES FOR THE SUM OF MULTIVARIATE DISCRETE INDEPENDENT RANDOM VARIABLES

A. S. Iskakova
Gumilyov Eurasian National University
Astana, KAZAKHSTAN
e-mail: ayman.astana@gmail.com

Abstract

In a multivariate discrete probability model of distribution of sum discrete random variables is proposed and studied. The concept of the most appropriate of the set of unbiased estimators, which has good asymptotic properties, is introduced.

1 Introduction

Multivariate probabilistic models, as a reflection of the current reality, are absolutely necessary for describe events and situations encountered in daily life. In recent years, a considerable amount of probabilistic models have been developed. However, there are many unsolved problems, for example, in the implementation of monitoring the it is clear only the sum of components, which as a result of observations can not be detected. So far, probabilistic models describing similar situations were not considered.

An exceptional example of the actual use of such a model is the advertising industry, where it is necessary to link the distribution of consumer interests with appropriate advertising in various sources. Similar problems are very common in meteorology and other fields. In this paper we present statistical evaluation of the distribution of sums of unobservable random matrices $\mathbf{L}_1, \ldots, \mathbf{L}_d$ by their amount. Thus, the results of the proposed work can solve many of these problems.

2 Multivariate discrete probability distribution of sum of discrete random variables

Assume that the true image can be represented as a matrix $\mathbf{l}_0 = \|l_{0_{i,j}}\|_{m\times q}$, which imposed distortion, consisting of four factors (matrices) of losses $\mathbf{u} = \|u_{i,j}\|_{m\times q}$, taking values from the set of $\mathbf{l}_1, \ldots, \mathbf{l}_d$.

Obviously, the factors (the matrix), the loss $\mathbf{l}_1, \ldots, \mathbf{l}_d$ are realizations of random matrices $\mathbf{L}_1, \ldots, \mathbf{L}_d$, appearing with probabilities $\mathbf{p} = (p_1, \ldots, p_d)$.

Assume that $V_{\mathbf{u}}$ is the number of possible combinations $r_{1_{v_{\mathbf{u}}}}\mathbf{L}_1, \ldots, r_{d_{v_{\mathbf{u}}}}\mathbf{L}_d$, which together form a matrix of $\mathbf{u}$, where $r_{1_{v_{\mathbf{u}}}}, \ldots, r_{d_{v_{\mathbf{u}}}}$ determine the possible number of balls taken out, which marked the relevant matrices $\mathbf{L}_1, \ldots, \mathbf{L}_d$. In other words, from [2] it follows that $V_{\mathbf{u}}$ is the number of partitions on the part of the matrix $\mathbf{u}$ on $\mathbf{L}_1, \ldots, \mathbf{L}_d$.

Theorem 1. *The distortion is distributed as follows:*

$$P(\mathbf{U}=\mathbf{u})=\sum_{v_{\mathbf{u}}=1}^{V_{\mathbf{u}}} n!\prod_{\alpha=1}^{d}\frac{p_{\alpha}^{r_{\alpha v_{\mathbf{u}}}}}{r_{\alpha_{v_{\mathbf{u}}}}!}. \tag{1}$$

3 Unbiased estimation of the probability distribution of the proposed model

In practice, as a rule, elements of the vector $\mathbf{p}=(p_1,\ldots,p_d)$ are not known. It is also not known matrix $\mathbf{L}_1,\ldots,\mathbf{L}_d$. Consequently, formula (1) does not find the actual application.

Assume that there are photos in the number of k particular locality with the distortions $\mathbf{x}=\{\mathbf{x}_1,...,\mathbf{x}_k\}$. In other words, a number of evidence-x can be interpreted as a realization of a sample of k, whose elements are subject to distribution (1). We denote $\mathbf{r}_{v_\beta}$ vector $(r_{1_{v_\beta}},\ldots,r_{d_{v_\beta}})$, which defines v_β-th solution of equation

$$\begin{cases}\sum\limits_{\alpha=1}^{d} L_\alpha r_{\alpha_{v_\beta}}=\mathbf{u},\\ \sum\limits_{\alpha=1}^{d} r_{\alpha_{v_\beta}}=n,\end{cases} \tag{2}$$

where $v_\beta=1,\ldots,V_\beta, V_\beta$ is the number of partitions of the matrix $\mathbf{x}_\beta$ on the matrices $\mathbf{L}_1,\ldots,\mathbf{L}_d$. Using the system of equations (2), the matrices $\mathbf{L}_1,\ldots,\mathbf{L}_d$, and the actual data $\mathbf{x}$, we define for each $\beta=1,\ldots,k$ the number of partitions V_β matrix $\mathbf{x}_\beta$ at $\mathbf{L}_1,\ldots,\mathbf{L}_d$, and vectors $\mathbf{r}_{1_\beta},\ldots,\mathbf{r}_{V_\beta}$.

Suppose that for each $j=1,\ldots,\mu$, where $\mu=\prod\limits_{\beta=1}^{k} V_\beta$, there is a vector $\mathbf{z}_j=(z_{1_j},\ldots,z_{d_j})$, defined as $\mathbf{z}_j=\sum\limits_{\beta=1}^{k}\mathbf{r}_{v_\beta}$, and the indices on the right and left side are linked one-to-one correspondence, which is not unique.

Thus, from the above lemma that if some element of the implementation of the sample $\mathbf{x}=(\mathbf{x}_1,\ldots,\mathbf{x}_k)$ of the distribution (1) has more than one partition on the submitted part, it is impossible, using the theorem Rao-Blackwell-Kolmogorov construct an unbiased estimate with minimum variance for the probability distribution (1).

Theorem 2. *The following statistics form an unbiased estimate for the probability distribution (1):*

$$W(\mathbf{u},\mathbf{z}_j)=\frac{\sum\limits_{v_{\mathbf{u}}=1}^{V_{\mathbf{u}}}\prod\limits_{\alpha=1}^{d}\binom{z_{\alpha_j}}{r_{\alpha v_{\mathbf{u}}}}}{\binom{nk}{n}},\ j=1,\ldots,\mu, \tag{3}$$

where $V_{\mathbf{u}}$ is the number of partitions on the part of the matrix $\mathbf{u}$, $\mathbf{L}_1,\ldots,\mathbf{L}_d$; for each partition $r_{1_{v_{\mathbf{u}}}},\ldots,r_{d_{v_{\mathbf{u}}}}$ determine the possible number of matrices $\mathbf{L}_1,\ldots,\mathbf{L}_d$; $k\geq 1$ and $z_{\alpha_j}\geq r_{\alpha_{v_{\mathbf{u}}}}$, when $\alpha=1,\ldots,d$, $v_{\mathbf{u}}=1,\ldots,V_{\mathbf{u}}$.

4 The most suitable unbiased estimates for the probability distribution of the proposed model and their properties

Thus, we have a lot of unbiased estimates of the probability of distortion.

Definition 1. Decision $\mathbf{z}_g$, based on observation, is the most appropriate set of $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$, if

$$\prod_{\beta=1}^{k} W(\mathbf{x}_\beta, \mathbf{z}_g) = \max_{j=1,\ldots,\mu} \prod_{\beta=1}^{k} W(\mathbf{x}_\beta, \mathbf{z}_j), \tag{4}$$

where for $\beta = 1, \ldots, k$ elements of $W(\mathbf{x}_\beta, \mathbf{z}) = \{W(\mathbf{x}_\beta, \mathbf{z}_1), \ldots, W(\mathbf{x}_\beta, \mathbf{z}_\mu)\}$ forms an unbiased estimate for the probability distribution (1) defined in (5).

Definition 2. Unbiased estimate of $W(\mathbf{x}_\beta, \mathbf{z}_g)$ for the probability distribution (1) is the most suitable from the entire set of unbiased estimates of $W(\mathbf{x}_\beta, \mathbf{z}) = \{W(\mathbf{x}_\beta, \mathbf{z}_1), \ldots, W(\mathbf{x}_\beta, \mathbf{z}_\mu)\}$ defined in (5), if $\mathbf{z}_g$ is the most appropriate solution, based on observation.

Theorem 3. *The most suitable unbiased estimate of $W(\mathbf{x}_\beta, \mathbf{z}_g)$ for the probability distribution (1) is consistent, asymptotically normal and asymptotically efficient.*

Let us summarize the results:

- proposed and studied a new probability distribution of discrete random variables;
- developed an algorithm for computing the probability and define the generating function for the distribution of the proposed model;
- the set of unbiased estimates for the probability distribution of the proposed model and the variance of these estimates;
- introduced a new concept of the most appropriate evaluation of the set of unbiased estimates, with asymptotic properties.

References

[1] Iskakova A.S. (2014). Construction of the most suitable unbiased estimate distortions of radiation processes from remote sensing data. *Journal of Physics: Conference Series*. Vol. **490**(1), id 012113.

[2] Andrews G. E. (1976). *The Theory of Partitions. Encyclopedia of Mathematics and Its Applications*. Addison-Wesley, London.

ON ONE GENERALIZATION OF MARKOV CHAIN WITH PARTIAL CONNECTIONS

Yu. S. Kharin[1], M. V. Maltsew[2]
Belarusian State University
Minsk, BELARUS
e-mail: [1]kharin@bsu.by, [2]maltsew@bsu.by

Abstract

The paper deals with homogenous s-order vector Markov chain with partial connections. Conditional distribution for this model depends only on finite number of components of previous vector states. Statistical estimators for model parameters are constructed.

1 Introduction

Markov chain is a broadly used mathematical model of discrete time series. It is applied in economics [1], biology [2], sociology [3] and other fields. Markov chain of the order s [4] is an adequate model for description of high-depth dependences in data. Since data is often represented in blocks, it is reasonable to use vector Markov chains. The state space for such models consists of fixed length vectors. Unfortunately, it is difficult to use s-order Markov chain in practice, because the number of parameters D for the model with N states increases exponentially when s growth: $D = (N-1)N^s$. That is why small-parametric or parsimonious models are used in applications [5]. For such models D depends polynomially on s. Markov chain of order s with r partial connections (MC(s,r)) is an example of a parsimonious model. It was developed in Belarusian state university [6]. Conditional distribution for this model does not depend on all s previous states but only on r selected states. In this paper we propose a generalization of the MC(s,r) for vector Markov chain.

2 Mathematical model

Introduce the notation: $\mathbb{N}$ is the set of positive integers; $A = \{0, 1, \ldots, N-1\}$ is the state space with N elements, $2 \le N < \infty$; $m \in \mathbb{N}$, $\bar{N} = N^m$, $\bar{A} = A^m$, $J_i = (j_{i1}, \ldots, j_{im}) \in \bar{A}$, $i = 1, 2, \ldots$ is a m-dimensional vector; $J_a^b = (J_a, \ldots, J_b)$, $a, b \in \mathbb{N}$, $a \le b$, is an ordered set of $b-a+1$ m-dimensional vectors; $\{x_t \in \bar{A} : t \in \mathbb{N}\}$ is a homogeneous vector Markov chain of the order s ($2 \le s < \infty$) with the following parameters:

$$\pi^{(0)}_{J_1,\ldots,J_s} = \mathrm{P}\{x_1 = J_1, \ldots, x_s = J_s\}$$

is an initial probability distribution;

$$P = (p_{J_1^s, J_{s+1}}) \tag{1}$$

is a $(s+1)$-dimensional matrix of transition probabilities:

$$p_{J_1^s,J_{s+1}} = \mathrm{P}\{x_t = J_{s+1} | x_{t-1} = J_s, \ldots, x_{t-s} = J_1\},\ t = s+1, s+2, \ldots$$

We will denote this Markov chain VMC(s) (Vector Markov Chain of the order s).

The number of independent parameters for the VMC(s) is determined by formula:

$$D_s = \bar{N}^s(\bar{N} - 1).$$

In the Table 1 we present the number of parameters for the binary VMC(s) when $m = 8$ for various values of s .

Table 1: The number of parameters for the binary VMC(s)

s	1	2	4	8	16
D_s	65280	16711680	$\approx 1,095 \cdot 10^{12}$	$\approx 4,704 \cdot 10^{21}$	$\approx 8,677 \cdot 10^{40}$

Table 1 illustrates the "curse of dimensionality" for s-order Markov chain. To overcome this difficulty we construct modification of the VMC(s) by analogy with [6]. We will use the notation:

$$M_r = \{(k_1, l_1), (k_2, l_2), \ldots, (k_r, l_r)\} \subseteq M_* = \{(k, l) : 1 \le k \le s, 1 \le l \le m\}$$

is an ordered set of $1 \le r \le sm$ pairs of indices, which we will call template-set, there exists a pair (k_i, l_i) such that $k_i = 1$; $\mathbf{M}_r$ is a set of all possible template-sets; $S_{M_r}(J_t, \ldots, J_{t+s-1}) = (j_{t+k_1-1,l_1}, \ldots, j_{t+k_r-1,l_r})$, $t = 1, 2, \ldots$ is a selector function, that associates s vectors with their r components: $S_{M_r} : \bar{A}^s \to A^r$; $Q = (q_{(i_1,\ldots,i_r),I_{r+1}})$ is a stochastic $N^r \times N^m$ matrix, $i_1, \ldots, i_r \in A$, $I_{r+1} \in \bar{A}$.

The Markov chain $\{x_t \in \bar{A} : t \in \mathbb{N}\}$ is called the vector Markov chain of the order s with r partial connections, if its transition probabilities have the following form:

$$p_{J_1^s,J_{s+1}} = q_{S_{M_r}(J_1,\ldots,J_s),J_{s+1}} = q_{(j_{k_1,l_1},\ldots,j_{k_r,l_r}),J_{s+1}}, \tag{2}$$

We will denote this model VMC(s, r).

The definition of the VMC(s, r) means that probability distribution of time series x_t at time t does not depend on all ms components of s previous states, but it depends only on r selected components determined by template-set M_r. If $r = sm$, then $M_r = M_*$ and we have fully-connected s-order Markov chain: VMC(s, ms) = VMC(s). If $m = 1$, then the VMC(s, ms) transforms into the Markov chain with partial connections [6].

The number of parameters for the VMC(s, r) is determined by formula:

$$d = N^r(N^m - 1) + 2r - 1. \tag{3}$$

In the Table 2 we present the number of parameters for the binary VMC(s, r) when $m = 8$ for various values of s and r .

Table 2: The number of parameters for the binary VMC(s,r)

(s,r)	(1, 2)	(2, 4)	(4, 6)	(8, 8)	(16, 10)	(32, 16)
d	1 023	4 087	16 331	65 295	261 139	16 711 711

3 Statistical Estimators for parameters

Let us construct now statistical estimators for VMC(s,r) parameters. Introduce the notation: $X^{(n)} \in \bar{A}^n$ is the observed time series of length n; $\mathbf{I}\{C\}$ is the indicator function of event C;

$$\nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1}) = \sum_{t=1}^{n-s} \mathbf{I}\{S_{M_r}(X_t,\ldots,X_{t+s-1}) = (i_1,\ldots,i_r), X_{t+s} = I_{r+1}\},$$

$$\nu_{s}^{M_r}(i_1,\ldots,i_r) = \sum_{I_{r+1}\in\bar{A}} \nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1}),\ (i_1,\ldots,i_r)\in A^r,\ I_{r+1}\in A^m,$$

are frequency statistics of VMC(s,r).

The loglikelihood function for the VMC(s,r) has the following form:

$$l_n(X^{(n)},Q,M_r) = \ln \pi^{(0)}_{X_1,\ldots,X_s} + \sum_{\substack{i_1,\ldots,i_r\in A,\\ I_{r+1}\in\bar{A}}} \nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1}) \ln q_{(i_1,\ldots,i_r),I_{r+1}}.$$

If the true values s, r and M_r are known, then the maximum likelihood estimators (MLE) for the one-step transition probabilities (2) are

$$\hat{q}_{(i_1,\ldots,i_r),I_{r+1}} = \begin{cases} \dfrac{\nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1})}{\nu_{s}^{M_r}(i_1,\ldots,i_r)},\ if\ \nu_{s}^{M_r}(i_1,\ldots,i_r) > 0, \\ \\ 1/\bar{N},\ if\ \nu_{s}^{M_r}(i_1,\ldots,i_r) = 0. \end{cases}$$

If s and r are known, then MLE for template set M_r is

$$\hat{M}_r = arg \max_{M_r\in\mathbf{M}_r} \sum_{\substack{i_1,\ldots,i_r\in A,\\ I_{r+1}\in\bar{A}}} \nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1}) \ln \frac{\nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1})}{\nu_{s}^{M_r}(i_1,\ldots,i_r)}.$$

In order to estimate the order s and the number of connections r we use Bayesian information criterion (BIC) [7]:

$$(\hat{s},\hat{r}) = arg \min_{2\le s'\le s_+,\ 1\le r'\le r_+} BIC(s',r'),$$

$$BIC(s',r') = -l_n(X^{(n)},Q,M_r) + 2d\ln(n-s') =$$

$$= -\sum_{\substack{i_1,\ldots,i_{r'}\in A,\\ I_{r'+1}\in\bar{A}}} \nu_{s'+1}^{M_{r'}}(i_1,\ldots,i_{r'},I_{r'+1}) \ln \frac{\nu_{s'+1}^{M_{r'}}(i_1,\ldots,i_{r'},I_{r'+1})}{\nu_{s'}^{M_{r'}}(i_1,\ldots,i_{r'})} + 2d\ln(n-s'),$$

where $s_+ \ge 1$, $1 \le r_+ \le ms_+$ are maximal admissible values of s and r respectively, d is the number of independent parameters of the VMC(s,r), defined by formula (3).

References

[1] Kemeny J., Snell J. (1963). *Finite Markov chains*. Princeton NJ: D. Van Nostrand.

[2] Gibson M.C. et al. (2006). The emergence of geometric order in proliferating metazoan epithelia. *Nature*. Vol. **442**(7106). pp. 1038-1041.

[3] Bonacich P. (2003). Asymptotics of a matrix valued Markov chain arising in sociology. *Stochastic Processes and their Applications*. Vol. **104**(1), pp. 155-171.

[4] Doob J.L. (1953). *Stochastic processes*. Wiley, NY.

[5] Kharin Yu. (2012). Parsimonious models for high-order Markov chains and their statistical analysis. *VIII World Congress on Probability and Statistics*. Publ. House of Koc. Univ.: Istanbul. pp. 168-169.

[6] Kharin Yu.S, Petlitskii A.I. (2007). A Markov chain of order s with r partial connections and statistical inference on its parameters. *Discrete Mathematics and Applications*. Vol. **17**(3), pp. 295-317.

[7] Csiszar I., Shields P.C. (1999). Consistency of the BIC order estimator. *Electronic research announcements of the American math. society*. Vol. **5**, pp. 123-127.

ON ROBUSTNESS OF CONFIGURATION GRAPHS WITH RANDOM NODE DEGREE DISTRIBUTION

M. M. Leri
Institute for Applied Mathematical Research, Karelian Research Centre of RAS
Petrozavodsk, RUSSIA
e-mail: leri@krc.karelia.ru

Abstract

We consider power-law configuration graphs with node degrees drawn from the power-law distribution with the parameter following the uniform distribution on a chosen interval. By computer simulation we study the robustness of these graphs from a viewpoint of link saving in the two cases of destruction process: the "random breakdown" and the "targeted attack".

1 Introduction

The study of random graphs with node degrees following the power-law distribution continues to attract special interest (see e.g. [3], [5]). The use of such models has been widening with the changes in the structure of massive data networks and with the appearance of new ones. Power-law random graphs used to be considered a good representation of the AS-level topology (see e.g. [4], [7], [9]) and, moreover, variations of these models could be used in other applications. Along with the studies of the structure of present-day complex networks, the problem of their robustness and vulnerability to various types of breakdowns remains rather pressing (see e.g. [2], [3], [8]).

2 Power-law configuration random graph

We consider power-law random graphs with the number of nodes N. Random variables $\xi_1, \xi_2, \ldots, \xi_N$ are independent identically distributed variables drawn from the power-law distribution:

$$\mathbf{P}\{\xi \geq k\} = k^{-\tau}, \qquad \tau > 1,\ k = 1, 2, \ldots. \tag{1}$$

We use the graph construction procedure introduced in [1], where such models were first called configuration graphs. Starting with a predefined number of nodes we draw node degrees from the distribution (1) with the parameter τ following the uniform distribution on a predefined interval $(a, b]$. The node degree gives the number of stubs for each node, numbered in an arbitrary order. Then all the stubs are joined one to another equiprobably forming links. The sum of node degrees has to be even, otherwise one stub is added to a randomly chosen node to form a lacking connection. The graph construction allows loops and multiple links.

3 Robustness in random environment: link saving

The distribution (1) with the parameter $\tau \in (1,2)$ has finite expectation and infinite variance. As the value of τ exceeds 2 the variance of the distribution (1) becomes finite. The power-law configuration graphs with $\tau \in (1,2)$ are known to contain a so-called giant component ([1], [3], [9] etc.) – a connected set of nodes, the number of nodes in which has the expectation proportional to the number of graph nodes N, as $N \to \infty$.

In [6] we considered power-law graphs with the values of the parameter $\tau \in (1,2)$ fixed for each node. With the evolution of networks the value of τ is regarded to change not only within the stated interval $(1,2)$. It may also happen to be a random variable. Therefore here we consider the parameter τ being drawn from the uniform distribution on the interval $(a,b]$. As it was mentioned above the interval $(1,2)$ is interesting due to its application to the Internet graphs and the existence of the giant component. Power-law graphs with the parameter $\tau \in (2,3)$ do not contain the giant component, but are useful for the studies of forest fire models [6]. The interval $(1,3)$ was chosen as a generalization. As in the previous work [6], here we also consider the two types of breakdowns: a "targeted attack" on the nodes with the highest degrees and the "random breakdown" meaning the removal of equiprobably chosen nodes.

To conduct simulations we modeled graphs of the sizes $N \in [1000, 10000]$ with the three ranges of the parameter τ: $(1,2]$, $(1,3]$ and $(2,3]$. The purpose was to look at how the graph structure changes with the destruction of its nodes. Let random variables $\eta_1, \eta_2, \ldots, \eta_s$ be equal to the sizes of graph components in decreasing order, thus η_1 is the percentage of nodes in the largest component, η_2 – the percentage of nodes in the second-sized component, etc. Let s be the number of graph components. Let us consider a graph being destroyed if $\{\eta_1 \leq 2\eta_2\}$, which means that the size of the second largest component becomes greater or equal to half the size of the largest component. Thus we derived the regression relations between the size of the largest component η_1 and the percentage of nodes removed from the graph r. In the case of a "targeted attack" relations were as follows:

$$\begin{aligned}
\eta_1 &= 53.2 - 8.9r - 6.2\ln r, && \tau \in (1,2],\\
\eta_1 &= 31.9 - 7.0r - 9.1\ln r, && \tau \in (1,3],\\
\eta_1 &= -1.3 + 2.5r - 3.9\ln r, && \tau \in (2,3].
\end{aligned}$$

The determination coefficients (R^2) of these regression models are equal to 0.99, 0.98 and 0.96, respectively. For the process of "random breakdowns" we derived the following relations:

$$\begin{aligned}
\eta_1 &= 88.1 - 1.5r, && \tau \in (1,2],\\
\eta_1 &= 73.3 - 1.3r, && \tau \in (1,3],\\
\eta_1 &= 20.2 - 2.7\sqrt{r}, && \tau \in (2,3].
\end{aligned}$$

with determination coefficients 0.97, 0.95 and 0.99, respectively. The results showed that in all cases the graph size N does not affect the size of the largest component. As for the sizes of second-sized components they will diminish slightly with the removal

of graph nodes and will not exceed 20% when $\tau \in (1,2]$, 15% when $\tau \in (1,3]$, and 6% when $\tau \in (2,3]$ of graph nodes. The number of graph components in the case of a "targeted attack" slightly increases with the removal of nodes, although in the case of a "random breakdown" this number decreases.

In Figures 1 and 2 we plot the results of the estimation of the regression relations between the probabilities $\mathbf{P}\{\mathrm{A}\}$ (where A is the following event: $\{\eta_1 \leq 2\eta_2\}$) of graph destruction, the percentage of nodes removed from the graph r and the graph size N.

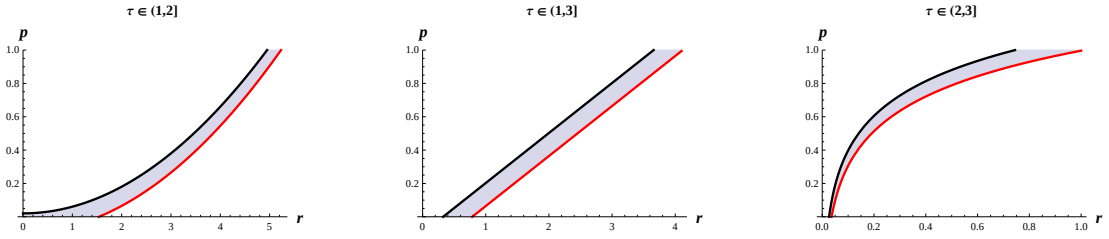


Figure 1: The probabilities of graph destruction in the case of a "targeted attack" (left-hand curve stands for $N = 10000$, right-hand curve – $N = 1000$).

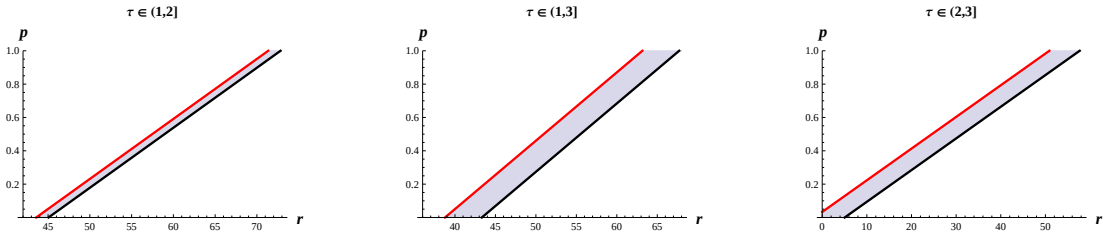


Figure 2: The probabilities of graph destruction in the case of a "random breakdown" (left-hand curve stands for $N = 1000$, right-hand curve – $N = 10000$).

Simulation results showed that power-law configuration graphs are much more robust to "random breakdowns" than to "targeted attacks" on the nodes with the highest degrees. To destroy such a graph by removing nodes with high degrees it is enough to take away $1 - 5\%$ of them. However, in the case of random nodes removal, the graph will be ruined by the destruction of $55 - 75\%$ of its nodes. The obtained results support previous conclusions [6] that the robustness of these graphs strongly depends on the value of the parameter τ. Thus, in the case when $\tau \in (2,3]$ graphs are more vulnerable to both targeted and random breakdowns than in the cases when $\tau \in (1,2]$ and $\tau \in (1,3]$.

4 Acknowledgements

The study is supported by the Russian Foundation for Basic Research, grant 16-01-00005.

References

[1] Bollobas B.A. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *Eur. J. Comb.*. Vol. **1**, pp. 311-316.

[2] Bollobas B., Riordan O. (2004). Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*. Vol. **1**, N **1**, pp. 1-35.

[3] Durrett R. (2007). *Random Graph Dynamics*. Cambridge Univ. Press, Cambridge.

[4] Faloutsos C., Faloutsos P., Faloutsos M. (1999). On power-law relationships of the Internet topology. *Computer Communications Rev.*. Vol. **29**, pp. 251-262.

[5] Hofstad R. (2011). *Random Graphs and Complex Networks*. Eindhoven University of Technology.

[6] Leri M., Pavlov Y. (2014). Power-law random graphs' robustness: link saving and forest fire model. *Austrian Journal of Statistics*. Vol. **43**, N **4**, pp. 229-236.

[7] Mahadevan P., Krioukov D., Fomenkov M., Huffaker B., Dimitropoulos X., claffy k., Vahdat A. (2006). The Internet AS-Level Topology: Three Data Sources and One Definitive Metric. *ACM SIGCOMM Computer Communication Review (CCR)*. Vol. **36**, N **1**, pp. 17-26.

[8] Norros I., Reittu H. (2008). Attack resistance of power-law random graphs in the finite mean, infinite variance region. *Internet Mathematics*. Vol. **5**, N **3**, pp. 251-266.

[9] Reittu H., Norros I. (2004). On the power-law random graph model of massive data networks. *Performance Evaluation*, Vol. **55**, pp. 3-23.

ASYMPTOTIC PROPERTIES OF BINARY SEQUENCES OBTAINED BY THE NEUMANN TRANSFORM

D. O. Menshenin
Lomonosov Moscow State University
Moscow, RUSSIA
e-mail: Dmitry.Menshenin@gmail.com

Abstract

In this paper we investigate the properties of binary sequences consisting of non-identically distributed dependent elements obtained by the Neumann transform. Some results for the asymptotic behaviour of joint distribution of such sequences are obtained.

1 Introduction

In the beginning of 1950s John von Neumann proposed a simple method to transform the sequence of independent identically distributed binary random variables into the sequence of independent random binary variables taking values 1 and 0 with probabilities $\frac{1}{2}$. This method was used to improve the quality of physical random number generators.

Let $\xi_1, \xi_2, \dots$ be a sequence of binary random variables and

$$\tau_1 = \min\{k \geq 1 : \xi_{2k-1} \neq \xi_{2k}\},\ \tau_{n+1} = \min\{k > \tau_n : \xi_{2k-1} \neq \xi_{2k}\}, n = 1, 2 \dots \quad (1)$$

The Neumann transform $\{\eta_t\}_{t=1}^{\infty}$ of the binary sequence $\{\xi_t\}_{t=1}^{\infty}$ is defined by the following rule:

$$\eta_t = \xi_{2\tau_t - 1}, \quad t = 1, 2 \dots \quad (2)$$

In what follows we suppose that $\xi_1, \xi_2, \dots$ are independent and $\mathbf{P}\{\xi_i = 1\} = p > 0$, $\mathbf{P}\{\xi_i = 0\} = q > 0$, $p + q = 1$, then

$$\mathbf{P}\{\xi_{2k-1} = 0 \,|\, \xi_{2k-1} \neq \xi_{2k}\} = \mathbf{P}\{\xi_{2k-1} = 1 \,|\, \xi_{2k-1} \neq \xi_{2k}\} = \frac{1}{2},$$

and therefore $\eta_1, \eta_2, \dots$ are independent and $\mathbf{P}\{\eta_t = 0\} = \mathbf{P}\{\eta_t = 1\} = \frac{1}{2}$.
Let us consider the Neumann transform $\{\eta'_t\}_{t=1}^{\infty}$ of the shifted binary sequence $\{\xi_t\}_{t=2}^{\infty}$ defined as follows:

$$\eta'_t = \xi_{2\tau'_t}, \quad t = 1, 2 \dots, \text{ where} \quad (3)$$

$$\tau'_1 = \min\{k \geq 1 : \xi_{2k} \neq \xi_{2k+1}\},\ \tau'_{n+1} = \min\{k > \tau'_n : \xi_{2k} \neq \xi_{2k+1}\}, n = 1, 2 \dots \quad (4)$$

It is clear that the elements of the sequences $\{\eta_t\}_{t=1}^{\infty}$ and $\{\eta'_t\}_{t=1}^{\infty}$ are dependent and joint distribution of the elements of the sequences is not trivial. In this paper we investigate the asymptotic behaviour of distributions of a pair (η_t, η'_t) and of vectors $(\eta_{t+1}, \dots, \eta_{t+l}; \eta'_{r+1}, \dots, \eta'_{r+s})$, as $t, r \to \infty$, where l, s are any finite numbers.

2 Results

Let us consider the first elements of the sequences $\{\eta_t\}_{t=1}^{\infty}$ and $\{\eta'_t\}_{t=1}^{\infty}$.

Lemma 1. *If* $\mathbf{P}_{\alpha}^{\beta} = \mathbf{P}\{\eta_1 = \alpha, \eta'_1 = \beta\}$, $\alpha, \beta \in \{0, 1\}$, *then* $\mathbf{P}_0^0 = \mathbf{P}_1^1 = pq/2$ *and* $\mathbf{P}_1^0 = \mathbf{P}_0^1 = \frac{q+p^2}{2}$.

The joint distribution of $(\eta_1, \eta'_1, \eta_2, \eta'_2)$ is more complicated. However the limit distribution of a pair (η_t, η'_t) as $t \to \infty$ is very simple.

Theorem 1. *The elements* η_t *and* η'_t *of the sequences* $\{\eta_t\}_{t=1}^{\infty}$ *and* $\{\eta'_t\}_{t=1}^{\infty}$ *obtained by the Neumann transform (2) and (3) are asymptotically independent as* $t \to \infty$.

The proof of independency is based on the fact that the distribution of a pair (τ_t, τ'_t) is asymtotically Gaussian. Then the relations

$$\left|\mathbf{P}\{(\eta_t, \eta'_t) = (\alpha, \alpha')\} - \frac{1}{4}\right| \leq \mathbf{P}\{|\tau_t - \tau'_t| < 2\} \to 0, \text{ when } t \to \infty, \text{ where } \alpha, \alpha' \in \{0, 1\},$$

shows that the distribution of a pair (η_t, η'_t) tends to the equiprobable one when $t \to \infty$.

The statement may be also applied to the sets of neighbouring elements of sequences $\{\eta_t\}_{t=1}^{\infty}$ and $\{\eta'_t\}_{t=1}^{\infty}$.

Theorem 2. *The elements* $(\eta_{t+1}, \ldots, \eta_{t+l}; \eta'_{r+1}, \ldots, \eta'_{r+s})$, $l, s \in \mathbb{N}$, *obtained from the sequences* $\{\eta_t\}_{t=1}^{\infty}$ *and* $\{\eta'_t\}_{t=1}^{\infty}$ *by Neumann transforms (2) and (3), are asymptotically independent at* $t, r \to \infty$.

References

[1] von Neumann J. (1951). Various techniques used in connection with random digits. *Applied Math, Washington, DC.* Vol. **12**, pp. 36-38.

[2] von Neumann J. (1963). von Neumann's Collected Works. *Oxford, U.K., Pergamon.* Vol. **5**, pp. 768-770.

[3] Lehmann E.L. (1979). *Testing Statistical Hypothesis.* Science, Moscow.

[4] Zubkov A.M., Menshenin D.O. (2004) Bernoilli sequence transform by Neumann method. *Obozr. Prikl. Prom. Math.*. Vol. **11**, pp. 820, (in Russian).

ON RANDOM GRAPHS IN RANDOM ENVIRONMENT

Yu. L. Pavlov
Institute for Applied Mathematical Research, Karelian Research Centre of RAS
Petrozavodsk, RUSSIA
e-mail: pavlov@krc.karelia.ru

Abstract

We consider a configuration graph with N vertices whose degrees are independent identically distributed according to power-law distribution under the condition that the sum of vertex degrees is equal to n. A random graph dynamics as $N, n \to \infty$ to take place in a random environment when parameter of vertex degree distribution following uniform distribution on the finite fixed interval. The limit distributions of the maximum vertex degree and the number of vertices with a given degree were obtained.

1 Introduction

The study of random graphs has been gaining interest in connection with the wide use of these models for the description of different complex networks (see e. g. [3]). One of the ways for constructed such models based on configuration graphs introduced in [2]. Configuration random graphs are being a good implementation of the social, telecommunication networks and Internet topology. While considering real networks it has been noted that they could be adequate representing by random graphs with the vertex degrees being independent identically distributed random variables following the power-law distribution [4]. In [7] it was shown that the distribution of a random variable ξ, being equal to an arbitrary vertex degree could be defined as follows:

$$\mathbf{P}\{\xi = k\} = k^{-\tau} - (k+1)^{-\tau}, \tag{1}$$

where $k = 1, 2, \ldots; \tau > 0$. Moreover in [4] it was found that for present-day complex telecommunication networks the typical values of the distribution (1) parameter τ belongs to the interval (1, 2). Research in the last years showed that configuration power-law random graphs could be used also for modeling forest fires as well as banking system defaults, but in these cases usually $\tau > 2$ [6]. Let N be a number of vertices in the graph and random variables $\xi_1, \ldots, \xi_N$ are equal to the degrees of vertices with the numbers $1, \ldots, N$. These variables are independent and following the distribution (1). The vertex degree is the number of its semiedges, i. e. edges for which adjacent vertices are not yet determined. All of semiedges are numbered in an arbitrary order. The graph is constructed by joining all of the semiedges pairwise equiprobably to form edges. Those models admit multiple edges and loops. The sum of vertex degrees in any graph has to be even, so if the sum $\xi_1 + \ldots + \xi_N$ is odd we add one extra vertex with degree one. In [7] it was note that addition of this vertex together with its semiedge does not influence the graph behaviour as $N \to \infty$. That is why further we will consider

only vertex degrees $\xi_1, \ldots, \xi_N$. An interesting fact (see e. g. [1]) that parameter τ of the distribution (1) can be depended on N and even can be random.

We consider the subset of random graphs under the condition that sum of vertex degrees is equal to n. It means that $\xi_1 + \ldots + \xi_N = n$ and $\xi_1, \ldots, \xi_N$ are not independent. Such conditional graphs can be useful for modeling of networks for which we can estimate the number of links. They are useful also for studying networks without conditions on the number of edges by averaging the results of conditional graphs with respect to the distribution of the sum of degrees. We assume that as $N \to \infty$ a dynamics of our graph to take place in a random environment when τ is a random variable following uniform distribution on the interval $[a, b], 0 < a < b < \infty$. Then from (1) we find

$$p_1 = \mathbf{P}\{\xi = 1\} = 1 - \frac{1}{(b-a)\ln 2}\left(\frac{1}{2^a} - \frac{1}{2^b}\right),$$

$$p_k = \mathbf{P}\{\xi = k\} = \frac{1}{(b-a)\ln k}\left(\frac{1}{k^a} - \frac{1}{k^b}\right) - \frac{1}{(b-a)\ln(k+1)}\left(\frac{1}{(k+1)^a} - \frac{1}{(k+1)^b}\right),$$

where $k = 2, 3, \ldots$

Denote by $\xi_{(N)}$ and μ_r the maximum vertex degree and the number of vertices with degree r respectively. We obtained the limit distributions of $\xi_{(N)}$ and μ_r as $N, n \to \infty$. The technique of obtaining these results is based on so called generalized allocation scheme supported by V. F. Kolchin [5].

2 Proof Strategy

Let $\eta_1, \ldots, \eta_N$ be auxiliary independent identically distributed random variables such that

$$p_k(\lambda) = \mathbf{P}\{\eta_i = k\} = \lambda^k p_k / B(\lambda), \tag{2}$$

where $k = 1, 2, \ldots;\ i = 1, \ldots, N;\ 0 < \lambda < 1$ and

$$B(\lambda) = \sum_{k=1}^{\infty} \lambda^k p_k.$$

It is readily seen that for our subset of graphs

$$\mathbf{P}\{\xi_1 = k_1, \ldots, \xi_N = k_N\} = \mathbf{P}\{\eta_1 = k_1, \ldots, \eta_N = k_N \mid \eta_1 + \ldots + \eta_N = n\}. \tag{3}$$

This equation means that for random variables $\xi_1, \ldots, \xi_N$ and $\eta_1, \ldots, \eta_N$ the generalized allocation scheme is valid and we can apply the known properties of this scheme to the study of conditional random graphs.

Let $\eta_i^{(r)}, \nu_i^{(r)}, i = 1, \ldots, N$, be two sets of random variables such that

$$\mathbf{P}\{\eta_i^{(r)} = k\} = \mathbf{P}\{\eta_i = k | \eta_i \le r\}, \quad \mathbf{P}\{\nu_i^{(r)} = k\} = \mathbf{P}\{\eta_i = k | \eta_i \ne r\}.$$

It is shown in [5] that from (3) it is not hard to get:

$$\mathbf{P}\{\xi_{(N)} \le r\} = (1 - \mathbf{P}\{\eta_1 > r\})^N \frac{\mathbf{P}\{\eta_1^{(r)} + \ldots + \eta_N^{(r)} = n\}}{\mathbf{P}\{\eta_1 + \ldots + \eta_N = n\}} \tag{4}$$

and

$$\mathbf{P}\{\mu_r = k\} = \binom{N}{k} p_r^k(\lambda)(1 - p_r(\lambda))^{N-k} \frac{\mathbf{P}\{\nu_1^{(r)} + \ldots + \nu_{N-k}^{(r)} = n - kr\}}{\mathbf{P}\{\eta_1 + \ldots + \eta_N = n\}}. \tag{5}$$

From (4) and (5) we see that to obtain the limit distributions of $\xi_{(N)}$ and μ_r it suffices to consider the asymptotic behaviour of the sums of independent random variables, binomial $(1 - \mathbf{P}\{\eta_1 > r\})^N$ and binomial probabilities. By this way we proved the main results of this paper (see the next section).

3 Results

Let parameter $\lambda = \lambda(N, n)$ of the distribution (2) be determined by the relation

$$m = \mathbf{E}\eta_1 = n/N$$

and let also $\sigma^2 = \mathbf{D}\eta_1$. We have the following results.

Theorem 1. *Let $N, n \to \infty$ in such a way that $n/N \to 1, (n - N)^3/N^2 \to \infty$ and sequence $r = r(N, n)$ are minimal natural numbers such that $N\lambda^r p_{r+1}/p_1 \to \gamma$, where γ is a non-negative constant. Then $\mathbf{P}\{\xi_{(N)} = r\} \to e^{-\gamma}$, $\mathbf{P}\{\xi_{(N)} = r + 1\} \to 1 - e^{-\gamma}$.*

Theorem 2. *Let $N, n \to \infty$ in such a way that $1 < C_1 \le n/N \le C_2 < \infty$ and $r = r(N, n)$ are chosen such that*

$$\frac{aN\lambda^{r+1}}{(b-a)B(\lambda)r^{a+1}\ln r} \to \gamma,$$

where γ is a positive constant. Then for any fixed $k = 0, \pm 1, \pm 2, \ldots$

$$\mathbf{P}\{\xi_{(N)} \le r + k\} = \exp\{-\gamma\lambda^k(1 - \lambda)^{-1}\}(1 + o(1)).$$

Theorem 3. *Let $N, n \to \infty$ in such a way that $n/N \to \infty, a \le 1$ and $N(1-\lambda)^{2+\delta} \to \infty$ for some $\delta > 0$. Then*

$$\mathbf{P}\{|\ln\lambda|\xi_{(N)} - u \le z\} \to e^{-e^{-z}},$$

where $-\infty < z < \infty$ and $u = u(N, n)$ are chosen so that

$$\frac{N|\ln\lambda|^a}{e^u u^{a+1}\ln(u/|\ln\lambda|)} \to \frac{b-a}{a}.$$

Theorem 4. *Let $N, n \to \infty$ in such a way that $n/N \to 1, n-N \to \infty$. Then for $r > 2$*

$$\mathbf{P}\{\mu_r = k\} = \frac{(Np_r(\lambda))^k}{k!} e^{-Np_r(\lambda)}(1 + o(1))$$

uniformly in the integer k such that $(k - Np_r(\lambda))/\sqrt{Np_r(\lambda)}$ lies in any fixed finite interval.

Theorem 5. *Let $N, n \to \infty$ and one of the following conditions hold:*

$$1. \quad 1 < C_1 \le n/N \le C_2 < \infty;$$

$$2. \quad a \le 1, \quad n/N \to \infty, \quad N(1-\lambda)^{2+\delta} \to \infty,$$

where δ is a some positive constant. Then for any fixed natural r

$$\mathbf{P}\{\mu_r = k\} = (\sigma_{rr}\sqrt{2\pi N})^{-1} e^{-u_r^2/2}(1 + o(1))$$

uniformly in the integer k such that $u_r = (k - Np_r(\lambda))/(\sigma_{rr}\sqrt{N})$ lies in any fixed finite interval, where

$$\sigma_{rr}^2 = p_r(\lambda)\left(1 - p_r(\lambda) - \frac{(m-r)^2}{\sigma^2} p_r(\lambda)\right).$$

4 Acknowledgements

The study is supported by the Russian Foundation for Basic Research, grant 16-01-00005.

References

[1] Bianconi G., Barabasi A.-L. (2001). Bose-Einstein condensation in complex networks. *Physical Review Letters*. Vol. **86**, pp. 5632-5635.

[2] Bollobas B. (1980). A probabilistic proof of an asymptotic formulae for the number of labelled regular graphs. *Eur. J. Comb.*. Vol. **1**, pp. 311-316.

[3] Durrett R. (2006). *Random Graph Dynamics*. Cambridge University Press, Cambridge.

[4] Faloutsos C., Faloutsos P., Faloutsos M. (1999). On power-law relationship of the Internet topology. *Computer Communications Rev.*. Vol. **29**, pp. 251-262.

[5] Kolchin V. F. (1999). *Random Graphs*. Cambridge University Press, Cambridge, New York.

[6] Leri M., Pavlov Yu. (2014). Power-law random graphs' robustness: link saving and forest fire model.*Austrian Journal of Statistics*. Vol. **43**, pp. 229-236.

[7] Reittu H., Norros I. (2004). On the power-law random graph model of massive data networks. *Performance Evaluation*. Vol. **55**, pp. 3-23.

HOEFFDING TYPE INEQUALITIES FOR LIKELIHOOD RATIO TEST STATISTIC

M. Radavičius
Institute of Mathematics and Informatics, Vilnius University
Vilnius, LITHUANIA
e-mail: marijus.radavicius@vu.mii.lt

Abstract

For Bernoulli trials, simple upper and lower bounds for tail probabilities of logarithmic likelihood ratio statistic are given. The bounds are exact up to a factor of 2. A problem of generalization of the results to the multinomial distribution is briefly discussed.

1 Introduction

Let $\mathbf{y} = (y_1, \dots, y_n)$ be a random vector having the multinomial distribution

$$\mathbf{y} \sim \text{Multinomial}_n(N, \mathbf{p}), \quad \mathbf{p} = (p_1, \dots, p_n).$$

For $n = 2$, $y_1 \sim \text{Binomial}(N, p_1)$. The maximum likelihood estimator of the unknown probabilities $\mathbf{p}$ is given by

$$\hat{\mathbf{p}} = \hat{\mathbf{p}}_N := N^{-1}\mathbf{y}.$$

Define scaled (logarithmic) likelihood ratio statistic

$$\ell_n(\hat{\mathbf{p}}, \mathbf{p}) := \sum_{i=1}^{n} \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i}\right).$$

Note that for $n = 2$,

$$\ell(\hat{p}_1, p_1) := \hat{p}_1 \log\left(\frac{\hat{p}_1}{p_1}\right) + (1 - \hat{p}_1) \log\left(\frac{1 - \hat{p}_1}{1 - p_1}\right) = \ell_2(\hat{\mathbf{p}}, \mathbf{p}). \tag{1}$$

Hoeffding (1965) proved the following inequality (see also Kallenberg, 1985): for $n = 2$,

$$\mathbf{P}\{\ell_n(\hat{\mathbf{p}}_N, \mathbf{p}) \geq x\} \leq 2\mathrm{e}^{-Nx}, \quad x > 0. \tag{2}$$

It is important to stress that (2) is *universal*: it holds for all $x > 0$, all $p_1 \in [0, 1]$ and all $N = 1, 2, \dots$. It is also *tight*, i.e. it cannot be improved without imposing some additional conditions.

The *problem* is to generalize this inequality to the case $n > 2$.

Generalizations of (2) to the case $n > 2$ have been obtained by Hoeffding himself (1965) and W.C.M. Kallenberg (1985). The inequality established by W.C.M. Kallenberg is tight up to a constant. However it holds only for $x \leq 0.15$ and impose some boundedness from below restriction on probabilities $\mathbf{p}$. Known universal bounds (i.e. bounds that are independent of $\mathbf{p}$) are loose and impractical for large n (W.C.M. Kallenberg, 1985, inequality (2.6)). The upper bound typically exceeds the corresponding lower bound by a factor of order $\sqrt{xN}$ (see W.C.M. Kallenberg, 1985, Theorem 2.1 on p. 1557). This applies to 2) as well.

2 Notation

Define the signed logarithmic likelihood ratio statistic for the binomial distribution

$$s_q(u) := \operatorname{sign}(u-q)\,\ell(u,q), \quad (u,q) \in (0,1)\times(0,1).$$

The logarithmic likelihood ratio statistic for the binomial distribution is defined in (1). The function $s_q(u)$ is strictly increasing and continuous with respect to $u \in (0,1)$. Let $\bar{s}_q$ denote the inverse function of s_q: $\bar{s}_q(s_q(u)) \equiv u$. In what follows, we reserve notation χ^2_m for a random variable which has χ^2 distribution with m degrees of freedom.

Let

$$b(t) = b(t;N,q) := \frac{\Gamma(N+1)}{\Gamma(t+1)\Gamma(N-t+1)}\, q^t(1-q)^{N-t}, \quad t \in [0,N].$$

Note that

$$b(k;N,q) = C_N^k q^k (1-q)^{N-k}, \quad k = 0,1,\ldots,N,$$

is the binomial probability density (mass) function.

3 Results

The proposition below gives upper and lower bounds (exact up to a factor of 2) for the tail probabilities of the logarithmic likelihood ratio statistic. In contrast to (2), they depend on the success probability of the binomial distribution.

Proposition 1. *Let* $\hat{p}_N := N^{-1}y$, $y \sim \text{Binomial}(N,p)$. *Then*

$$\begin{aligned} \mathbf{P}\{\ell(\hat{p}_N,p) \ge x\} &\le \mathbf{P}\{\chi^2_1 \ge 2\,xN\} \\ &+ b(N\bar{s}_p(-x);N,p) + b(N\bar{s}_p(x);N,p) \\ &\le 2\,\mathbf{P}\{\ell(\hat{p}_N,p) \ge x\}. \end{aligned} \tag{3}$$

The inequalities for the upper tail probability presented below are more apprehensible than (3). Let $x_k := s_p(k/N)$ with $k/N > p$. Then

$$\begin{aligned} \max(2^{-1}\mathbf{P}\{\chi^2_1 \ge 2\,x_kN\}, b(k;N,p)) &\le \mathbf{P}\{s_p(\hat{p}_N) \ge x_k\} \\ &\le 2^{-1}\mathbf{P}\{\chi^2_1 \ge 2\,x_kN\} + b(k;N,p). \end{aligned} \tag{4}$$

Note that the first term in the right-hand side of (4) is just the tail probability of the asymptotic distribution of the logarithmic likelihood ratio statistic.

The inequalities (4) as well as the Proposition 1 are simple corollaries of results by Zubkov and Serov [4].

Remark. We expect that, for arbitrary $n > 2$, exact (up to a constant factor) upper bounds for tail probabilities of logarithmic likelihood ratio statistic can be obtained by making use of the Proposition 1 and induction with respect to n.

References

[1] Hoeffding W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*. Vol. **58**, pp. 13-30.

[2] Hoeffding W. (1965). Asymptotically Optimal Tests for Multinomial Distributions. *Ann. Math. Statist.* Vol. **36**, pp. 369-401.

[3] Kallenberg W.C.M. (1985). On moderate and large deviations in multinomial distributions. *Annals of Statistics*. Vol. **13**, pp. 1554-1580.

[4] Zubkov A.M., Serov A.A. (2012). A complete proof of universal inequalities for distribution function of binomial law. *Teoriya Veroyatnostei i ee Primeneniya*. Vol. **57**, pp. 597-602.

STEGANOGRAPHIC CAPACITY OF LOCALLY UNIFORM MARKOV COVERS

Valeriy Voloshko
Research Institute for Applied Problems of Mathematics and Informatics
Minsk, BELARUS
e-mail: ValeraVoloshko@yandex.ru

Abstract

Proposed a new correction algorithm for the standard steganographic model of binary message embedding into binary cover. Embedding is known to be influential on cover's statistical characteristics and, thus, to be statistically detectable. The proposed correction algorithm does not affects the embedded message and under certain model assumptions restores the cover's histogram of n-subwords frequencies. The key condition for the n-subwords histogram restorability limits the ratio of message to cover: it should not exceed some value called n-capacity of cover. Some capacities found theoretically for locally uniform Markov covers.
Keywords: steganographic capacity, histogram correction, embedding

1 Introduction

Such tools for multimedia copyright protection, as digital watermarking or digital signature, use steganographic methods of covert embedding. The standard embedding model is very simple [1, 2]. We have three binary sequences of $\{0,1\}$ values: *cover* $\mathbf{c} = (\mathbf{c}_i)_{i=1}^N$, *selector* $\mathbf{s} = (\mathbf{s}_i)_{i=1}^N$ and *message* $\mathbf{m} = (\mathbf{m}_i)_{i=1}^{N_1}$, where N_1 is the number of ones in selector $\mathbf{s}$. Cover values $\mathbf{c}_i$ corresponding to N_1-subset of indices $\{i : \mathbf{s}_i = 1\}$ are then being replaced with N_1 message values. Cover sequence $\mathbf{c}$ with embedded message $\mathbf{m}$ we call *stego* and denote $\mathbf{c}^* = (\mathbf{c}_i^*)_{i=1}^N$. After message embedding we may want to correct stego $\mathbf{c}^*$ for some goal. Of course, correction should not affect the embedded message values $\{\mathbf{c}_i^* : \mathbf{s}_i = 1\}$. Corrected stego sequence we denote $\mathbf{c}^{**} = (\mathbf{c}_i^{**})_{i=1}^N$.

The goal of correction is usually to somehow restore certain features of cover $\mathbf{c}$, distorted by message embedding. Obviously, we are not talking about restoration of the cover $\mathbf{c}$ itself, because stego $\mathbf{c}^*$ can not even be moved closer to it in Hamming metric d by correction:

$$d(\mathbf{c}, \mathbf{c}^*) \le d(\mathbf{c}, \mathbf{c}^{**}).$$

Nevertheless, the cover's statistical characteristics turn quite repairable. Here we aim to approximately restore the histogram of frequencies of cover's n-subwords:

$$\|\mathbf{H}_n(\mathbf{c}) - \mathbf{H}_n(\mathbf{c}^{**})\| \to \min, \tag{1}$$

where

$$\mathbf{H}_n(x) ::= \left(\frac{\#\{0 \le i \le N-n : (x_{i+1}, \dots, x_{i+n}) = q\}}{N+1-n} \right)_{q \in \{0,1\}^n}, \ x \in \{0,1\}^N. \tag{2}$$

In its formulation the problem (1) looks rather combinatorial, but it appears to be effectively treatable by probabilistic and statistical methods based on few model assumptions. The following probabilistic model assumptions are used to be standard in literature [1, 2]:

A1: cover $\mathbf{c}$, selector $\mathbf{s}$ and message $\mathbf{m}$ are mutually independent random binary sequences;

A2: selector $\mathbf{s}$ is a Bernoulli process with success probability $0 < \varepsilon < 1$;

A3: message $\mathbf{m}$ is a uniformly distributed sequence (Bernoulli process with parity of successes and failures);

A4: cover $\mathbf{c}$ is a Markov chain.

Based on [1], we use extended versions of **A2** and **A4**:

XA2: selector $\mathbf{s}$ is a stationary n-ergodic process;

XA4: cover $\mathbf{c}$ is a stationary n-ergodic process.

Remark. Compared to **XA4**, extension **XA2** is more exotic and analytically harder to work with, but it may sufficiently increase the capacity of stegosystem [1].

Remark. Under ergodicity we mean almost sure convergence of frequencies to probabilities. Namely, let $x : \mathbb{Z} \to \{0, 1\}$ be a stationary random binary process. Then we call it n-ergodic, if in (2):

$$\mathbf{H}_n(x) \xrightarrow[N\to+\infty]{\text{a.s.}} [x]^n = \left([x]^n_q\right)_{q\in\{0,1\}^n}, \ [x]^n_q ::= \mathbb{P}\{(x_1, \ldots, x_n) = q\}. \tag{3}$$

Remark. We call $[x]^n$ in (3) an *n-projection* of x's probability measure $[x] ::= [x]^\infty$.

Thus the distance (1) between histograms almost surely vanishes at $N \to +\infty$, if the following two conditions hold for cover $\mathbf{c}$ and corrected stego $\mathbf{c}^{**}$:

- they are both n-ergodic;
- they have the same n-projections of probability measures: $[\mathbf{c}]^n = [\mathbf{c}^{**}]^n$.

The correction algorithm proposed in [1] provide these conditions.

Remark. Under n-ergodicity the restoration of n-projection $[\mathbf{c}]^n$ of cover's measure guarantees the asymptotic restoration of cover's n-subwords histogram $\mathbf{H}_n(\mathbf{c})$. So further under histogram restoration we understand the asymptotic one.

2 Correction algorithm

Thus the embedding leads to deformation of the cover's probability measure $[\mathbf{c}]$, and we want to restore it (at least up to probabilities of n-subwords) by correction. It is shown in [1] that under assumptions **A1**, **XA2**, **A3** and **XA4** the considered deformation of $[\mathbf{c}]$ turns out to be a convolution with the specially transformed selector's measure. Namely, with the measure $[\mathbf{su}]$ of selector $\mathbf{s}$ multiplied by independent from it uniformly distributed random binary sequence $\mathbf{u}$ (in the standard **A2** case $[\mathbf{su}]$ is Bernoulli measure with success probability $\varepsilon/2$). So the idea of correction algorithm is

clear now: we just have to replace cover $\mathbf{c}$ with another stationary n-ergodic random binary sequence $\mathbf{k}$ (let us call it *corrector*), whose measure (n-projection, to be precise) $[\mathbf{k}]^n$ convoluted with $[\mathbf{su}]^n$ gives n-projection $[\mathbf{c}]^n$ of the cover's measure.

From the computational point of view, we need an inverse convolution. To obtain it, one may use Fourier transform $\mathbf{F}$, which provides correspondence between convolution and multiplication. Without going into technical details, the object of our interest, n-projection $[\mathbf{k}]^n$ of the corrector's measure, has the form [1]:

$$[\mathbf{k}]^n = 2^{-n}\mathbf{F}\left(\frac{\mathbf{F}[\mathbf{c}]^n}{\mathbf{F}[\mathbf{su}]^n}\right),\ (\mathbf{F}f)_q = \sum_{q'\in\{0,1\}^n} f_{q'}(-1)^{|qq'|},\ q\in\{0,1\}^n, \tag{4}$$

reducing in the standard **A2** case to:

$$[\mathbf{k}]^n_q = \left(\frac{1-\varepsilon/2}{1-\varepsilon}\right)^n \sum_{q'\in\{0,1\}^n} [\mathbf{c}]^n_{q\oplus q'}\left(\frac{\varepsilon}{\varepsilon-2}\right)^{|q'|},\ q\in\{0,1\}^n, \tag{5}$$

where $|q'|$ means Hamming weight of q' and $\oplus$ means elementwise XOR. The uniqueness of $[\mathbf{k}]^n$ is guaranteed by strict positiveness of $[\mathbf{s}]^n$ (every binary n-word $q\in\{0,1\}^n$ has nonzero probability to appear as a subword in selector $\mathbf{s}$, holds for **A2**). The existence of n-ergodic corrector itself is guaranteed by strict positiveness of $[\mathbf{k}]^n$ in (4) (or (5) for **A2**). The last thing we should say about the algorithm is that we have to use histogram $\mathbf{H}_n(\mathbf{c})$ instead of n-projection $[\mathbf{c}]^n$, which is unknown on practice.

Thus the correction algorithm is of the form:

- **step 1:** compute cover's histogram of n-subwords frequencies $\mathbf{H}_n(\mathbf{c})$;
- **step 2:** using $\mathbf{H}_n(\mathbf{c})$ instead of $[\mathbf{c}]^n$ in (4) (or (5) for **A2**), compute n-projection $[\mathbf{k}]^n$ of the corrector's measure;
- **step 3:** generate the corrector $\mathbf{k}$ by pseudorandom stationary Markov chain of order $n-1$ with transfer probabilities, providing computed $[\mathbf{k}]^n$, or state the fail of correction, if $[\mathbf{k}]^n$ is not strictly positive;
- **step 4:** correct the values of stego $\mathbf{c}^*$: replace $\mathbf{c}^*_i$ with $\mathbf{k}_i$ at the positions i, not occupied by message ($\mathbf{s}_i = 0$).

3 Capacity of cover

The assumption **XA2** means, in particular, that the portion of ones N_1/N in selector $\mathbf{s}$ almost surely tends to $[\mathbf{s}]_1 = \mathbb{P}\{\mathbf{s}_i = 1\}$ when the cover's volume N grows. For this reason under **XA2** the value $[\mathbf{s}]_1$ can be thought of as a data transfer rate (DTR for brevity) of stegosystem. Maximization of DTR seems rather natural objective, next after cover's histogram restoration. Hence the idea of capacity [1, 3] as a steganographic characteristic of cover. Informally, capacity is a maximum achievable DTR among stegosystems providing histogram restorability for some particular cover. More precisely, in considered model capacity characterizes cover's distribution $[\mathbf{c}]$.

Following [1], consider two cases. If selector $\mathbf{s}$ is an arbitrary one (**XA2** case), providing restoration of cover's n-subwords histogram, then maximum DTR is called

absolute n-capacity of cover's measure $[\mathbf{c}]$ and denoted by $\varepsilon_n^*[\mathbf{c}]$. And it is called *plain n-capacity* and denoted by $\varepsilon_n[\mathbf{c}]$, if selector is chosen among Bernoulli processes (**A2** case). Obviously, both absolute and plain capacities of a fixed cover's probability measure $[\mathbf{c}]$ do not increase in n and $\varepsilon_n[\mathbf{c}] \le \varepsilon_n^*[\mathbf{c}]$.

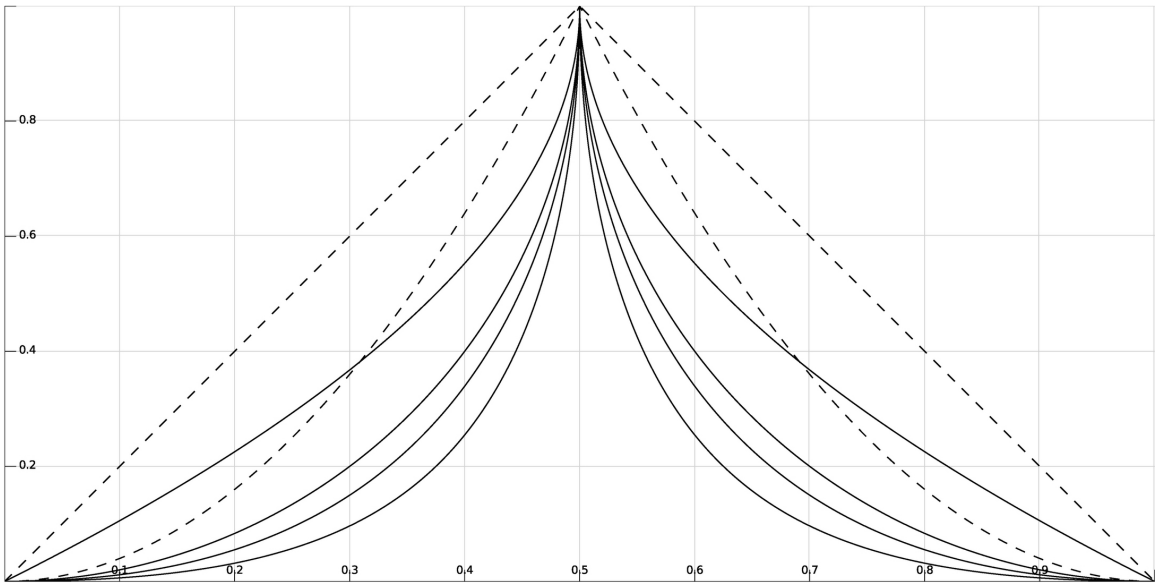


Figure 1: Capacities of $\mathrm{MC}_{\mathrm{U}}^1(p)$ against p: plain $\varepsilon_2 > \varepsilon_3 > \varepsilon_4 > \varepsilon_\infty$ (solid lines) and absolute $\varepsilon_2^* > \varepsilon_3^*$ (dashed lines).

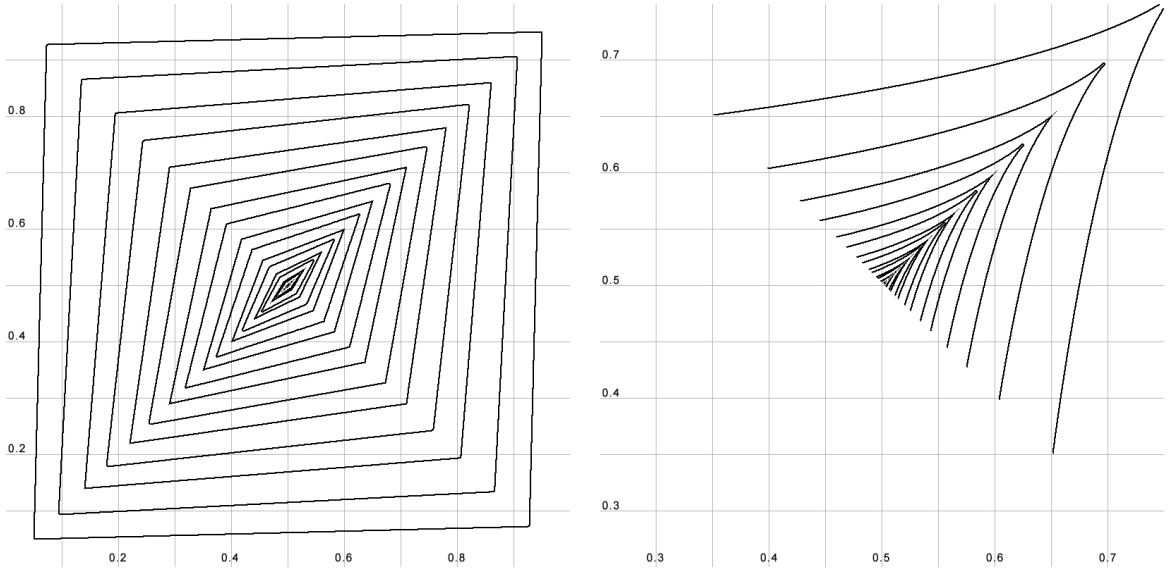


Figure 2: Contour maps for plain capacities of $\mathrm{MC}_{\mathrm{U}}^2(p,s)$ on the (p,s) plane: ε_3 on the left and ε_∞ (only for $p+s>1$) on the right. Lines correspond to multiples of 0.05.

Consider now two Markov models of locally uniform covers: the first order Markov chain with uniformly distributed 1-subwords and the second order one with uniformly

distributed 2-subwords. We call them $\mathrm{MC}^1_{\mathrm{U}}$ and $\mathrm{MC}^2_{\mathrm{U}}$ respectively. The first one appears one-parametric with parameter:

$$p = \mathbb{P}\{0 \to 1\} = \mathbb{P}\{1 \to 0\}.$$

The second one is two-parametric with parameters:

$$p = \mathbb{P}\{00 \to 1\} = \mathbb{P}\{10 \to 0\},$$
$$s = \mathbb{P}\{11 \to 0\} = \mathbb{P}\{01 \to 1\}.$$

The arrows mean transfers within the cover's sequence $\mathbf{c}_i$.

Theorem 1. *[1] Absolute and plain capacities of* $\mathrm{MC}^1_{\mathrm{U}}(p)$*-distributed cover are:*

$$\varepsilon^*_2 = 2\hbar,\ \varepsilon^*_3 = 4\hbar^2,$$
$$\varepsilon_2 = 1 - \sqrt{1-2\hbar},\ \varepsilon_3 = 1 - \sqrt{1-4\hbar^2},\ \varepsilon_4 = 1 - \sqrt{\kappa + \kappa^2 - \kappa^3},\ \varepsilon_\infty = 1 - \frac{\sqrt{1-2\hbar}}{1-\hbar},$$

where $\hbar = \min\{p, 1-p\}$, $\kappa = \hbar + \sqrt{1+\hbar^2}$.

Theorem 2. *Third and limiting (for* $p+s>1$*) plain capacities of* $\mathrm{MC}^2_{\mathrm{U}}(p,s)$*-distributed cover are:*

$$\varepsilon_3 = \begin{cases} 1 - \sum_{\pm} \sqrt[3]{\hbar_- \pm \sqrt{\hbar_-^2 - \hbar_+^3}}, & \hbar_-^2 \geq \hbar_+^3, \\ 1 - 2\sqrt{\hbar_+} T_{\frac{1}{3}}(\hbar_- \hbar_+^{-\frac{3}{2}}), & \hbar_-^2 < \hbar_+^3, \end{cases},\ \varepsilon_\infty = 1 - \frac{\sinh^3\Phi - \tanh^3\Psi}{\sinh^2\Phi - \tanh^2\Psi} \cdot \frac{1}{\cosh\Phi},$$

where $\hbar_+ = \frac{1}{3}|1-p-s|$, $\hbar_- = \frac{1}{2}|p-s|$, $T_\nu(x) = \cos(\nu \arccos x)$ *is a fractional analogue of Chebyshev polynomial,*

$$\Psi = \operatorname{arcsinh} \sqrt[3]{\frac{|p-s|}{p+s-2ps}},\ \Phi = \operatorname{arccosh}\left(\frac{p+s-2ps}{(1-p)(1-s)} \cdot \frac{7+\cosh(4\Psi)}{16\cosh\Psi}\right).$$

Remark. Both limiting plain capacities ε_∞ for the considered cover models were obtained based on some unproven hypotheses, confirmed by numerical experiments.

Remark. Comparison of absolute and plain capacities (Figure 1) shows that more sophisticated choice of positions for embedding may sufficiently increase the data transfer rate of stegosystem.

References

[1] Voloshko V.A. (2016). Steganographic Capacity for One-dimensional Markov Cover. *Diskretnaya Matematika*. Vol. **28**(1), pp. 19–43 (in Russian).

[2] Kharin Yu.S., Vecherko E.V. (2016). Detection of Embeddings in Binary Markov Chains. *Discrete Mathematics and Applications*. Vol. **26**(1), pp. 13–29.

[3] Harmsen J.J., Pearlman W.A. (2009). Capacity of Steganographic Channels. *IEEE Transactions on Information Theory*. Vol. **55**, pp. 1775–1792.

ON COINCIDENCES OF TUPLES IN A BINARY TREE WITH RANDOMLY LABELED VERTICES

A. M. Zubkov[1], V. I. Kruglov[2]
Steklov Mathematical Institute, Russian Academy of Sciences
Moscow, RUSSIA
e-mail: [1]zubkov@mi.ras.ru, [2]kruglov@mi.ras.ru

Abstract

Let all vertices of a complete binary tree of finite height be independently and equiprobably labeled by the elements of some finite alphabet. We consider the numbers of pairs of identical tuples of labels on chains of subsequent vertices in the tree. Exact formulae for the expectations of these numbers are obtained. Convergence to the compound Poisson distribution is proved.

The work was supported by the Russian Science Foundation under grant 14-50-00005.

Let T_2^n be a complete binary tree of height n with root $*$ and n layers of vertices; we enumerate 2^k elements of the set $I^{(k)}$ of the k-th layer vertices ($k = 1, 2, \ldots, n$) by binary strings $i = (i_1, i_2, \ldots, i_k) \in \{0,1\}^k$. So the unique vertex $*$ of layer $I^{(0)}$ is connected by two outcoming edges with vertices of layer $I^{(1)}$ and any vertex $i = (i_1, i_2, \ldots, i_k) \in I^{(k)}, 1 \le k \le n-1$, is connected by two outcoming edges with vertices $i' = (i_1, i_2, \ldots, i_k, 0)$ and $i'' = (i_1, i_2, \ldots, i_k, 1)$ of layer $I^{(k+1)}$. Vertex $i = (i_1, i_2, \ldots, i_k)$ has incoming edge from vertice $i^- = (i_1, i_2, \ldots, i_{k-1})$ for $k > 1$ and from root $* = (0)^- = (1)^-$ for $k = 1$. Each vertex i of the tree T_2^n defines subtree consisting of this vertex and all vertices of next layers that are connected to i with edges.

We can define natural lexicographical order on the set of vertices of T_2^n : $i = (i_1, \ldots, i_k) \prec j = (j_1, \ldots, j_h)$ if either $i = *, j \ne *$, or $1 \le k < h$, or $1 \le k = h$ and $\sum_{m=1}^k i_m 2^{k-m} < \sum_{m=1}^k j_m 2^{k-m}$. For vertex $i = (i_1, i_2, \ldots, i_k) \in I^{(k)}, k \ge 0$, the chain C_i of length l is a sequence of l vertices

$$(i_1, i_2, \ldots, i_k), (i_1, i_2, \ldots, i_k, i_{k+1}), \ldots, (i_1, i_2, \ldots, i_k, i_{k+1}, \ldots, i_{k+l-1})$$

connected by edges. Denote these vertices of the chain C_i by $C_i[0], C_i[1], \ldots, C_i[l-1]$. We will refer to the vertex i as the initial vertex of chain C_i and to the vertex $(i_1, i_2, \ldots, i_k, i_{k+1}, \ldots, i_{k+l-1})$ as its final vertex. It's easy to see that the final vertex and length l explicitly define the chain, so we can introduce order on the set of chains of the fixed length l: $C_i \prec C_j$ if and only if $C_i[l-1] \prec C_j[l-1]$. Denote by $\mathcal{P}$ the set of ordered pairs of nonintersecting chains $(C_i, C_j), i \prec j$.

It is easy to check that chains C_i and C_j intersect if and only if either $C_i[0] \in C_j$ or $C_j[0] \in C_i$. The total number of vertices in the tree T_2^n is equal to $1 + 2 + \ldots + 2^n = 2^{n+1} - 1$, and the total number of chains of the length l in the tree T_2^n is equal to the number of their final vertices $|\cup_{j=l-1}^n I^{(j)}| = 2^{l-1} + \ldots + 2^n = 2^{n+1} - 2^{l-1}$.

Let any vertex i in tree T_2^n be assigned with a random label $m(i)$ from the set $\{1, \ldots, d\}$ so that variables $m(i)$, $i \in T_2^n$, are independent and $\mathbf{P}\{m(i) = j\} = \frac{1}{d}, j \in$

$\{1,\dots,d\}$, for all $i \in T_2^n$. So, for any chain C_i of length l we have a random tuple of labels

$$M(C_i) = (m(C_i[0]), m(C_i[1]), \dots, m(C_i[l-1])).$$

Obviously, if all chains $C_{i_1}, \dots, C_{i_s}$ are nonintersecting, then the corresponding tuples of labels $M(C_{i_1}), \dots, M(C_{i_s})$ are independent and equiprobably distributed on the set $\{1,\dots,d\}^l$.

We consider the distribution of the number of pairs $(C_i, C_j), i \prec j$, of chains of length l in the tree T_2^n with identical tuples of labels (i.e. $M(C_i) = M(C_j)$). Total number of such pairs is equal to

$$V_{n,l} = \sum_{(C_i,C_j)\in\mathcal{P}} \mathbf{I}\{M(C_i) = M(C_j)\};$$

the alphabet size d is supposed to be fixed.

Probability of the event $\{M(C_i) = M(C_j)\}$ depends on the character of intersection of chains C_i and C_j, so we divide the sum $V_{n,l}$ into several parts: sum $V_{n,l}^{(0)}$ over the nonintersecting chains, sum $V'_{n,l}$ over intersecting chains with different initial vertices, sum $V''_{n,l,k}$ over chains with common initial vertices:

$$V_{n,l} = V_{n,l}^{(0)} + V'_{n,l} + \sum_{k=1}^{l-1} V''_{n,l,k},$$

$$V_{n,l}^{(0)} = \sum_{(C_i,C_j)\in\mathcal{P}:\, C_i\cap C_j=\emptyset} \mathbf{I}\{M(C_i) = M(C_j)\},$$

$$V'_{n,l} = \sum_{(C_i,C_j)\in\mathcal{P}:\, C_i\cap C_j\neq\emptyset, C_i[0]\neq C_j[0]} \mathbf{I}\{M(C_i) = M(C_j)\},$$

$$V''_{n,l,k} = \sum_{(C_i,C'_i)\in\mathcal{P}:\, |C_i\cap C'_i|=k, C_i[0]=C'_i[0]} \mathbf{I}\{M(C_i) = M(C'_i)\},\ 1 \le k < l.$$

Theorem 1. *The following equalities are valid*

$$\mathbf{E}V_{n,l}^{(0)} = \begin{cases} \frac{1}{d^l}\left(2^{2n+1} - 5\cdot 2^{n-1+l} + 2^{n+1} + 2^{2l-2}l\right), & \text{if } 2l-1 \le n, \\ \frac{1}{d^l}\left(2^{2n+1} - 5\cdot 2^{n-1+l} + 2^{2l-2}(n-l+4)\right), & \text{if } 2l-1 > n, \end{cases}$$

$$\mathbf{E}V'_{n,l} = \begin{cases} \frac{1}{d^l}\left((2^{l-1}-1)2^{n+1} - 2^{2l-2}(l-1)\right), & \text{if } 2l-1 \le n, \\ \frac{1}{d^l}\left(2^{l-1}(2^{n+1}-2^l) - 2^{2l-2}(n-l+1)\right), & \text{if } 2l-1 > n, \end{cases}$$

$$\mathbf{E}V''_{n,l,k} = \tfrac{1}{d^{l-k}}\,(2^{n-l+2}-1)2^{2l-k-3},\ 1 \le k < l,$$

$$\sum\nolimits_{k=1}^{l-1} \mathbf{E}V''_{n,l,k} = \frac{2^n - 2^{l-2}}{d}\,\frac{1-(\frac{2}{d})^{l-1}}{1-\frac{2}{d}}\,.$$

If $M(C_i) = M(C_j)$ and $i^- \neq j^-$, then $M(C_{i^-}) = M(C_{j^-})$ with probability $1/d = \mathbf{P}\{m(i^-) = m(j^-)\}$, and $\mathbf{P}\{M(C'_i) = M(C'_j)\} = 1/d$ if $C'_i[0] = C'_j[0], C'_i[l-2] = C'_j[l-2], C'_i[l-1] \neq C'_j[l-1]$. In theorem 2 we propose sufficient conditions and estimate the weak convergence rate of the number of pairs of nonintersecting chains C_i, C_j with $M(C_i) = M(C_j), m(i^-) \neq m(j^-)$ to the compound Poisson distribution. Such pairs of tuples may be interpreted as coincidences which cannot be shifted to the root.

Let

$$X_{C_iC_j} = \mathbf{I}\{M(C_i) = M(C_j), m(i^-) \neq m(j^-)\}, (C_i, C_j) \in \mathcal{P};$$

if $i = *$, then the condition $m(i^-) \neq m(j^-)$ is supposed to be satisfied. Labels of vertices are independent and equiprobable, so for $(C_i, C_j) \in \mathcal{P}$ we have

$$\mathbf{E}X_{C_iC_j} = \mathbf{E}\,\mathbf{I}\{M(C_i) = M(C_j)\}\mathbf{I}\{m(i^-) \neq m(j^-)\} = \begin{cases} \frac{d-1}{d^{l+1}}, & \text{if } i^- \neq j^-, \\ 0, & \text{if } i^- = j^-. \end{cases}$$

Let $\widetilde{\mathcal{P}} \subset \mathcal{P}$ be the set of pairs $(C_i, C_j), i \in I^{(v_i)}, j \in I^{(v_j)}$, of nonintersecting chains such that if the vertex j belongs to a subtree with root i, then $v_j \geq v_i + 2l - 1$. Define

$$V_{n,l}^{(0)-} = \sum_{(C_i,C_j)\in\mathcal{P}:\, C_i\cap C_j=\emptyset} X_{C_iC_j}, \quad \widetilde{V}_{n,l} = \sum_{(C_i,C_j)\in\widetilde{\mathcal{P}}} X_{C_iC_j}.$$

Lemma 1. *The following equalities are valid*

$$\mathbf{E}V_{n,l}^{(0)-} = \begin{cases} \frac{d-1}{d^{l+1}}\left(2^{2n+1} - 6\cdot 2^{n+l-1} + 2^{n+1} + 2^{2l-2}(l+1)\right), & \textit{if } 2l-1 \leq n, \\ \frac{d-1}{d^{l+1}}\left(2^{2n+1} - 6\cdot 2^{n+l-1} + 2^{2l-2}(n-l+5)\right), & \textit{if } 2l-1 > n. \end{cases}$$

$$\mathbf{E}V_{n,l}^{(0)-} - \tfrac{l}{d^l}\,2^{n-l+2} < \mathbf{E}\widetilde{V}_{n,l} < \mathbf{E}V_{n,l}^{(0)-}.$$

Corollary 1. *If $n, l \to \infty$ in such a way that $\mathbf{E}V_{n,l}^{(0)-}$ is bounded, then $\mathbf{P}\{\widetilde{V}_{n,l} = V_{n,l}^{(0)-}\} \to 1$.*

Comparing formulae for $\mathbf{E}V_{n,l}^{(0)}$ and $\mathbf{E}V_{n,l}^{(0)-}$ we can mention that under the conditions of corollary 1 for any coincidence which cannot be shifted to the root there exist in average $\frac{1}{d-1}$ additional coincidences that may be shifted to root.

Definition 1. Consider a pair of chains $(C_i, C_j) \in \mathcal{P}$ such that subtrees of height $l-1$ with roots in vertices i an j do not intersect. Define

$$\pi_k = \frac{1}{k}\,\mathbf{P}\left\{\sum\nolimits_{(C'_i,C'_j)\in\mathcal{P}} X_{C'_iC'_j} = k \;\middle|\; X_{C_iC_j} = 1\right\}, \quad k = 1, 2, \ldots$$

Definition 2. The compound Poisson distribution $CP(\pi)$ is the distribution of random variable

$$\Xi_\pi = \sum\nolimits_{k=1}^{\infty} k\xi_k,$$

where $\xi_1, \xi_2, \ldots$ are independent and for any $k \geq 1$ random variable ξ_k has Poisson distribution with parameter π_k.

Theorem 2. *If $n, l \to \infty$ in such a way that $2^{2l} = o(2^n)$ and*

$$\mathbf{E}\widetilde{V}_{n,l} = \frac{d-1}{d} \cdot \frac{2^{2n+1}}{d^l}(1 + o(1)) \to \lambda \in (0, \infty),$$

then there exists $\varepsilon(l, n)$ such that $\varepsilon(l, n) = o(1)$ and

$$d_{\mathrm{tv}}(\mathcal{L}(\widetilde{V}_{n,l}), CP(\pi)) = \frac{1}{2}\sum_{k=0}^{\infty} |\mathbf{P}\{\widetilde{V}_{n,l} = k\} - \mathbf{P}\{\Xi_\pi = k\}| \leq$$

$$\leq 2H_1(\pi)\left(\mathbf{E}\widetilde{V}_{n,l}\right)^2 \frac{2^{2l}}{2^n}\left(1 + \varepsilon(l, n)\right) \to 0,$$

where $H_1(\pi) \leq \min\left(1, \dfrac{1}{\pi_1}\right) \cdot \exp\left(\sum\limits_{k=1}^{\infty} \pi_k\right)$.

References

[1] Erhardsson T. Stein's method for Poisson and compound Poisson approximation. — In Barbour A. D., Chen L. H. Y. (ed.) An introduction to Stein's method, Singapore Univ. Press, 2005, p.61–113.

[2] Hoffmann C.M., O'Donnell M.J. (1982). Pattern matching in trees. *J. ACM.* Vol. **29:1**, pp. 68-95.

[3] Steyaert J.-M., Flajolet P. (1983). Patterns and pattern-matching in trees: an analysis. *Inf. & Control.* Vol. **58:1**, pp. 19-58.

[4] Zubkov A.M., Mikhailov V.G. (1974). Limit distributions of random variables associated with long duplications in a sequence of independent trials. *Teoriya veroyatn. primen.* Vol. **19:1**, pp. 173-181 (in Russian; translated: *Theory Probab. Appl.* Vol. **19**:1, pp. 172-179).

Section 4

ECONOMETRIC MODELING AND FINANCIAL MATHEMATICS

EXACT D-OPTIMAL DESIGNS EXPERIMENTS FOR LINEAR MODEL WITH HETEROSCEDASTIC OBSERVATIONS

V. P. Kirlitsa
Belarusian State University
Minsk, BELARUS
e-mail: Kirlitsa@bsu.by

Abstract

The problem of construction exact D-optimal designs of experiments for linear model with heteroscedastic observations is investigated

Consider the linear model of heteroscedastic observations

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i(x_i),\ i = 1, ..., n,\ n \geq 2, \tag{1}$$

where y_i are observed variables; $\theta_0.\theta_1$ are unknown parameters; x_i are controllable variables from the interval $[-1, 1]$, $\varepsilon_i(x_i)$ are uncorrelated random errors of observations with mean zero and limited variances $D\{\varepsilon_i(x_i)\} = d_i(x_i) > 0$ for each realization x_i. Functions $d_i(x_i)$ satisfy to inequality:

$$d_i(x_i) \geq \frac{1}{4}\left([d_{i1} + d_{i2}]x_i^2 + 2[d_{i2} - d_{i1}]x_i + d_{i1} + d_{i2}\right), d_{i1} = d_i(-1),\ d_{i2} = d_i(1). \tag{2}$$

It is easy to check up that to these inequalities (2) satisfy constant functions ($d_i(x_i) = d = const$), with linear change ($d_i(x_i) = a_i + b_i x_i, |b_i| < a_i, a_i > 0$) and also all concave functions positive on [-1,1].

In [1] designs of experiments (1) are constructed in a case when variances $d_i(x_i)$ are defined by the same function $d(x_i)$. In this article result obtained in [1] is generalized on a case when the variance of each observation can be defined by own function.

Theorem 1. *For model of observations (1) at which variances of observations $d_i(x_i)$ satisfy to inequalities (2) there is an D-optimal exact design of experiment at which all spectrum points lay on interval ends of* $[-1.1]$.

Theorem 1 helps to obtain the following.

Theorem 2. *Spectrum points of exact D-optimal design of experiments for model of observations (1),(2) are co-ordinates of one of tops of n-dimensional cube $x_i = \pm 1, i = \overline{1, n}$ and these co-ordinates maximize function*

$$f(x_1, ..., x_n) = \left(\sum_{i=1}^{n} \frac{1 - x_i}{d_i(x_i)}\right)\left(\sum_{i=1}^{n} \frac{1 + x_i}{d_i(x_i)}\right),$$

provided that this maximum is calculated on all tops of this cube.

Proof. From Theorem 1 follows that it is necessary to search exact D-optimal design ε_n^0 among designs ε_n at which spectrum points lay in cube tops. Determinants of information matrix such designs are equal:

$$|M(\varepsilon_n)| = f(x_1, ..., x_n).$$

Exact D-optimal design is that design which maximizes function$f(x_1, ..., x_n)$. □

It is possible to receive a number of corollaries from the Theorem 2 in which the problem of construction D-optimal designs becomes simpler.

Corollary 1. *For model of observation (1). (2) with variances $d_i(x_i) = d_i$, $i = \overline{1,n}$ not dependent on x_i spectrum points x_i of D-optimal design lay on the ends of interval $[-1, 1]$ and these are such combinations of points for which the absolute value*

$$\left| \sum_{i=1}^{n} \frac{x_i^0}{d_i} \right|$$

take the minimum value.

From Corollary 1 follows that if design ε_n^0 is D-optimal design then the symmetric to him design $\varepsilon_1 = -\varepsilon_n^0$ is also the D-optimal design.

Let's consider now a case when a series from n independent observations breaks on two series of observations: $y_1, ..., y_{n_1}$; $\overline{y}_1, ..., \overline{y}_{n_2}$, $n = n_1 + n_2$. Observations $y_1, ..., y_{n_1}$ are realized in points $x_1, ..., x_{n_1}$ with equal variance $d_1 > 0$ and other observations $\overline{y}_1, ..., \overline{y}_{n_2}$ are realized in points $\overline{x}_1, ..., \overline{x}_{n_2}$ with equal variance $d_2 > 0$. In this case the problem of construction D-optimal designs is reduced minimizing expression

$$\left| \sum_{i=1}^{n_1} \frac{x_i}{d_1} + \sum_{i=1}^{n_2} \frac{\overline{x}_i}{d_2} \right| = \frac{1}{d_2} \left| \overline{d} \sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} \overline{x}_i \right|$$

on variables $x_i = \pm 1$, $i = \overline{1, n_1}$; $\overline{x}_i$, $i = \overline{1, n_2}$, where $\overline{d} = \frac{d_2}{d_1}$. Let is k number of x_i accepting value -1 and m is number of $\overline{x}_i$ accepting value -1. Then

$$\min \left| \overline{d} \sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} \overline{x}_i \right| = \min |\overline{d}(n_1 - 2k) + n_2 - 2m|, \tag{3}$$

where minimum on variables k, m is calculated on set: $0 \le k \le n_1$, $0 \le m \le n_2$. Solution of optimizing problem (3) defines structure of D-optimal design for heteroscedastic observations which are broken on two groups of homoscedastic observations. Construction of D-optimal designs can be generalized on a case when n heteroscedastic observations are broken into some series of homoscedastic observations but with different variances in each of series.

In that specific case when all functions $d_i(x)$ coincide, $d_i(x) = d(x)$ and function $d(x)$ satisfies to inequality

$$d(x) \ge \frac{1}{4}([d_1 + d_2]x^2 + 2[d_2 - d_1]x + d_1 + d_2),\ x \in [-1, 1]\, d_1 = d(-1),\ d_2 = d(1), \tag{4}$$

then process of construction D-optimal designs becomes easier and Theorem 3 takes place.

Theorem 3. *For model of heteroscedastic observations (1), (2) with variance $d(x)$ satisfying to inequality (4) exact D-optimal design ϵ_n^0 as well as for homoscedastic observations and has structure*

$$\varepsilon_n^0 = \begin{pmatrix} -1, & 1 \\ m, & n-m \end{pmatrix},$$

where m is number of observations in point -1. If $n = 2s$ is even number then $m = s$. If $n = 2s + 1$ is odd number then $m = 2s$.

Let's now consider a special case of Theorem 3. Let K is a set of points from interval [-1.1] in which inequality (4) turns to equality. If the set K contains other points from interval [-1,1] except points $-1, 1$ then in this case D-optimal design for odd number of observations can contain not only two but also three points.

Theorem 4. *If set $K\backslash\{-1, 1\}$ is not empty then for odd number $n = 2s + 1$ of observations D-optimal design has structure:*

$$\varepsilon_{2s+1}^0 = \begin{pmatrix} -1, & 1, & x \\ s, & s, & 1 \end{pmatrix}, x \in K.$$

Proof of this theorem you can find in article [1]. Consider now situation when in the condition of Theorem 3 function $d(x)$ is equal to zero on one of the interval ends of [-1.1]. Let's consider for definiteness that this function is equal to zero on the left end of interval [-1,1] and $d_1 = 0$, $d_2 > 0$. In this case the set of functions $d(x)$ defining change of variance of observations satisfies to inequality:

$$d(x) \geq \frac{d_2}{4}(x+1)^2,\ x \in [-1, 1]. \tag{5}$$

Let K_1 there is a set of points from interval $(-1, 1]$ in which inequality (5) turns to equality.

Theorem 5. *With probability 1 for model of heteroscedastic observation (1) which has variance $d(x)$ satisfying to inequality (5) exact D-optimal design of observations look like*

$$\varepsilon_n^0 = \begin{pmatrix} -1, & x_i, \\ 1, & 1, \end{pmatrix}, i = \overline{2, n}, \ x_i \in K_1. \tag{6}$$

The estimations of unknown parameters constructed under design (6) are following:

$$\overline{\theta}_1 = \left(\sum_{i=2}^{n} \frac{(1+x_i)^2}{d(x_i)} \right)^{-1} \sum_{i=2}^{n} \frac{\overline{y}_i}{d(x_i)}, \overline{\theta}_0 = \overline{\theta}_1 + y_{(-1)}, \tag{7}$$

where $\overline{y}_i = y_i - y_{(-1)}$; y_i are observations in the points x_i and $y_{(-1)}$ is observation in point -1. Variances of estimations (7) are equal to

$$D(\overline{\theta}_1) = D(\overline{\theta}_0) = \frac{d_2}{4(n-1)},\ d_2 = d(1).$$

Special case of Theorem 5 is the following.

Corollary 2. *If the set* $K_1\backslash\{1\}$ *is empty then in conditions of Theorem 5 D-optimal design looks like*

$$\varepsilon_n^0 = \begin{pmatrix} -1, & 1, \\ 1, & n-1 \end{pmatrix}. \tag{8}$$

The estimations of unknown parameters constructed under design (8)not dependent from variances and are equal to

$$\overline{\theta}_1 = \frac{1}{2}\left(\frac{1}{n-1}\sum_{i=2}^{n} y_i - y_{(-1)}\right),\ \overline{\theta}_0 = \theta_1 + y_{(-1)}$$

where y_i *are observations in point* 1.

There are full proof of Theorem 5 and corollary of this theorem in article [1].

Let's consider an example. We will construct designs of experiments for model (1), (2) for various cases of change of variances for $n = 4, 5$.

a) If the variances are

$$d_1(x) = \frac{3}{2} + \frac{1}{2}x,\ d_2(x) = \frac{3}{2} - \frac{1}{2}x,\ d_3(x) = \frac{7}{2} + \frac{1}{2}x,\ d_4(x) = 5,\ x \in [-1, 1],$$

then spectrum points of the unique D-optimal design are the following: $x_1^0 = -1$, $x_2^0 = 1$, $x_3^0 = -1$, $x_4^0 = 1$.

b) If the variances are $d_i = \frac{3}{2} + \frac{1}{2}x$, $i = \overline{1,4}$, $x \in [-1, 1]$ then in D-optimal design of experiments two spectrum points should lay on the left end of interval [-1,1] and other points to lay on the right end of this interval.

c) If the variances are $d_i(x) = \frac{3}{4}x^2 + \frac{1}{2}x + \frac{3}{4}$, $x \in [-1, 1]$, $i = \overline{1,5}$, then D-optimal designs should be built according to Theorem 4 in which $s = 2$, $K = [-1, 1]$. Such designs there are an infinite, incalculable set of power a continuum.

References

[1] Kirlitsa V.P. (2015). Exact D-optimal designs of experiments for linear regression with heteroscedastic observations. *Vestnik BSU*. Vol. **1**, pp. 97-102.

[2] Kirlitsa V.P. (2015). Analysis structure D-optimal designs of experiments with heteroscedastic observations. *Theory of probability, random processes, mathematical statistic and applications. Minsk*, pp. 35-37.

SOME APPROACHES TO CLASSIFICATION OF SUBJECTS OF FOREIGN ECONOMIC ACTIVITY BY RISK LEVEL

P. M. Lappo[1], T. A. Yakushava[2]
[1]*Belarusian State University*
[2]*Minsk Central Customs*
Minsk, BELARUS
e-mail: [1]lappopm@bsu.by

Abstract

Analyzed two approaches to classification of foreign economic activity subjects by the level of risk. Approaches using as risk measures the probability of customs legislation violation and the expected financial loss are considered.
Keywords: classification, customs risk, k nearest neighbors

1 Introduction

Currently the system of customs risk management in the Republic of Belarus is under development and improvement. As part of the work to a unified identification and analysis, providing for a minimum participation of the subjective human factor to be used probabilistic and statistical methods. One possible approach to risk management at the customs is a division of subjects of foreign economic activity (FEA) into three categories with high, medium and low risk. Appropriate type of customs control can be applied to each of these categories.

Different authors use various measures of customs risks. For example, in the paper [2] as a measure of risk the probability of violation of the customs legislation is used. The probability of violation is estimated using principal component method, cluster and regression analysis. Some authors use expert estimation for risk [3].

On the basis of data on subjects of foreign economic activity of Republic of Belarus who were exposed to check on customs legislation violation we have considered two approaches to classification of subjects by risk level. At the first approach as a measure of risk the probability of violation of the customs legislation was used, at the second — the expected losses for the budget.

2 Legislation violation probability approach

In this approach all subjects of foreign economic activity divided into classes depending on an estimator of probability of a violation of the legislation. For classification on three classes (low, medium and high risk levels) with use of a method of k nearest neighbors [1]. For the training sample the results are given in Table 1. Initially the subjects were distributed by classes with levels depending on the estimators of probability of violation: low for $[0, \frac{1}{3})$, medium for $[\frac{1}{3}, \frac{2}{3})$ and high for $[\frac{2}{3}, 1]$.

Note that only subjects with low risk level are classified more or less well. For the subjects with medium and high risk levels classification is rather unsatisfactory.

		Predicted			
		low	medium	high	Correct, %
Training	low	150	26	31	72.5
	medium	75	45	19	32.4
	high	63	23	51	37.2
	General, %	59.6	19.5	20.9	50.9

Table 1: Risk measure: probability of a violation of the legislation.

3 Expected losses approach

In this approach the risk is measured by means of the expected losses which the customs can incur. Mathematically we can represent the expected loss L_i from subject i in the form $L_i = S_i \cdot X \cdot p_i$, where S_i is the cost of the volume of goods moved by subject i, X is the average income from the detection of violation on a unit of value of the moved goods, in the presence of violation, p_i is the probability of violation for subject i. Having ordered all subjects who had checks (the training sample) by values of the expected losses, we can receive their classification. Initially every class contained equal number of subjects.

For classification of other subjects it is possible to use method of k nearest neighbors. Results of classification for the training sample by this method are given in Table 2. We find them more or less satisfactory.

		Predicted			
		low	medium	high	Correct, %
Training	low	125	68	44	52.7
	medium	26	168	44	70.6
	high	0	14	225	94.1
	General, %	21.1	35.0	43.8	72.5

Table 2: Risk measure: expected losses.

References

[1] Aivazyan S.A. et al. (1989). *Applied Statistics: Classification and Dimensionality Reduction*. Finansy i Statistika, Moskow. (in Russian)

[2] Afonin D.N., Afonin P.N., Scribe S.V. (2008). Computing methods of the analysis of customs risks. *Information customs technologies*. pp. 177-212. (in Russian)

[3] Solovyova I.V. (2008). *Statistical methods for evaluation in the system by the customs risk management: the example of the Southern Customs Department. PhD dissertation*. Rostov-on-Don. (in Russian)

MODELING THE REGIONS OF BELARUS COMPETITIVENESS BASED ON PANEL DATA

V. I. Lialikova[1], G. A. Khatskevich[2]
[1]*Grodno State University*
[2]*Institute for Economics, National Academy of Sciences of Belarus*
[1]*Grodno and* [2]*Minsk, BELARUS*
e-mail: [1]vlialikova@tut.by

Abstract

Comparative analysis of the regions of Belarus competitiveness based on panel data for 2011–2014 years was conducted. A system of indicators that reflect the competitiveness in the regions under study was built. It consists of five units: quality of the population, living standards, quality of social services, quality of the ecological niche, cultural condition of society, investment attractiveness. Integral indicator of the competitiveness for regions was built using the factor analysis. All baseline indicators were sorted according to their impact on the rating.

1 Introduction

Countries competitiveness is estimated annually by international non-governmental organization, the World Economic Forum (WEF). The Republic of Belarus has not taken part in the WEF ratings. Improving the Republic of Belarus competitiveness and the participation in the WEF ratings is scheduled for 2016–2020 by the Government program.

The competitive advantages of the country directly depend on the competitiveness of its regions. In this regard, forming of region competitiveness is the main goal in the task of improving the competitiveness of the country.

The region's competitiveness will mean the ability of the regional economy to stably produce and consume goods and services in competition with the goods and services produced in other regions, while ensuring the continued growth of quality of life [1].

This definition highlights two fundamental directions for providing the growth of the region competitiveness: achieving the high quality of life and improving the region investment attractiveness. Accordingly, the region competitiveness estimate is suggested to be performed based on these two groups of indicators.

2 The system of indicators

The system of indicators composed in this work consists of five units: quality of population (8 indicators), standard of living (4 indicators), quality of social sphere (4 indicators), quality of the environment (3 indicators), investment attractiveness (7 indicators).

Table 1: Factor loadings of indicators related to the first principal factor

Indicator	**Factor 1**
Percentage of employees with higher education organizations	**0.98**
Paid services for population, per capita	**0.91**
Age dependency rate	**−0.90**
Population provision with housing	**−0.89**
Ratio of per capita income to the minimum subsistence budget	**0.89**
Rate of migration increase	**0.85**
Retail turnover of trade	**0.84**
Provision with doctors	**0.82**
Registered unemployment rate	**−0.76**
Rate of natural increase	**0.69**
Share of innovation-active organizations in companies	**0.68**
Life expectancy	**0.65**

Official statistics, published in the collections of National Statistical Committee of the Republic of Belarus [2], are used for selected indicators.

The integral indicator of the Grodno region districts competitiveness was built based on panel data in [3] according to the 2008–2010 period.

Such an important factor in investment attractiveness, as the innovative activity of industrial organizations, has been being recorded in the official statistics since 2011. The regions of Belarus competitiveness estimation is built here taking into account this factor according to 2011 data. Competitiveness rating of regions is obtained base on panel data for 2011–2014 years.

A technique based on the methods of applied statistics was used for the construction of integral indicator [3].

Comparability of data was carried out by matching to the minimum consumer budget by the end of the year.

3 Competitiveness in the regions of Belarus

Original 26 indicators were scaled on the interval $[0, 1]$ for comparability of indicators, measured in different units. The indicators were then transformed according to the principal components method of factor analysis into the 4 principal factors. Thus all the indicators were associated with one of the 4 main factors. The total percentage of variance, saved by them, is 73.4 % (the first factor saves 35.9 % of the variance). The factor loadings values of the first principal factor are listed in table 1.

Integral indicator of the quality of life was obtained using the equation

$$R = 35.9F_1 + 15.7F_2 + 12.1F_3 + 9.7F_4,$$

where R is the competitiveness integral indicator, F_1, F_2, F_3, F_4 — values of the first principal factors. The percentage of the dispersion, saved by them, is taken as weight.

Table 2: The regions of Belarus competitiveness rating for 2011–2014 years

Region	**2011**	**Region**	**2012**	**Region**	**2013**	**Region**	**2014**
Minsk city	71.0	Minsk city	85.0	Minsk city	105.1	Minsk city	101.1
Brest	−16.0	Grodno	−3.2	Brest	18.6	Grodno	17.5
Grodno	−19.0	Mogilev	−8.9	Grodno	4.3	Brest	10.4
Mogilev	−31.7	Brest	−10.5	Mogilev	−3.9	Mogilev	3.3
Vitebsk	−43.1	Minsk	−30.7	Gomel	−5.0	Gomel	−5.2
Gomel	−48.2	Gomel	−33.1	Minsk	−14.2	Vitebsk	−13.1
Minsk	−55.2	Vitebsk	−35.3	Vitebsk	−22.8	Minsk	−17.2

Table 3: The regions of Belarus competitiveness dynamics for 2011–2014

Region	**Year**	R	**Region**	**Year**	R
Minsk city	2013	105.1	Minsk	2012	−30.7
Minsk city	2014	101.1	Minsk	2011	−55.2
Minsk city	2012	85.0	Vitebsk	2014	−13.1
Minsk city	2011	71.0	Vitebsk	2013	−22.8
Brest	2013	18.6	Vitebsk	2012	−35.3
Brest	2014	10.4	Vitebsk	2011	−43.1
Brest	2012	−10.5	Grodno	2014	17.5
Brest	2011	−16.0	Grodno	2013	4.3
Gomel	2013	−5.0	Grodno	2012	−3.2
Gomel	2014	−5.2	Grodno	2011	−19.0
Gomel	2012	−33.1	Mogilev	2014	3.3
Gomel	2011	−48.2	Mogilev	2013	−3.9
Minsk	2013	−14.2	Mogilev	2012	−8.9
Minsk	2014	−17.2	Mogilev	2011	−31.7

4 Comparative analysis of the competitiveness in the regions of Belarus

Comparative analysis on the basis of panel data allows you to not only build a rating of regions, but also to analyze the dynamics of the competitiveness of each region for the period of study. As a result regions can be sorted by years (table 2) and by their dynamics (table 3).

Minsk city is the permanent leader. Grodno and Brest regions are highly competitive due to the standard of living (table 2).

The lowest values of the integral indicators for the regions of the Republic of Belarus are observed in 2011. For Brest, Gomel, Minsk regions and Minsk city integral indicator of competitiveness decreased in 2014 comparing to 2013 year. Positive dynamics remained in Grodno, Vitebsk and Mogilev regions in 2014 (table 3).

5 Conclusion

The most important competitiveness growth factors of Grodno region districts for the period under review were revealed.

The quality of population: the proportion of employees with higher education in organizations, age dependency rate, rate of natural increase, rate of migration increase, life expectancy.

The investment attractiveness: share of the shipped innovative products.

The standard of living: population provision with housing, the ratio of per capita income to the minimum subsistence budget, retail turnover of trade, paid services for population.

Quality of social services: provision with doctors.

The same results were obtained by the study of the competitiveness in Grodno region [4].

In order to solve the identified problems it is, first of all, necessary to create new jobs and thus attract young working population to districts, as well as to implement a package of measures stimulating the development of small and medium businesses in the fields of material production, innovation and provision of public services.

References

[1] Pechatkin V.V., Perfilov V.A. (1985). Competitiveness of the Russian regions: Theoretical and methodological aspects of evaluation. *Problems of Modern Economics*. Num. **3**, pp. 285-290.

[2] National Statistical Committee of the Republic of Belarus (2015). *Regions of the Republic of Belarus 2015: main socio-economic indicators of cities and districts: statistical book*. Vol. **2**. National Statistical Committee of the Republic of Belarus, Minsk.

[3] Lialikava V.I. (2010). Methodological aspects of economic objects ranging by applied statistics methods. *Vestnik of Yanka Kupala State University of Grodno. Series 5*. Num. **2**, pp. 29-35.

[4] Lialikova V.I, Khatskevich G.A. (2013). Modeling the relationship of the quality of life and the investment attractiveness in Grodno region. Computer data analysis and modeling: Theoretical and applied stochastics. Vol. **2**, pp. 209–213.

STATISTICAL ESTIMATION AND TESTING OF TURNING POINTS IN MULTIVARIATE REGIME-SWITCHING MODELS

V. I. Malugin[1], A. Yu. Novopoltsev[2]
Belarusian State University
Minsk, BELARUS
e-mail: [1]malugin@bsu.by, [2]novopsacha@gmail.com

Abstract

For vector autoregressive models with Markov switching states (MS-VARX) we propose the algorithms of classification of states based on classified and non-classified learning samples. We also suggest the procedure to exclude short-term (acyclic) fluctuations in system states. It is based on successive application of algorithms implementing the Bayesian plug-in decision rule of point-wise classification and a statistical test for expected probability of misclassification. Accuracy of the algorithms is examined by means of computer simulation experiments.

1 Models and tasks of the research

Regime-switching models (RS-Models) are convenient for analyzing complex systems with cyclic changes of state [1]. Most studies are devoted to *Markov-switching* vector autoregressive model (MS-VAR) [2]. In the case of independent states *independent regime-switching* autoregressive and regressive models (IS-Models) should be used. These models are also preferable under the Markov dependence condition when there are high uncertainty about the future state of a system. The models of this type were thoroughly studied in [3, 4]. In this paper, the object of study is the vector autoregressive model with Markov-switching states including exogenous variables (MS-VARX), thus allowing a multivariate linear regressive model (MS-MLR) as its special case.

Let a complex system at time t is characterized by a random observation vector defined on the probability space $(\Omega, \mathfrak{F}, \mathbf{P})$, where Ω — space of elementary objects $\omega \in \Omega$; $\mathbf{P}$ — probability measure: $\mathbf{P}(A) = \mathbf{P}\{\omega \in A\}$, $A \in \mathfrak{F}$. Let $\{\Omega_0, \ldots, \Omega_{L-1}\}$ – decomposition of Ω into a finite number of non-empty disjoint subsets, such that $\Omega_l \in \mathfrak{F}$, $\mathbf{P}\{\Omega_l\} = \mathbf{P}(\{\omega \in \Omega_l\}) > 0$, $\bigcup_{l \in S(L)} \Omega_l = \Omega$, $S(L) = \{0, \ldots, L-1\}$. These subsets are the classes of states of a complex system, the number of which is L.

A random vector of observation $y_t = (x'_t, z'_t)' \in \Re^n$ can be partitioned into subsectors of endogenous variables $x_t = (x_{tj}) \in \Re^N$ and exogenous variables $z_t = (z_{tk}) \in \mathfrak{X} \subset \Re^M$. It is assumed that, in general, the time series is described by a model RS-VARX(p)($p \geq 1$):

$$x_t = \sum_{i=1}^{p} A_{d(t),i} x_{t-i} + B_{d(t)} z_t + \eta_{d(t)_t}, \quad t = 1, \ldots, T, \tag{1}$$

where $x_{1-p}, \ldots, x_0 \in \Re^N$ — the set of the given initial values; $\eta_{d(t),t} \in \Re^N$ — random disturbances or innovation process; $d(t) \in S(L) = \{0, \ldots, L-1\}$ — the class of state number.

Model (1) should satisfy with the following conditions:

M.1. Segmented-stationary condition: matrices $A_{l,i}(i = 1, \ldots, p)$ satisfy with the stationarity condition of VAR(p) model for each class of states $l \in S(L)$.

M.2. Disturbance assumptions: $\mathbf{E}\eta_{l,r} = 0_N \in \Re^N, \mathbf{E}\{\eta_{l,r}\eta'_{l,s}\} = \delta_{r,s}\Sigma_l$ $(r, s = 1, ..., T,\ l \in S(L))$, where $\delta_{r,s}$ — the Kronecker delta.

M.3. Exogenous variables $z_t = (z_{t1}, ..., z_{tM})' \in \mathfrak{X} \subseteq \Re^M$ are deterministic or stationary time series.

M.4. Structural heterogeneity conditions: $A_l \neq A_k$ and (or) $B_l \neq B_k\ \forall k \neq l,\ k, l \in S(L)$.

We consider a model with $L\,(2 \leq L < s + 1)$ classes of states, where $s \geq 1$ — number of state switching points $1 < \tau_1 < \ldots < \tau_s < T$. Concerning the sequence of states $d(t) \equiv d_t \in S(L)$ $(t = 1, ..., T)$ there are two types of assumptions:

d.1. d_t $(t = 1, ..., T)$ — independent identically distributed random variables with probability distribution $\mathbf{P}\{d_t = l\} = \pi_l > 0\,(l \in S(L))\,,\ \sum_{l \in S(L)} \pi_l = 1$; $\mathbf{P}\{d_t = l\} = \pi_l > 0\,(l \in S(L))\,,\ \sum_{l \in S(L)} \pi_l = 1$;

d.2. $d_t\,(t = 1, ..., T)$ — homogeneous ergodic Markov chain (GCM) with the distribution, which is determined by the vector of probability of the initial state π and a matrix of one-step transition probabilities P:

$$\pi = (\pi_l),\ \pi_l = \mathbf{P}\{d_1 = l\} > 0\ \ (l \in S(L))\,,\ \sum\nolimits_{l \in S(L)} \pi_l = 1;$$

$$P = (p_{kl})\,,\ p_{kl} = P\{d_{t+1} = l |\, d_t = k\} \geq 0,\ \sum\nolimits_{l \in S(L)} p_{kl} = 1,\ k \in S(L).$$

Under the conditions of d.1 and d.2, we get the models IS-VARX and MS-VARX, respectively. Model (1) allows for a number of special cases: a model of multivariate linear regression RS-MLR, if $p = 0,\ M \geq 1$ [4]; a model RS-VAR without exogenous variables, if $p > 0,\ M = 0$ [2].

The true values of a model parameters $\{A_l, B_l, \Sigma_l\}(l \in S(L))$, π, P and the moments of switching state $\{\tau_i\}(i = 1, ..., s)$ are unknown. There are a classified or a non-classified sample of observations $(\bar{X},\ \bar{Z})$ $(\bar{X} = (x_t) \in \Re^{NT},\ \bar{Z} = (z_t) \in \mathfrak{X}^T \subseteq \Re^{MT})$ when a vector of states $\bar{d} = (d_t) \in S^T(L)$ is known and unknown, respectively. We presented two statistical classification algorithms for MS-VARX model in these cases: an EM-algorithm for joint estimation of the parameters and the vector of states for a non-classified sample and a discriminant analysis algorithm in the case of a classified sample for classification of out-of-sample observations. To eliminate short-term fluctuations of states arising from misclassification, we propose a statistical test based on a pointwise classification decision rule. For IS-MLR and IS-VARX models the problems mentioned are solved in [3, 4].

2 Representations for the model parameter estimates

Model (1) under the assumptions M.1–M.4, d.1 or d.2 can be represented in the regression form

$$x_t = \Pi_{d(t)} u_t + \eta_{d(t),t}, \tag{2}$$

where $\Pi_{d(t)} = (A_{d(t),1}, \ldots, A_{d(t),p}, B_{d(t)})$ is the block $N \times (pN+M)$-matrix of parameters; $u_t = (x'_{t-1}, \ldots, x'_{t-p}, z'_t)' \in \Re^{Np+M}$ — the stacked vector of predetermined variables formed from lagged endogenous and exogenous variables whose values are known at time t.

In this case we use a sample of observations $(\bar{X}, \bar{U})$, where $\bar{X} = (x'_1, \ldots, x'_T)' \in \Re^{NT}$ — the values of the endogenous variables, which correspond to the values $\bar{U} = (u'_1, \ldots, u'_T)' \in \Re^{NpT} \times \mathfrak{X}^T \subseteq \Re^{(Np+M)^T}$ of predefined exogenous variables. For the model (2) we will also denote:

$\theta_l \in \Re^m$ $(m = N \times (pN + M) + N\,(N + 1)/2\,)$ — stacked vector of the parameters for the class $l \in S(L)$ formed of independent elements of matrices $\{\Pi_l,\ \Sigma_l\}$ $(l \in S(L))$;

$\phi \in \Re^q$ $(q = Lm + (L - 1)\,(L + 1))$ — parameters of a mixture of distributions including $\{\theta_l\}$ and π, P, $\hat{\phi} \in \Re^q$ — statistical estimate of $\phi \in \Re^q$;

$D = (d_1, \ldots, d_T)' \in S^T(L)$ — state vector for the period under observation;

$\tilde{\gamma}_{l,t} = \mathbf{P}\{d_t = l|\bar{X},\, \bar{U};\, \tilde{\phi}\}$ — posteriori probability of the class $l \in S(L)$ at the moment t;

$\tilde{\xi}_{kl,t} = \mathbf{P}\{\, d_{t+1} = l|\, d_t = k;\, \bar{X},\, \bar{U};\, \tilde{\phi}\}$ — posteriori probability of a transition from class $k \in S(L)$ to class $l \in S(L)$ at the moment t $(t = 1, \ldots, T - 1)$.

For a joint estimation of all the parameters and the state vector an EM-algorithm (*Expectation-Maximization algorithm*) is proposed. This algorithm belongs to the family of Baum–Welch algorithms for splitting a mixture of multivariate distributions, controlled by a hidden Markov chain [5]. In accordance with the general approach [4, 5], we obtain an analytical representation for the estimated characteristics.

The representation for an estimate $\hat{\phi} \in \Re^q$ is obtained by maximization of the conditional expectation of the log-likelihood function for some given initial value $\tilde{\phi} \in \Re^q$:

$$\hat{\phi} = \arg\max_{\phi \in \Re^q} \Lambda(\phi, \tilde{\phi}) = \arg\max_{\phi \in \Re^q} E_{\tilde{\phi}}\{l(\phi; \bar{X}, \bar{U}, D)|\bar{X}, \bar{U}; \tilde{\phi}\}, \tag{3}$$

$$l(\phi; \bar{X}, \bar{U}, D) = \ln(\pi_{d_1} p_X(x_1; u_1, \theta_{d_1})) + \sum_{t=2}^{T} \ln(p_{d_{t-1},d_t} p_X(x_t; u_t, \theta_{d_t})). \tag{4}$$

Theorem 1. *If the model (1), (2) satisfy the assumptions of M.1–M.4, d.2, then the estimates* $\{\hat{\Pi}_l, \hat{\Sigma}_l\}$ $(l \in S(L))$, $\hat{\pi}, \hat{P}$ *on a sample* $(\bar{X}, \bar{U})$ *are the solution of the problem (3), (4) for the given vector of parameters* $\tilde{\phi} \in \Re^q$*:*

$$\hat{\pi}_l = \tilde{\gamma}_{l,1}, \hat{p}_{kl} = \sum_{t=2}^{T} \tilde{\xi}_{kl,t} \left(\sum_{t=2}^{T} \tilde{\gamma}_{k,t-1} \right)^{-1}, \hat{\Pi}_l = \sum_{t=1}^{T} \tilde{\gamma}_{l,t} x_t u'_t \left(\sum_{t=1}^{T} \tilde{\gamma}_{l,t} u_t u'_t \right)^{-1}, \tag{5}$$

$$\hat{\Sigma}_l = \sum_{t=1}^{T} \tilde{\gamma}_{l,t}(x_t - \hat{\Pi}_l z_t)(x_t - \hat{\Pi}_l z_t)' \left(\sum_{t=1}^{T} \tilde{\gamma}_{l,t} \right)^{-1}, \tag{6}$$

where analytical representations for the posterior probabilities $\{\tilde{\gamma}_{l,t}\}$, $\{\tilde{\xi}_{kl,t}\}$ *are obtained as specified above.*

Corollary 1. *Using the known block structure for the matrices* $\hat{\Pi}_l$*, we can get the estimates* $\{\hat{A}_{l,1}, \ldots, \hat{A}_{l,p}, \hat{B}_l\}$ $(l \in S(L))$.

3 Classification and testing procedure

Bayesian decision rules (BDR) of pointwise and groupwise classification of multivariate observations described by IS-VARX model, have been proposed in [4]. In the case of a Markov-switching model we propose a decision rule of groupwise classification based on the dynamic programming approach described in [6].

Lemma 1. *If the model (1), (2) satisfy the assumptions of M.1–M.3, d.2, and parameters* $\phi \in \Re^q$ *are known, then a BDR of groupwise classification is determined by the condition*

$$\hat{D} \equiv \hat{D}(\bar{X}_1^T, \bar{U}_1^T) = \arg\max_{D \in S^T(L)} l(\phi; \bar{X}_1^T, \bar{U}_1^T, D), \tag{7}$$

where $(\bar{X}_1^T, \bar{U}_1^T)$ $(\bar{X}_1^T = (x'_1, ..., x'_T)' \in \Re^{NT}, \bar{U}_1^T = (u'_1, ..., u'_T)' \in \Re^{NpT} \times \mathfrak{X}^{MT} \subseteq \Re^{(Np+M)T})$ *is a sample of observations to be classified.*

To solve the problem (7) with a help of a dynamic programming approach, we use a special representation of the log-likelihood function $l(\phi; \bar{X}, \bar{U}, D)$ through the Bellman function [7].

Theorem 2. *Under the conditions of Lemma 1, a BDR of groupwise classification of sample* $(\bar{X}_1^T, \bar{U}_1^T)$ *is implemented using dynamic programming method in accordance with the following relationships:*

$$\hat{d}_T = \arg\max_{k \in S(L)} F_T(k), \ \hat{d}_t = \arg\max_{k \in S(L)} \left(f_t(k, \hat{d}_{t+1}) + F_t(k) \right), \ t = T-1, T-2, ..., 1, \tag{8}$$

$$F_1(l) \equiv 0, \ F_{t+1}(l) = \max_{k \in S(L)} \left(f_t(k, l) + F_t(k) \right), \ l \in S(L), \quad t = 1, ..., T-1, \tag{9}$$

where $\{F_t(k)\}$ *are Bellmans functions and* $\{f_t(k, l)\}$ *are determined by formula*

$$f_t(k,\, l) = \delta_{t,1} \left(\ln \pi_k + \ln p_X \left(x_1;\, u_1,\, \theta_k \right) \right) + \ln p_{kl} + \ln \left(x_{t+1};\, u_{t+1},\, \theta_l \right), \tag{10}$$

δ_{t1} — *Kronecker symbol,* $t = 1, \ldots, T-1$.

If $\{\hat{\theta}_l\}$ $(l \in S(L)), \hat{\pi}, \hat{P}$ are estimates of a model parameters, then using them in (9) we obtain a consistent "plug-in" decision rule. To find these estimates, it is advisable to apply the EM-algorithm proposed here. The "plug-in" BDR of group classification

can be used to classify out-of-sample observations (x_τ, u_τ) $(\tau = T+1, \ldots, T+h)$, that is, to forecast future states of a system.

We also suggest a procedure that allows to eliminate short-term (acyclic) fluctuations in system states, which caused by errors of classification of the proposed decision rules. It is based on application of algorithms implementing the Bayesian plug-in decision rule of pointwise classification and subsequent use of a statistical test for expected probability of misclassification [8].

References

[1] Hamilton J.D. (2008). Regime-switching models. *New Palgrave Dictionary of Economics. 2nd Edition. Palgrave Macmillan, Basingstoke*. pp. 1755-1804.

[2] Krolzig H.M. (1997). *Markov-switching vector autoregressions. Modelling statistical inference and application to business cycle analysis*. Springer, Berlin.

[3] Malugin V.I., Kharin Yu.S. (1986). On optimal classification or random observations different in regression equations. *Automation and Remote Control*. Vol. **7**, pp. 61-69 (in Russian).

[4] Malugin V.I. (2014). *Methods of analysis of multivariate econometric models with heterogeneous structure*. BSU, Minsk (in Russian).

[5] Bilmes J.A. (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models: Technical Report. *Int. Computer Science Institute*. Berkeley, USA.

[6] Kharin Yu.S. (1996). *Robustness in statistical pattern recognition*. Kluwer Academic Publishers, Dordrecht.

[7] Malugin V.I., Novopoltsev A.Yu. (2015). Analysis of multivariate statistical models with heterogeneous structure in the case of hidden Markov dependence of states. *Proc. National Academy of Science of Belarus: physics and mathematics series*. Vol. **2**, pp. 26-36 (in Russian).

[8] Malugin V.I. (2015). Algorithms of testing the cyclic structural changes in the vector autoregression models with switching states. *Informatica*. Vol. **4**, pp. 5-16 (in Russian).

ON THE PROBABILITY DISTRIBUTION PROCESSES SOME MODELS OF INTEREST RATES

G. A. Medvedev
Belarusian State University
Minsk, BELARUS
e-mail: MedvedevGA@bsu.by

Abstract

Presents the probability density and their properties for some stochastic models of short-term interest rates of yield, the authors previously proposed constructively without probabilistic analysis of their properties.

1 Introduction

There are many different models of short-term interest rates of the class of diffusion processes. Most of them are well documented by the authors, which offered them, or those who use them for their studies. However, there is a set of models tend to be fairly complex, probabilistic description of the properties which are absent in the literature. It is they who are the subject of our consideration. The main problem that we are interested is getting analytical expressions for the stationary probability densities and its main moments. Some models, such as models Vasiek (1977), Cox - Ingersoll - Ross (CIR) (1985), Duffie - Kan (1996), Ahn - Gao (1999), are well documented in the literature, therefore are not described here and not mentioned in the list of references. All considered models belong to the class of diffusion models, that generate processes $X(t)$, described by the equation

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dW(t), t > t_0, X(t_0) = X_0,$$

where a specific determination of drift $\mu(x)$ and volatility $\sigma(x)$ defines one or another particular model.

2 The Ait-Sahalia model [1]

Ait-Sahalia has tested the based models of short interest rates (including described here) by fitting them to the actually time series of rates. It was found that an acceptable level of goodness-of-fit all these rates were rejected because the drift and volatility properties. As a result he proposed the following functions drift and diffusion

$$\mu(r) = \alpha_0 + \alpha_1 r + \alpha_2 r^2 + \alpha_{-1}\frac{1}{r}, \sigma^2(r) = \beta_0 + \beta_1 r + \beta_2 r^2.$$

In this model, the non-linear functions of drift and diffusion allow a wide variety of forms. To $\sigma^2(r) > 0$ for any r, it is necessary that the diffusion function parameters ensure the fulfilment of inequalities

$$\beta_0 > 0, \beta_2 > 0, \gamma^2 \equiv 4\beta_0\beta_2 - \beta_1^2 \geq 0.$$

Relevant in this function a probability density is given by expression

$$f(x) = Nx^B(\beta_0 + \beta_1 x + \beta_2 x^2)^{C-1} e^{Ax + Garctg(E+Fx)}, x > 0,$$

where N is normalization constant,

$$A = \frac{2\alpha_2}{\beta_2} < 0, B = \frac{2\alpha_{-1}}{\beta_0} > 0, C = \frac{\alpha_1}{\beta_2} - \frac{\alpha_2\beta_1}{\beta_2^2} - \frac{\alpha_{-1}}{\beta_0},$$

$$G = 2\left(2\alpha_0 + \frac{\alpha_2\beta_1^2}{\beta_2^2} - \frac{\alpha_1\beta_1}{\beta_2} - \frac{2\alpha_2\beta_0}{\beta_2} - \frac{\alpha_{-1}\beta_1}{\beta_0}\right) \Big/ \gamma,$$

$$E = \beta_1/\gamma,\ F = \beta_2/\gamma.$$

Since the density $f(x)$ at $x \to 0$ has order $O(x^B), B > 0$, and at $x \to \infty$ its order is $O(x^{B+C}e^{Ax}), A < 0$, then for every finite m the moments $E[X^m]$ are exist, but their analytical expressions can not be obtained, and they can be calculated only numerically.

3 The CKLS model [2]

In Chan - Karolyi - Longstaff - Sanders (CKLS) model it is assumed that $\mu(x) = k(\theta - x), \sigma^2(x) = \sigma^2 x^3$. It turns out that a random process corresponding to this model has a stationary density

$$f(x) = \frac{n}{x^3} e^{-c((\frac{\theta}{x})^2 - 2\frac{theta}{x})}, x > 0,$$

where $c = \frac{k}{\theta\sigma^2}, n$ is normalization constant. Note that such random process has only the first stationary moment $E[X] = \theta$.

4 The unrestricted model I [3]

In "unrestricted model I"

$$dr = (\alpha_1 + \alpha_2 r + \alpha_3 r^2)dt + \sqrt{\alpha_4 + \alpha_5 r + \alpha_6 r^3} dW$$

are embedded some known models, that is, at a certain setting parameters $\{\alpha\}$ can get any of these known models. Table of according in this case has the form

Stationary probability density "unrestricted I" process has the form

$$f(x) = \frac{c(w)}{\sigma^2(x)} e^{\int_w^x \frac{2\mu(u)}{\sigma^2(u)} du} = \frac{c(w)}{\alpha_4 + \alpha_5 x + \alpha_6 x^3} e \int_w^x \frac{2(\alpha_1 + \alpha_2 u + \alpha_3 u^2)}{\alpha_4 + \alpha_5 u + \alpha_6 u^3} du,$$

where $c(w)$ is normalization constant, w is a fixed number from the set of possible values of a random process, the specific value of which does not play some role.

To get the explicit form of expression for $f(x)$ is possible, but it will be in general case quite cumbersome, and we restrict ourselves to the case when the values of the

Restrictions of parameters	Model	Equation of processes
$\alpha_3 = \alpha_5 = \alpha_6 = 0$	Vasicek	$dr = k(\theta - r)dt + \sigma dW$
$\alpha_3 = \alpha_4 = \alpha_6 = 0$	CIR	$dr = k(\theta - r)dt + \sigma\sqrt{r}dW$
$\alpha_3 = \alpha_6 = 0$	Duffie - Kan	$dr = k(\theta - r)dt + \sqrt{\alpha + \beta r}dW$
$\alpha_1 = \alpha_4 = \alpha_5 = 0$	Ahn - Gao	$dr = k(\theta - r)rdt + \sigma r^{1.5}dW$
$\alpha_3 = \alpha_4 = \alpha_5 = 0$	CKLS	$dr = k(\theta - r)dt + \sigma r^{1.5}dW$

parameters $\{\alpha\}$ provide performance properties of the probability density $f(x)$. First, we note that the volatility of the real process needs to be a real function, so $\sigma^2(r) = \alpha_4+\alpha_5 r+\alpha_6 r^3 \geq 0$ for all values of r. At the same analytic properties of the probability density depends on the type of the roots of equation $\alpha_4 + \alpha_5 r + \alpha_6 r^3 = 0$,$\alpha_6 > 0$. The sign of the discriminant $\Delta = (\frac{\alpha_5}{3\alpha_6})^3 + (\frac{\alpha_4}{2\alpha_6})^2$ specifies the number of real and complex roots of the equation. When $\Delta > 0$, there is one real and two complex conjugate roots. When $\Delta < 0$, there are three different real roots. When $\Delta = 0$, real roots are multiples.

Let $\Delta > 0$ and the real root is $r = r_0$, then we can write

$$\alpha_4 + \alpha_5 r + \alpha_6 r^3 = \alpha_6(r - r_0)(r^2 + pr + q),$$

where r_0, p and q are relatively sophisticated analytical expression and because of that are not listed here. However, if $\alpha_4 = 0$, then $r_0 = 0, p = 0, q = \frac{\alpha_5}{\alpha_6}$. In this case, the probability density is given by

$$f(x) = \frac{c(w}{\alpha_6 x(x^2 + \frac{\alpha_5}{\alpha_6})} e^{\int_w^x \frac{2(\alpha_1+\alpha_2 u+\alpha_3 u^2)}{\alpha_6 u(u^2+\frac{\alpha_5}{\alpha_6})} du} =$$

$$n x^{\frac{2\alpha_1}{\alpha_5}-1}(\alpha_6 x^2 + \alpha_5)^{\frac{\alpha_3}{\alpha_6}-\frac{\alpha_1}{\alpha_5}-1} e^{\frac{2\alpha_2}{\sqrt{\alpha_5\alpha_6}} arctg[x\sqrt{\frac{\alpha_6}{\alpha_5}}]},$$

where n is the normalization constant. For the existence of the probability density its parameters must satisfy the inequalities: $\frac{\alpha_1}{\alpha_5} > 1, \frac{\alpha_3}{\alpha_6} < 1$. In order to at the same time there exist stationary moments it is necessary for the expectation $\frac{\alpha_3}{\alpha_6} < 0,5$, for variance $\frac{\alpha_3}{\alpha_6} < 0$, for the third moment $\frac{\alpha_3}{\alpha_6} < -0,5$ and for the fourth moment $\frac{\alpha_3}{\alpha_6} < -1$.

If $\Delta < 0$, denote the roots of the equation $r_0 > r_1 > r_2$ so

$$\alpha_4 + \alpha_5 r + \alpha_6 r^3 = \alpha_6(r - r_0)(r - r_1)(r - r_2).$$

Then the probability density is expressed in the form

$$f(x) = n \prod_{i=0}^{2} (x - r_i)^{-1+2(\alpha_1+\alpha_2 r_i+\alpha_3 r_i^2)/\alpha_6 \prod_{j\neq i}(r_i - r_j)}. \quad (9)$$

In this case must be performed the inequalities

$$2(\alpha_1 + \alpha_2 r_0 + \alpha_3 r_0^2) > \alpha_6(r_0 - r_1)(r_0 - r_2),\ \alpha_3/\alpha_6 < 1.$$

For the existence of the m-th moment other than that necessary to perform the conditions $\frac{m}{2}+\frac{\alpha_3}{\alpha_6} < 1$. Unfortunately, the analytical expression of the normalization constant n and moments $E[r^m]$ very cumbersome, they include hypergeometric functions. Under these assumptions the process with such density has a bottom line equal to the largest root, i.e. $r(t) \geq r_0$.

Model	γ	$E[X]$	$Var[X]$	Skewness	Kurtosis
Vasicek	0	θ	$\frac{\sigma^2}{2k}$	0	3
CIR	0.5	$\frac{q}{c} = \theta$	$\frac{q}{c^2} = \frac{\sigma^2\theta}{2k}$	$2\sqrt{q}$	$3 + \frac{6}{q}$
Brennan - Schwartz	1.0	$\frac{q}{c} = \theta$	$\frac{\theta^2}{c-1}$	$\frac{4\sqrt{c-1}}{c-2}$	$\frac{3(c-1)(c+6)}{(c-2)(c-3)}$
CKLS	1.5	$\frac{q}{c} = \theta$	not exist	not exist	not exist

5 The unrestricted model II [2]

In the "unrestricted model II" process of short rate follows the equation

$$dr = k(\theta - r)dt + \sigma r^{\gamma} dW,\ \gamma > 0. \tag{1}$$

Therefore $\mu(x) = k(\theta - x), \sigma^2(x) = \sigma^2 x^{2\gamma}$ and the stationary density $f(x)$ has form

$$f(x) = \frac{n}{x^{2\gamma}} e^{\frac{1}{x^{2\gamma}}(\frac{qx}{1-2\gamma} - \frac{cx^2}{2-2\gamma})},\ x > 0, \tag{2}$$

where $q = \frac{2k\theta}{\sigma^2}, c = \frac{2k}{\sigma^2}, n$ is the normalization constant. Values of parameter γ, allowing the convergence of the integral of $f(x)$ on the interval $(0, \infty)$, determined by the inequality $\gamma > 0.5$. At the same time, there are two critical points: $\gamma = 0.5$ (in this case, the model is transformed into a short-term rate model CIR) and $\gamma = 1$, when the probability density is reduced to form that corresponds to process of the Brennan - Schwartz model [4]

$$f(x) = \frac{q^{1+c}}{x^{2+c}\Gamma(1+c)} e^{-\frac{q}{x}},\ x > 0.$$

When $\gamma = 1.5$, model "unrestrictions II" is known as the model CKLS. Vasicek model is also a model embedded in the model "unrestrictions II" at $\gamma = 0$. For existence of moments of order m, it is necessary the fulfilment of inequality $2\gamma > m + 1$. Unfortunately, the expression for the probability density in general case does not allow to calculate moments in analytical form, although for referred particular cases they simply calculated. For the model CIR

$$E[X^m] = \Gamma(m+q)/c^m\Gamma(q);$$

for Brennan - Schwartz model

$$E[X^m] = q^m\Gamma(1+c-m)/\Gamma(1+c),$$

the moments of order m exist if the inequality $m < 1 + c$ is fulfilled. So that

Even before the appearance of the model "unrestrictions II" there were used models, which then turned out to be special cases of this model. This is the model of the CIR (1980) [5], which is obtained from the equation (1), if we assume that $\gamma = 1.5$ and $k = 0$. Another particular version is the CEV model, i.e. model of constant elasticity of variance that was proposed J. Cox and S. Ross (1976) [6], as in equation (1) made

$\theta = 0$. Properties of the processes generated by these models can be understood by considering the limiting transition $k \to 0$ in the first model or $\theta \to 0$ in the second. When k and θ still finite the stationary regimes in the models exist and the probability density of processes for these models is expressed in the form (2). However, in the limiting case $k = 0$ or $\theta = 0$ stationary regimes of processes no longer exist, and the probability density can not be expressed in the form (2), and can be obtained as solutions of partial differential equations

$$\frac{\partial f(x,t|y,s)}{\partial t} - \frac{1}{2}\frac{\partial^2[\sigma^2 x^3 f(x,t|y,s)]}{\partial x^2} = 0$$

for model CIR (1980) and

$$\frac{\partial f(x,t|y,s)}{\partial t} + \beta\frac{\partial[xf(x,t|y,s)]}{\partial x} - \frac{\sigma^2}{2}\frac{\partial^2[x^{2\gamma} f(x,t|y,s)]}{\partial x^2} = 0$$

for model CEV at the boundary condition for both equations

$$\lim_{t\to s} f(x,t|y,s) = \delta(x-y).$$

Unfortunately, these equations can not be solved analytically, but we can say that for $k = 0$ or $\theta = 0$ the process generated by the equation (1) becomes unsteady for the CIR model (1980) with the constant expectation and increasing with time variance, and for model CEV changing with time both the expectation and the variance.

References

[1] Ahn D.-H., Gao B. (1999). A parametric nonlinear model of term structure dynamics. *The Review of Financial Studies*. Vol. **12**(4), pp. 721–762.

[2] At-Sahalia Y. (1996). Testing continuous-time models of the spot interest rate. *Review of Financial Studies*. Vol. **9**(2), pp. 385–426.

[3] Brennan M.J., Schwartz E.S. (1979). A continuous time approach to the pricing of bond. *Journal of Banking and Finance*. Vol. **3**, pp. 135–155.

[4] Chan K.C., Karolyi G.A., Longstaff F.A., Sanders A.S. (1992). An empirical comparison of alternative models of the short-term interest rate. *J. Finance*. Vol. **47**, pp. 1209–1227.

[5] Cox J.C., Ingersoll J.E., Ross S.A. (1980). An analysis of variable rate loan contracts.*J. Finance*. Vol. **35**, pp. 389–403.

[6] Cox J.C., Ross S.A. (1976). The valuation of options for alternative stochastic processes. *J. Financial Economics*. Vol. **3**, pp. 145–166.

MINIMUM DISTANCE FROM POINT TO LINEAR VARIETY IN EUCLIDEAN SPACE OF THE TWO-DIMENSIONAL MATRICES

V. S. Mukha
Belarusian State University of Informatics and Radioelectronics
Minsk, BELARUS
e-mail: mukha@bsuir.by

Abstract

This work relates to the problem of linear approximation of multidimensional statistical data. Instead of the approach of regression analysis, we want to use another approach which is to minimize of the sum of the squares of the perpendicular distances from the system of points to the approximating plane. We receive the formula of minimum distance from point to linear variety in Euclidean space of the two-dimensional matrices as a first step in solving the problem.

1 Introduction

The approximation of statistical data by linear regression function minimizes the sum of the squares of deviations between observations of endogenous variables and variables predicted by regression function [1, 3, 7]. The another approach is to minimize of the sum of the squares of the perpendicular distances from the system of points to the approximating plane. This approach was considered in works [2, 5], however hasn't got the wide illumination in statistical literature. We want to apply this approach to matrix statistical data. We solve the first part of this problem. We give the formula of minimum distance from point to linear variety in Euclidean space of the two-dimensional matrices. Unlike the works [2, 5] we receive a new independent multidimensional-matrix solution of the problem.

2 Linear varieties in matrix arithmetical space

Let us denote $R_{[n_1 n_2]}$ the linear space of $(n_1 \times n_2)$ -matrices with real elements and operations of addition and multiplication on the real numbers and let us call it arithmetical matrix linear space. Any element $X \in R_{[n_1 n_2]}$ let us call a vector or point in $R_{[n_1 n_2]}$. The system of vectors $\{X_1, X_2, ..., X_m\}$ we will call linear dependent if there are the real numbers $\alpha_1, \alpha_2, ..., \alpha_m$ such that at least one of them not equal zero and $\alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_m X_m = 0$. If this equation is possible only when $\alpha_1 = 0, \alpha_2 = 0, ..., \alpha_m = 0$, then system of vectors is called linear independent.

We define also the linear varieties in parametric form in $R_{[n_1 n_2]}$:

$$X = C_0 + t_1 C_1 + t_2 C_2 + ... + t_{n_1 n_2 - r_1 r_2} C_{n_1 n_2 - r_1 r_2}, \tag{1}$$

where $C_0 = (c_{i_1,i_2,0})$, $C_1 = (c_{i_1,i_2,1})$, $C_2 = (c_{i_1,i_2,2})$, , $C_{n_1 n_2 - r_1 r_2} = (c_{i_1,i_2,n_1 n_2 - r_1 r_2})$, $i_1 = \overline{1, n_1}$, $i_2 = \overline{1, n_2}$, – linear independent $(n_1 \times n_2)$ -matrices in $R_{[n_1 n_2]}$, $t_1, t_2, ..., t_{n_1 n_2 - r_1 r_2}$

– scalar real parameters. By analogy with vector space R^m we will call the variety (1) $(n_1n_2 - r_1r_2)$-dimensional plane in $R_{[n_1n_2]}$, and matrices C_1 , C_2 , , $C_{n_1n_2-r_1r_2}$ – direction matrices of this plane [6].

Relationship between r_1 and r_2 can by any in the framework of inequality $1 \leq r_1r_2 \leq n_1n_2$, but more easy to interpretation is case when $r_1 = n_1$, $1 \leq r_2 \leq n_2$.

For the case $r_1 = n_1$, $1 \leq r_2 \leq n_2$ we receive a new form of linear variety (1). We rewrite (1) in form

$$X = C_0 +^{0,2} (CT), \tag{2}$$

where

$$C = (c_{i_1,i_2,i'_1,i'_2}) = ((c_{i_1,i_2})_{i'_1,i'_2}) = (\tilde{C}_{i'_1,i'_2}), \quad i_1, i'_1 = \overline{1, n_1}, \quad i_2, i'_2 = \overline{1, n_2 - r_2}, \tag{3}$$

is four-dimensional matrix with sections $C_1 = \tilde{C}_{1,1}$, $C_2 = \tilde{C}_{1,2}$, , $C_{n_1n_2-r_1r_2} = \tilde{C}_{n_1,(n_2-r_2)}$, and $T = (t_{i'_1,i'_2})$, $i'_1 = \overline{1, n_1}$, $i'_2 = \overline{1, n_2 - r_2}$, – $(n_1 \times (n_2 - r_2))$-matrix, that contains the parameters t_1, t_2, , $t_{n_1n_2-n_1r_2}$ as its elements, $^{0,2}(CT)$ is $(0, 2)$-convolute product of matrices C and T [4]. We present the matrices X , C_0 in (2) in form of the block matrices: $X = [X_{n_2-r_2}, X_{r_2}]$, $C_0 = [C_{n_2-r_2,0}, C_{r_2,0}]$, where

$$X_{n_2-r_2} = (x_{i_1,i_2}), \quad C_{n_2-r_2,0} = (c_{i_1,i_2,0}), \quad i_1 = \overline{1, n_1}, \quad i_2 = \overline{1, n_2 - r_2},$$

$$X_{r_2} = (x_{i_1,i_2}), \quad C_{r_2,0} = (c_{i_1,i_2,0}), \quad i_1 = \overline{1, n_1}, \quad i_2 = \overline{n_2 - r_2 + 1, n_2}.$$

The block $X_{n_2-r_2}$ is matrix, that contains the first $n_2 - r_2$ columns of matrix X, and block X_{r_2} is matrix, that contains the last X_{r_2} columns of X. We present also the matrix C in form of the block matrix $C = \{C_{n_2-r_2}, C_{r_2}\}$, and its blocks we define as follows:

$$\overline{C}_{n_2-r_2} = (c_{i_1,i_2,i'_1,i'_2}), \quad i_1, i'_1 = \overline{1, n_1}, \quad i_2, i'_2 = \overline{1, n_2 - r_2},$$

$$\overline{C}_{r_2} = (c_{i_1,i_2,i'_1,i'_2}), \quad i_1, i'_1 = \overline{1, n_1}, \quad i_2, i'_2 = \overline{n_2 - r_2 + 1, n_2}.$$

Now we can write two equations instead of equation (2):

$$\begin{cases} X_{n_2-r_2} = \overline{C}_{n_2-r_2,0} +^{0,2} (\overline{C}_{n_2-r_2} T), \\ \\ X_{r_2} = \overline{C}_{r_2,0} +^{0,2} (\overline{C}_{r_2} T). \end{cases} \tag{4}$$

Because the matrices C_1 , C_2, , $C_{n_1n_2-r_1r_2}$ are linear independent, the matrix $\overline{C}_{n_2-r_2}$ is not singular, and we can get the matrix T from first equation of system (4):

$$T =^{0,2} (\overline{C}^{-1}_{n_2-r_2} (X_{n_2-r_2} - \overline{C}_{n_2-r_2,0})),$$

where $\overline{C}^{-1}_{n_2-r_2}$ is the $(0, 2)$-inverse matrix to the matrix $\overline{C}_{n_2-r_2}$. Substitution this solution to the second equation of system (4) gives

$$X_{r_2} = \overline{C}_{r_2,0} +^{0,2} (\overline{C}_{r_2} {}^{0,2}(\overline{C}^{-1}_{n_2-r_2} (X_{n_2-r_2} - \overline{C}_{n_2-r_2,0}))). \tag{5}$$

The last expression shows that in the case of $(n_1n_2 - n_1r_2)$-dimensional plane in $R_{[n_1n_2]}$ the block X_{r_2} of the matrix X is linear expressed via its block $X_{n_2-r_2}$. The expression

(5) gives this dependence in explicit form for the second block X_{r_2} of matrix X. By analogy with a vector space R^m we can call the variety (1) when $n_1n_2 - n_1r_2 = 0$ ($r_2 = n_2$) as point in $R_{[n_1n_2]}$. When $n_1n_2 - n_1r_2 = n_1$ ($r_2 = n_2 - 1$) the linear variety (1) means that $n_2 - 1$ sections of matrix X (last its columns) linear depends on one its section (first column). When $n_1n_2 - n_1r_2 = n_1(n_2 - 1)$ ($r_2 = 1$) the linear variety (1) means that one its column (last column) linear depends on all previous its columns.

3 Distance from point to linear variety in Euclidean space of the two-dimensional matrices

We denote $E_{[n_1n_2]}$ Euclidean space of the two-dimensional $(n_1 \times n_2)$-matrices with the scalar product

$$(X, Y) = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} x_{i_1,i_2} y_{i_1,i_2} =^{0,2} (XY), \quad X, Y \in E_{[n_1n_2]}. \tag{6}$$

We call orthogonal a vectors X and Y from $E_{[n_1n_2]}$, if $(X, Y) =^{0,2} (XY) = 0$, and we call normalized a vector $X \in E_{[n_1n_2]}$, if $(X, X) =^{0,2} (XX) = 1$. We call orthonormal the system of vectors $X_1, X_2, ..., X_m \in E_{[n_1n_2]}$, if this vectors are pairwise orthogonal and each of them has single length, i.e. if

$$(X_i, X_j) =^{0,2} (X_iX_j) = \delta_{i,j},$$

$\delta_{i,j}$ – the Kronecker symbol.

Let $\xi = (\xi_{i_1,i_2})$, $i_1 = \overline{1, n_1}$, $i_2 = \overline{1, n_2}$, – matrix from $E_{[n_1n_2]}$. We formulate the task of finding the minimum distance from point $\xi \in E_{[n_1n_2]}$ to linear variety (1). In accordance with the scalar product (6) the square of distance is determined by formula

$$\rho^2(\xi, X) = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} (\xi_{i_1,i_2} - x_{i_1,i_2})^2 =^{0,2} (\xi - X)^2.$$

If we use in this formula the expression (2) for X, then we receive the optimization task:

$$\rho^2(\xi, X) =^{0,2} (\xi - X)^2 =^{0,2} (\xi - C_0 -^{0,2} (CT))^2 \to \min_T. \tag{7}$$

Now we go to the solving the task (7). We note, that we can write the variety (1) in form

$$X = C_0 +^{0,2} (TC^{T_1}),$$

where C^{T_1} is transposed matrix C in accordance with substitution $T_1 = \begin{pmatrix} i, j, k, l \\ k, l, i, j \end{pmatrix}$ [4]. Then the task (7) get form

$$\rho^2(\xi, X) =^{0,2} (\xi - X)^2 =^{0,2} ((\overset{o}{\xi} -^{0,2} (TC^{T_1}))(\overset{o}{\xi} -^{0,2} (CT))) \to \min_T, \tag{8}$$

where $\overset{o}{\xi} = \xi - C_0$. Because in (8)

$$\rho^2(\xi, X) =^{0,2} (\overset{o}{\xi}\overset{o}{\xi}) - 2^{0,2}(^{0,2}(\overset{o}{\xi}\, C)T) +^{0,2} (T^{0,2}(^{0,2}(C^{T_1}C)T)), \tag{9}$$

then necessary conditions for a minimum are next equation

$$\frac{d}{dT}\rho^2(\xi, X) = -2^{0,2}(\overset{o}{\xi}\, C) + 2^{0,2}(^{0,2}(C^{T_1}C)T) = 0.$$

From this equation we get

$$T =^{0,2} (^{0,2}(C^{T_1}C)^{-1\ 0,2}(\overset{o}{\xi}\, C)),$$

where $^{0,2}(C^{T_1}C)^{-1}$ is matrix $(0,2)$-inverse to the matrix $^{0,2}(C^{T_1}C)$. If we substitute this solution to the expression (9), then we get the square of minimum distance:

$$\rho^2_{min}(\xi, X) =^{0,2} (\overset{o}{\xi}\overset{o}{\xi}) -^{0,2} (\overset{o}{\eta}\ ^{0,2}(\overset{o}{\eta}\ ^{0,2}(C^{T_1}C)^{-1})), \tag{10}$$

where

$$\overset{o}{\eta} =^{0,2} (C^{T_1}\ \overset{o}{\xi}).$$

We have proved the following theorem.

Theorem 1. *Let $E_{[n_1n_2]}$ is Euclidean space of the two-dimensional $(n_1 \times n_2)$-matrices with the scalar product (6) and ξ is point in $E_{[n_1n_2]}$. The square of distance from point ξ to the linear variety (2) in $E_{[n_1n_2]}$ is defined by expression (10).*

References

[1] Draper N., Smith H. (1998). *Applied Regression Analysis. 3d edition*. Wiley.

[2] Cramer H. (1999). *Mathematical Methods of Statistics*. Princeton University Press.

[3] Mukha V.S. (2014). Multidimensional-matrix linear regression analysis. Distributions and properties of the parameters. *Proc. National Academy of Sciences of Belarus. Physic and Mathematics Series*. Vol. **2**, pp. 71–81.

[4] Mukha V.S. (2004). *Analysis of mulnidimensional data*. Technoprint, Minsk.

[5] Pearson K. (1901). On lines and planes of closets fit to systems of points in space. *Philosophical Magazine. Series 6*. Vol. **2**(11), pp. 559–572.

[6] Utesheu A.E. (2015). *Notebook on virtual faculty*.
`http://www.apmath.spbu.ru/ru/staff/uteshev/index.html`

[7] Vuchkou I, Boyadjieva I, Solakou E. (1987). *Applied Linear Regression Analysis*. Finansy and Statistika, Moskva.

MODELING OF REGIONAL SOCIO-ECONOMIC DEVELOPMENT OF BELARUS[1]

K. Navitskaya[2], B. Zhalezka
Yanka Kupala State University of Grodno
Grodno, BELARUS
e-mail: [2]navickaya@tut.by

Abstract

This article presents the results of the analysis of regional socio-economic development of Belarus based on calculated integral criterion. The membership functions of fuzzy clusters of Belarus regions by this criterion were developed and compared. The Cobb-Douglas production functions of Belarus region were constructed and analyzed. The results of research can be used for regional socio-economic prediction.

1 Introduction

The problems of evaluation and analysis of regional development are actual both in Belarus and abroad. A lot of different valuation techniques of estimation were developed and proposed considering the complexity and diversity of the object. Among the topic of research should be noted methodology of international assessment agencies in Europe (for example, Eurostat [1]), Russian scholars O. Kuznetsova, A. Bakhtizin, domestic rating agency under the supervision of prof. M. Kovalev, regional assessments of V. Lyalikova and others. In this paper the evaluation and modeling is not made by time series because their stationary is subject of doubt. Thats why its made by variation. The purpose of this paper is modeling of regional socio-economic development of Belarus based on calculated integral criterion and fuzzy clusters of level of regional development.

2 The fuzzy clusters of regional development

The theoretical basis of the modeling and analysis is the author's technique of multi-agent situation analysis. It allows to determine main factors of regional socio-economic development and to develop a methodology for calculating the index, which characterizes the impact of the local area at the result of socio-economic development of the high region. Previous testing of this technique was carried out on the areas of Grodno Region for the years 2008-2014 [4]. In this study calculations of this indicator of all areas (118 cases) and the cities of regional subordination (10 cases) of the Republic of Belarus in 2014 were carried out. The necessity of integral criterions calculating by the author's methodology is defined by the absence of a formal comprehensive index

[1]This work is supported by the TEMPUS project Fostering the Knowledge Triangle in Belarus, Ukraine and Moldova (FKTBUM 543853-TEMPUS-1-2013-1-DE-TEMPUS-SMHES).

of regional socio-economic development at local level. At the high level of regions this indicator is gross regional product. The presence of an integrated indicator allowing to assess the impact of various factors on the level of social and economic development [2]. The value of budget of the administrative-territorial unit, investments in fixed capital, retail turnover of trade and net exports of goods and services in current prices were summed for calculating the value of the integral criterion.

A lot of attention in regional economys research is given to inter-regional comparisons. In this regard, all local areas and cities of regional subordination of each region were divided into 4 fuzzy clusters: very high, high, medium and low level of development. Graphical representation of membership functions built on the example of Brest region are given on the figure.

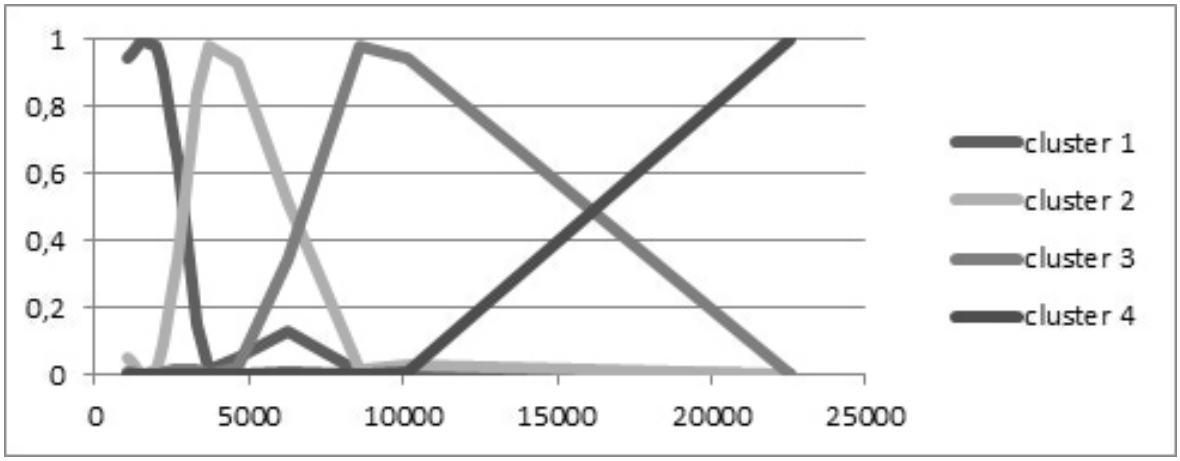


Figure 1: Graphical representation of the membership functions of local regions (of Brest region) of fuzzy clusters of socio-economic development

The Construction of membership functions provides its mathematical representation in the form of a piecewise-linear functions. For example, Brest regions membership functions get following formula:

$$F_1(x) = \begin{cases} 1, x \in (0; 2200) \\ -0,0006x + 2,213, x \in (2200; 3500) \\ 0, x \in (3500; \infty) \end{cases} \quad F_2(x) = \begin{cases} 0, x \in (0; 2200) \\ 0,0007x - 1,396, x \in (2200; 3200) \\ 1, x \in (3200; 4400) \\ -0,0002x + 2,001, x \in (4400; 8500) \\ 0, x \in (8500; \infty) \end{cases}$$

$$F_3(x) = \begin{cases} 0, x \in (0; 4700) \\ 0,0003x - 1,157, x \in (4700; 8550) \\ 1, x \in (8550; 9450) \\ -0,00008x + 1,719, x \in (9450; 22550) \\ 0, x \in (22550; \infty) \end{cases} \quad F_4(x) = \begin{cases} 0, x \in (0; 10000) \\ 0,00008x - 0,833, x \in (10000; 22550) \\ 0, x \in (22550; \infty) \end{cases}$$

where F_i(x) - the value of membership functions of cluster i, x - calculated value of gross regional product of local area.

Assuming stability of the economy it will be possible to determine the membership of the administrative-territorial unit to the particular cluster depending on the value of the estimated gross regional product in the future.

3 The production functions of regional development

The factor analysis revealed that the main factors affected to the level of socio-economic development are population (or the number of employed in the economy, among which the correlation coefficient is 0.99) and the value of investments. These factors may be taken as the base of economic developments modeling with the Cobb-Douglas production function. In general it can be represented by the formula:

$$Q = a_0 L^{\alpha} K^{\beta}$$

where Q - gross regional product; a_0 - factor which declare the level of technology development; L and K - numerical expression of labor and capital resources; α and β - characteristics of efficiency of resources using.

The model was linearized by taking the logarithm for assess the value of the regression coefficients. The method of least squares was used for models constructing. As a result, we obtained the production functions for each area. The table below shows the production functions coefficients.

Table 1: Coefficients of production functions of regions of Belarus.

Region	a_0	α	β
Brest region	23,4022	0,7158	0,3754
Vitebsk region	39,9304	0,7377	0,2954
Gomel region	42,8437	0,8984	0,2329
Grodno region	31,0446	0,8724	0,2736
Minsk region	45,0027	0,9978	0,1625
Mogilev region	30,3998	0,5682	0,4237
Source: author's own development.			

The determination coefficient for all constructed models is above 0.9, and p-value less than 0.001, the model is adequate. The free factor in all areas is quite close. In this article the simplest model of the Cobb-Douglas was used. The free factor of the model in the developed economic theory includes the level of technological development. Thus, theres no significant difference in the level of technological development in regions of the Republic of Belarus.

However the use of labor and capital factors in the production of gross regional product has noticeable difference. The analysis showed that Minsk regions capital flexibility is the lowest in the country, its much lower than the rate of labor flexibility.

This means that most of the cash resources of the region is spent on salaries, rather than to finance investment. This conclusion partly correspondes to the findings obtained by K. Rudyj [3].

Theoretically the values of the coefficients are stable over time if there is not major changes in the economy. In this case the models can be used for predicting social and economic development of regions.

4 Conclusion

In the research the integral criterion that takes into account the costs of households and the state, the value of businesss investments and net exports was proposed based on the technique of multi-agent situational analysis of regional development. The values of this criterion of all local areas and cities of regional subordination of the Republic of Belarus in 2014 were calculated.

All areas of each region were divided into 4 fuzzy cluster: very high, high, medium and low levels of development. In all regions, except Minsk, a cluster of very high development represented by only one object - the regional center. Comparing of membership functions of low development cluster in all regions allowed determining that Minsk and Brest regions are characterized by the development of higher and Vitebsk and Mogilev significantly lower than republican level.

Identification of key factors of socio-economic development was taken into account in the model of production functions of the gross regional product. We got that functioning of the economy in the regions of Belarus is mostly laboriousness. This is expressed by the fact that more than half of financial funds spent on salaries, rather than on investments. Thus the higher level of the regional development leads to the greater flexibleness of labor factor. This can be explained that the less developed areas have lower level of innovation and investments can significantly increase the impact of production and productivity. Highly developed areas need for this purpose the breakthrough technologies, which have higher cost and complexity of implementation.

References

[1] Eurostat (2014). Cohesion indicators.
Mode of access: `epp.eurostat.ec.europa.eu` Date of access: October 09, 2014.

[2] Navickaya K., Zhelezko B. (2013). The analysis of the influence of the information society's development on the economic performance in regions of Belarus. *Computer Data Analysis and Modeling. Theoretical and Applied Stochastics : Proceedings of the Tenth International conference*. Vol **2**, pp. 130-133.

[3] Rudyj K. (2014). *Financial diet: public finance reform in Belarus*. Zvezda, Minsk (in Russian).

[4] Zhelezko B, Navickaya K. (2015). Multy-criteria fuzzy analysis of regional development. *ECONTECHMOD*. Vol **4**(3), pp. 39-46.

MULTIVARIATE LINEAR REGRESSION WITH HETEROGENEOUS STRUCTURE AND ASYMMETRIC DISTRIBUTIONS OF ERRORS

A. Yu. Novopoltsev
Belarusian State University
Minsk, BELARUS
e-mail: novopsacha@gmail.com

Abstract

Multivariate econometric models with heterogeneous structure arise in economic and financial processes influenced by exogenous shocks. If structural heterogeneity is driven by presence of several classes of states in modeled complex systems, multivariate regime-switching econometric models is a choice. An assumption of normally distributed errors, which is traditionally held for such models, is often violated on real data. Therefore it is actual to develop multivariate regime-switching econometric models in presence of non-gaussian errors. In this paper, a multivariate regression model with switching regimes and asymmetrically distributed errors is proposed. A maximum likelihood approach is used to estimate the parameters of the model.

1 The model

Let us introduce an independent-switching multivariate linear regression model with errors distributed according to a class SNI [1] of asymmetric distributions, hereafter the IS-MLR-SNI model. The relation between endogenous and exogenous variables in the IS-MLR-SNI is expressed as follows:

$$x_t = B_{d(t)} z_t + \eta_{d(t),t}, \quad t = 1, \ldots, T, \tag{1}$$

where for a period of time t: $x_t = (x_{t1}, ..., x_{tN})' \in \mathbf{X}$, $\mathbf{X} \subset \Re^N$ $(N \geq 1)$ — vector of endogenous variables, $z_t = (z_{t1}, ..., z_{tM})' \in \mathbf{Z}$, $\mathbf{Z} \subset \Re^M$ $(M \geq 1)$ — vector of exogenous variables, $d(t) \in S(L) = \{1, ..., L\}$ — a state of a system modeled, $B_{d(t)}$ — regression coefficients matrix with a dimension $N \times M$, $\eta_{d(t),t} \in \Re^N$ — a random vector of heterogeneous errors.

For the model (1) the following assumptions are used.

1. Assumptions about observation errors:

a) observation errors have zero means and are mutually uncorrelated:

$$\mathbf{E}\{\eta_{d(t),t}\} = 0_N, \ \mathbf{E}\{\eta_{d(t),t}(\eta_{d(\tau),\tau})'\} = 0_{M\times N}, \quad t \neq \tau, \ (t, \tau = 1, \ldots, T); \tag{2}$$

b) observation errors have asymmetric distribution from a class SNI:

$$\eta_{d(t),t} \sim SNI_N(b\Delta_{d(t)}, \Sigma_{d(t)}, \lambda_{d(t)}, \nu), \quad t = 1, \ldots, T, \tag{3}$$

where $SNI_N(\mu, \Sigma, \lambda, \nu)$ — a class of multivariate asymmetric distributions [1] including skewed normal distribution and skewed t-distribution. The distributions from the

SNI class have the following parameters: $\mu \in \Re^N$ – location parameter; Σ – scale parameter, a covariance matrix with a dimension of $N \times N$; $\lambda \in \Re^N$ – skewness parameter; $H(u|\ \nu)$ – mixing distribution with a parameter $\nu \in \Re^{m_\nu}(m_\nu \geq 1)$; $\Sigma_{d(t)}$ $\lambda_{d(t)}$ – covariance matrix and skewness parameter for a state $d(t) \in S(L)$; $b\Delta_{d(t)}$ – parameter ensuring the condition $\mathbf{E}\{\eta_{d(t),t}\} = 0_N$, $b = -K_1\sqrt{2/\pi}$, $K_1 = \mathbf{E}\{U^{-1/2}|\nu\}$ – expectation of $U^{-1/2}$, where the random variable U is distributed according to $H(u|\nu)$, $\Delta_l = \Sigma_l^{1/2}\delta_l$, $\delta_l = \lambda_l/\sqrt{1+\lambda'_l\lambda_l}$, $l \in S(L)$.

2. Assumptions about the regime-switching model: the sequence of states following discrete time and space process with the distribution

$$P\{d_t = l\} = \pi_l > 0\,(l \in S(L))\,,\ \sum\nolimits_{l=1}^{L} \pi_l = 1, \tag{4}$$

where parameters $\{\pi_l\}$ $(l \in S(L))$ correspond to prior probabilities of states.

3. Condition of structural parametric heterogeneity:

$$B_k \neq B_l,\ k \neq l,\ k, l \in S(L). \tag{5}$$

4. Assumption about exogenous variables.

A vector of exogenous variables z_t is fixed for all realizations $\{z_t\},\ t = 1, \ldots, T$.

With assumption (3), the model IS-MLR-SNI may be represented in the form of the mixture of distributions with the following density function:

$$p(x_t|\Theta; z_t) = \sum\nolimits_{l=1}^{L} \pi_l sni_N\,(x_t|B_l z_t + b\Delta_l, \Sigma_l, \lambda_l, \nu), \quad t = 1, \ldots, T, \tag{6}$$

where $sni_N\,(x_t|B_l z_t + b\Delta_l, \Sigma_l, \lambda_l, \nu)$ — distribution density function for a random vector $x_t \in \Re^N$ against parameters Θ and fixed vector $z_t \in \Re^M$.

For model (1) in assumptions (2)–(5), let $\Theta = (\pi_1, \ldots, \pi_{L-1}, \theta'_1, \ldots, \theta'_L)' \in \Re^m$ be the stacked vector of all independent parameters, where $\theta_l = (b'_l, S'_l, \lambda'_l)' \in \Re^K$, $b_l = vec(B_l)$ denotes the vector of all elements of matrix B_l, S_l denotes the vector with the elements of upper triangular matrix of Σ_l, $\nu \in \Re^{m_\nu}$. Then the overall number of parameters equals $m = L - 1 + LK + m_\nu$, where $K = NM + N(N+1)/2 + N$.

2 Parameter estimation

To estimate the parameters of the model, we use an approach based on maximizing the likelihood function for the parameters Θ given a sample of regression observations $\{x_t, z_t\},\ t = 1, \ldots, T$. For derivation of the parameter estimates introduce the parameterization:

$$\Delta_l = \Sigma_l^{1/2}\delta_l, \qquad \Gamma_l = \Sigma_l^{1/2}(I_N - \delta_l\delta'_l)\Sigma_l^{1/2} = \Sigma_l - \Delta_l\Delta'_l, \quad l \in S(L) \tag{7}$$

where $\delta_l = \lambda_l/\sqrt{1+\lambda'_l\lambda_l}$, $\lambda_l \in \Re^N$ – skewness parameter for class l.

Let $X = (x'_1, \ldots, x'_T)' \in \Re^{TN}$ – stacked vector of all endogenous variables' realizations from the sample, $Z = (z'_1, \ldots, z'_T)' \in \Re^{TM}$ – stacked vector of all exogenous variables' realizations; $v = (v_1, \ldots, v_T)'$, $u = (u_1, \ldots, u_T)'$ – vectors of all realizations of

random variables V_t, U_t accordingly, where random variable V_t has the truncated univariate normal distribution with mean b and variance u_t^{-1} on the interval $(0, \infty)$ and depends on realization u_t of random variable U_t distributed according to the mixing distribution $H(u|\nu)$, and $\alpha_t = (\alpha_{t1}, \ldots, \alpha_{tL})'$ denotes the state indicator that has the multinomial distribution $M(1;\ \pi_1, \ldots, \pi_L)$.

Define the following expectations:

$$\begin{gathered}\rho_{tl} = \mathbf{E}_\Theta \left\{\alpha_{tl} | x_t, z_t\right\}, \beta_{tl} = \mathbf{E}_\Theta \left\{\alpha_{tl} U_t |\ x_t, z_t\right\}, \\ \xi_{tl} = \mathbf{E}_\Theta \left\{\alpha_{tl} U_t V_t |\ x_t, z_t\right\}, \omega_{tl} = \mathbf{E}_\Theta \left\{\alpha_{tl} U_t V_t^2 |\ x_t, z_t\right\}, t = 1, \ldots, T,\ l \in S(L),\end{gathered} \tag{8}$$

where α_{ti}, V_t, U_t – random variables, and the expectations (8) derived against fixed vector of parameters Θ and regression observations x_t, z_t with the following formula [1]:

$$\begin{gathered}\rho_{tl} = \frac{\pi_l sni_N\left(x_t | B_l z_t + b\Delta_l, \Sigma_l, \lambda_l, \nu\right)}{\sum_{j=1}^{L} \pi_j sni_N\left(x_t | B_j z_t + b\Delta_j, \Sigma_j, \lambda_j, \nu\right)}, \quad t = 1, \ldots, T,\ l \in S(L), \\ \beta_{tl} = \rho_{tl}\beta(x_t, z_t, \theta_l),\ \xi_{tl} = \rho_{tl}\xi(x_t, z_t, \theta_l), \\ \omega_{tl} = \rho_{tl}\omega(x_t, z_t, \theta_l), \quad t = 1, \ldots, T,\ l \in S(L),\end{gathered} \tag{9}$$

where $\beta(\cdot),\ \xi(\cdot),\ \omega(\cdot)$ are defined for basic distributions from SNI class as in [1].

Theorem 1. *Let $\tilde{\rho}_{tl}, \tilde{\beta}_{tl}, \tilde{\xi}_{tl}, \tilde{\omega}_{tl}$ be conditional expectations (8) derived against fixed vector of parameters $\tilde{\Theta}$ and regression observations sample $\{x_t, z_t\},\ t = 1, \ldots, T$. Then the maximum likelihood estimates of the parameters $\{\pi_l, B_l, \Delta_l, \Gamma_l\},\ l \in S(L)$ have the following representation:*

$$\hat{\pi}_l = 1/T \sum\nolimits_t^T \tilde{\rho}_{tl}, \tag{10}$$

$$\hat{B}_l = \sum\nolimits_{t=1}^{T} \left(\tilde{\beta}_{tl} x_t z_t' - \tilde{\xi}_{tl}\tilde{\Delta}_l z'_t\right) / \left(\sum\nolimits_{t=1}^{T} \tilde{\beta}_{tl} z_t z'_t\right)^{-1}, \tag{11}$$

$$\hat{\Delta}_l = \left[\sum\nolimits_{t=1}^{T} \tilde{\xi}_{tl}\left(x_t - \hat{B}_l z_t\right)\right] / \sum\nolimits_{t=1}^{T} \tilde{\omega}_{tl}, \tag{12}$$

$$\begin{gathered}\hat{\Gamma}_l = \left(\sum\nolimits_{t=1}^{T} \tilde{\rho}_{tl}\right)^{-1} \sum\nolimits_{t=1}^{T} \left\{\tilde{\beta}_{tl}\left(x_t - \hat{B}_l z_t\right)\left(x_t - \hat{B}_l z_t\right)' + \tilde{\omega}_{tl}\hat{\Delta}_l\left(\hat{\Delta}_l\right)' - \right. \\ \left. - \tilde{\xi}_{tl}\left[\left(x_t - \hat{B}_l z_t\right)\left(\hat{\Delta}_l\right)' + \hat{\Delta}_l\left(x_t - \hat{B}_l z_t\right)'\right]\right\}, \quad l \in S(L).\end{gathered} \tag{13}$$

To prove the theorem we follow the corresponding results from [1] considering density (6) from model (1) based on assumptions (2)–(5).

To recover the initial parameters λ_l, Σ_l the following formulae are used:

$$\begin{gathered}\lambda_l = \left(\Gamma_l + \Delta_l\Delta'_l\right)^{-1/2} \Delta_l / \left[1 - \Delta'_l(\Gamma_l + \Delta_l\Delta'_l)^{-1}\Delta_l\right]^{1/2}, \\ \Sigma_l = \Gamma_l + \Delta_l\Delta'_l, \quad l \in S(L).\end{gathered} \tag{14}$$

3 Objectives of the study

Regime-switching models are widely used in such applications as macroeconomics (real business cycles modeling), microeconomics (company credit risk modeling), financial markets (modeling and analysis of cyclical changes on financial markets) [2]. The problem of cyclical changes analysis with a help of the models mentioned may be considered in a context of more general problem of structural breaks analysis [3]. Structural breaks may be partial or full, that is parameters $\{B_l,\, \Sigma_l,\, \lambda_l,\, \pi_l\},\ l \in S(L)$ of the IS-MLR-SNI model may partially or fully distinguish across the states. The changes in the parameters may take place in any period of time $t = 1, \ldots, T$. A vector of states $d = (d_1, \ldots, d_T)'$ is unobserved.

To estimate structural breaks, a classification based approach is proposed. Therefore for the IS-MLR-SNI model (1) on the assumptions (2)–(5) we have the following problems to solve: 1) estimation of the parameters Θ of the model and the vector of states $d = (d_1, \ldots, d_T)'$ on unclassified sample of regression observations $\{x_t, z_t\},\ t = 1, \ldots, T$; 2) classification of new observations $\{x_\tau, z_\tau\},\ \tau = T{+}1, \ldots, T{+}h\ (h \geq 1)$ with the model estimated on the train data sample of size T. Problems 1 and 2 are solved with cluster and discriminant analysis algorithms respectively. For cluster analysis we use Expectation-Maximization (EM) algorithm. Earlier these problems were solved for multivariate regression models with switching regimes and normally distributed errors [4]. In [5] algorithms for analysis of multivariate regression observations with markov-switching regimes were presented.

In this study, for the solution of the problems an EM-type algorithm has been developed for the model (1) on the assumptions (2)–(5). An experimental study of the proposed algorithm is conducted on the simulated data.

References

[1] Cabral C.R.B., Lachos V.H., Prates M.O. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Statistics & Data Analysis*. Vol. **56**, pp. 126-142.

[2] Hamilton J.D. (2008). Regime switching models *New Palgrave Dictionary of Economics. 2nd Edition*. Palgrave Macmillan, Basingstoke.

[3] Peron P. (2006). Dealing with structural breaks *Palgrave handbook of econometrics. Volume 1: Econometric Theory*. Palgrave Macmillan, Basingstoke. pp. 376-389.

[4] Malugin V.I. (2014). *Methods for analysis of multivariate econometric models with heterogeneous structure*. Minsk, BSU. (in Russian)

[5] Novopoltsev A.Yu, Malugin V.I. (2015). Econometric forecasting with multivariate regression models with several classes of states *Econometrics, Modelling, Forecasting. Minisry of Economics of the Republic of Belarus, National Economic Research Institute*. Vol. **8**, pp. 206-214. (in Russian)

AUTOMATED REPORT ON THE BUSINESS PLAN OF THE INVESTMENT PROJECT

A. I. Zmitrovich[1], A. V. Krivko-Krasko[2], T. V. Lysenko[3]
[1]*Institute of IT and Business Administration*
[2]*School of Business and Management of Technology, Belarusian State University*
[3]*Belagroprombank*
Minsk, BELARUS
e-mail: [1]azmitrovich@iba.by, [2]sbmt@mail.ru, [3]t_poleschuk@tut.by

Abstract

The article is dedicated to the concepts and components of decision making support system for the possibility of providing credit support to enterprise for realization of investment project

1 Introduction

Currently in banks, consulting and financial companies are actual decision making support systems for the possibility of providing credit support to enterprises for realization of investment projects. Analysis of the conclusions of solving this problem infer that there is a possibility of building a computer system for automated building of conclusions on the business plan of the investment project . It seems that such a system should include the following components.

2 Information on the commercial organization

This section shall include: the abbreviated name of the organization (the project proponent); legal address; the current accounts and the name of the servicing bank; date of the last registration of the organization; the amount of the statutory fund; ownership; founder members; distribution of the statutory fund in shares; the average number of persons; industry affiliation; main activities; primary suppliers, consumers, competitors; position (market share) in the product market; special features of the technology used; main types of products; Purpose and main characteristics of the product.

3 Key performance indicators of the company

Further table is formed on the main indicators of the company for the last two years, with automatic calculation of the growth rate on the following parameters: the annual production capacity by product type; average number of persons; proceeds from realization of production; production and distribution costs; profit (loss) from sales of the products; net profit (loss); profitability of products sold; the share of non-cash payments in the revenue; the proportion of sales by markets: the domestic market,

neighboring countries, the far abroad. According to the formed table the system produces analytical conclusions on prefabricated frames in variable parts which fit values of the indicators and the conclusions: “more”, “above” or “less”, “better”, “worse”, “growth”, “loss”, etc.

Filled frame system might look like this. In 2015 the company produced 7.1 thousand tons of meat (including 3.7 thousand tons of beef and 3 thousand tons of pork), which is 7.8% more than in the previous year; 4 tons of sausage products (101.2% in 2014); 1.7 thousand tons of meat products (40.6% more than in 2014). The Group’s sales, excluding VAT was 8,474,600,000 rubles, the growth rate in 2014 — 114%. Net working capital — 492.7 million rubles.

4 Balance sheet structure of the organization

There follows a balance sheet structure of the organization for the last two years with also the indication of the proportion of this indicator in percentage to balance currency. Based on these data, the system calculates growth (+) or reduction (−) in absolute terms and growth rate in percent. The conclusions are based on these data and the corresponding frames.

For example, in the structure of circulating assets the largest share (49% on 1.1.2014 and 54% on 1.1.2015) took the receivables. The amount of buyers and customers debts to the company on 1.1.2015 is 3,212 million rubles, which 815 million rubles or 34% more than at the beginning of 2014. The growth of receivables is connected with an increase in the economic turnover of the enterprise: the turnover of receivables for the period from 1.1.2014 to 1.1.2015 decreased from 30 days to 24 days.

During the period under consideration the amount of inventories and costs decreased: from 5 010 million rubles to 4 631 million rubles. The reduction took place due to the reduction on 28% of the residues of production at the enterprise warehouses. The turnover of finished products decreased from 17 to 9 days.

It should be noted that the negative aspect in the analysis of calculation was the increase of the amount of receivables over the accounts payables more than 1.6 times, that indicates the use of bank loans as a source of free resources for the debtors.

5 Analysis of solvency ratios

Name of the ratio	**01.01.14**	**01.01.15**	**Standard**
Current ratio	0.82	0.75	≥ 1.7
Ratio of own current assets	-0.30	-0.36	≥ 0.3
Production ratio of financial liabilities assets	0.58	0.55	< 0.85

Current ratio characterizes the security of the enterprise’s own funds for business activities and timely repayment of urgent liabilities. During the analyzed period (as well as on 1.1.2015), the value of the current liquidity ratio was significantly lower than the standard.

Ratio of own working capital was also lower than the standard.

The value of the ratio of sufficiency of financial liabilities shows the independence of the enterprise from borrowed funds. Thus, the structure of the balance sheet "Enterprise" can be considered satisfactory, while the company is bankrupt.

6 Debt situation on credits and loans

Since January 2015, the companys debts on loans increased on 7368,3 million rubles or 45.3%, including credits on the current activity, the debt increased on 48.4%, while the debt on investment loans increased by 33.0%.

7 Information about the investment project

In this section, the following characteristics of the investment project are indicated in tabular form: project goals and objectives; horizon of business plans calculation; discount rate; currency; specific measures for implementation of the project with an indication of their value, the alleged suppliers, contractors; total investment costs including investment in fixed assets; VAT (value-added tax) on the amount of investment in fixed assets; net working capital growth; duration of the project; terms of development of capital expenditures; a period of performance producing on the scope of its planned production capacity; sources of financing of investment expenses.

According to the formed table the system produces analytical conclusions, for example on the following criteria: the conformity of investment costs volume to sources of funding (investment expenses coverage ratio), the conformity of investments volumes to reminder of the specific activities, the degree of readiness of the project to the date.

When calculating the effectiveness of the investment project the following indicators are analyzed: net present value; internal rate of return; profitability index; discounted pay-back period.

For example. Net present value shows the absolute value of net income, given to the beginning of the project. During the period of the project (2008-2015) the amount of remaining at the enterprise's disposal of the net present value will be 1716,8 million rubles.

Internal rate of return will be 13.9% for the under planned discount rate at a rate of 12.68%. Profitability index characterizes the impact of the project on the money invested in it. Effective are considered projects whose profitability index is greater than 1. For this project, the profitability index is 1.15.

Simple payback period is 4 years 11 months. Discounted pay-back Period is 6 years 3 months.

8 Planned indicators of financial-economic activity

The indicators are analyzed in two ways: without taking into account the implementation of the project and taking into account its implementation. It requires a comparison

of the main indicators on the options and the conclusion on the feasibility of the investment.

Sales revenue; sales revenues growth in relation to the base period,%; production and distribution costs, production and distribution costs in relation to the base period,%; semi-variable costs; growth of semi-variable costs in relation to the base period,%; fixed costs; the growth of fixed costs in relation to the base period,%; net profit; net income; debt service on credits and loans; cumulative balance (deficit) in cash; debt coverage ratio; break-even position; profitability of invested capital; profitability of production; profitability of sales; current ratio; size of working capital; a ratio of its own working capital; a ratio of financial obligations to assets; capital structure ratio; the share of short-term and long-term liabilities in the revenue; capital turnover period; current assets turnover term; finished goods turnover term; receivables turnover term; payables turnover term.

9 Updating the knowledge base and training decision making support system

In the process of working with the proposed system, there is a need to improve it by making changes and additions. This process involves the collection of requirements from the system user with further programmers realization for software changes. The form for submission of such a requirement is a production rule.

References

[1] Lobanova E.N., Zmitrovich A.I., Krivko-Krasko A.V., Voshevoz A.A. (2010). Financial decision making support system. *XX International Congress - AEDEM'2010*, pp. 341-346.

CALCULATION OF EUROPEAN OPTIONS WITH ABSOLUTE CRITERIA

N. M. Zuev
Belarusian State University
Minsk, BELARUS
e-mail: ZuevNM@bsu.by

Abstract

The problem of European options calculation is considered. The recurrent equations for the major characteristics are obtained.
Keywords: European option, absolute criteria, starting capital

1 Introduction

The problem of European options calculation is an important problem not only from the practical point, but also for the theory. In this paper we consider a (B, S)-market [1] and construct a sequence of recurrent equations for calculation of major characteristics of European type options for a self-financing portfolio.

2 Main Result

Let S_n be the cost of a risky active unit at the time moment n. Suppose it follows the model:

$$S_n = S_0(1+\rho_1)\dots(1+\rho_n),$$

where ρ_k, $k = 1, \dots, n$, are the interest rates with stochastic changes. Let B_n be the cost of a non-risky active unit at the time moment n that depends on the random variables $\rho_1, \dots, \rho_{n-1}$ only. Denote by $\pi_n = (\beta_n, \gamma_n)$ the portfolio at the time moment n, where β_n is the number of non-risky active units at the time moment n, γ_n is the number of risky active units at the time moment n, $n = 1, \dots, N$; N is the terminal moment, at this moment the option is executed. Denote by f_N the payment function that depends on random variables $\rho_1, \dots, \rho_N$ only. In case of a standard purchase of the European option, $f_N = (S_N - K)^+$ are the losses of the option seller for a risky active unit. Let K be the contract price for the purchase of a risky active unit at the time moment N. The variables β_n, γ_n are under prediction, and are supposed to depend on $\rho_1, \dots, \rho_{n-1}$ only. Let $X_n = \beta_n B_n + \gamma_n S_n$ be the portfolio cost at the time moment n.

The problem of calculation of a European option is concentrated on the choice of the starting capital $X_0 > 0$ (the option cost) and the portfolio $\pi_n = (\beta_n, \gamma_n)$, $n = 1, \dots, N$, so that $X_N - f_N = 0$ would be minimal.

In [2] – [4] the formulae to calculate the options with quadratic criteria are obtained.

In this paper we give the formula for calculation of options while minimizing $E\{|X_N - f_N|\}$, the minimum is 0.

Let us denote: $\tilde{X}_n = X_n/B_n$, $\tilde{S}_n = S_n/B_n$, $\tilde{f}_n = f_n/B_n$.

Theorem 1. *For the self-financing portfolio* $\pi_n = (\beta_n, \gamma_n)$, $n = 1, \ldots, N$, *the values* X_0, β_n, γ_n *are calculated recursively from the following equations:*

$$E\left\{|\tilde{f}_{n-1} + \gamma_n \Delta(\tilde{S}_n) - \tilde{f}_n|/\bar{\rho}_{n-1}\right\} = \min, \tag{1}$$

for $n \le N$*;*

$$\beta_n = \beta_{n-1} - \tilde{S}_{n-1}(\gamma_n - \gamma_{n-1}),\ \tilde{X}_0 = \tilde{f}_0, \tag{2}$$

$$\tilde{f}_{n-1} = E\{\tilde{f}_n \mid \bar{\rho}_{n-1}\} - \gamma_n E\{\Delta(\tilde{S}_n) \mid \bar{\rho}_{n-1}\},\ n = 1, \ldots, N.$$

In (1) the expectation is taken conditionally w.r.t. $\bar{\rho}_{n-1} = (\rho_1, \ldots, \rho_{n-1})$.

Proof. In case of a self-financing portfolio the value of X_n follows the law: $X_n = \beta_n B_n + \gamma_n S_n = \beta_{n+1} B_n + \gamma_{n+1} S_n$. From it we get $\frac{X_{n+1}}{B_{n+1}} - \frac{X_n}{B_n} = \gamma_{n+1}\left(\frac{S_{n+1}}{B_{n+1}} - \frac{S_n}{B_n}\right)$, or

$$\Delta\left(\frac{X_n}{B_n}\right) = \gamma_n \Delta\left(\frac{S_n}{B_n}\right).$$

From here we obtain:

$$\frac{X_N}{B_N} = \frac{X_0}{B_0} + \sum_{n=1}^{N} \gamma_n \Delta\left(\frac{S_n}{B_n}.\right) \tag{3}$$

From (3) we get $\tilde{X}_N - \tilde{f}_N = \tilde{X}_0 - \tilde{f}_0 + \sum_{n=1}^{N}\left(\tilde{f}_{n-1} + \gamma_n \Delta(\tilde{S}_n) - \tilde{f}_n\right)$, where the functions $\tilde{f}_n$, $n = 1, \ldots, N-1$ that depend on $\bar{\rho}_n$, are chosen below. □

Choose the values $\tilde{f}_{n-1}$ and γ_n from relations (1). From (1), (2) we find $\tilde{f}_n$, γ_n, β_n for $n \le N$. The value of $\tilde{X}_0$ is set to $\tilde{f}_0$. The values β_n are found from (2) and the portfolio cost at the time moment n.

In case where the random variables ρ_n take only two values for all n, equation (1) for arbitrary probabilities of these values turn into two linear equations $\tilde{f}_{n-1} + \gamma_n \Delta(\tilde{S}_n) - \tilde{f}_n = 0$ with $\tilde{f}_{n-1}$ and γ_n unknown.

References

[1] Shiryaev A.N. (1998). *Financial Mathematics.* Vol. 2. Moscow.

[2] Zuev N.M. (2013). Calculation of European type options. *Procceedings of the 10th International Conference on Computer Data Analysis and Modeling.* Minsk: BSU. Vol. **2**, pp. 188-189.

[3] Zuev N.M. (2014). Calculation of European and American type options. *Procceedings of the International Conference on Probability Theory, Random Processes, Mathematical Statistics and Applications.* Minsk: RIHS, pp. 64-66.

[4] Zuev N.M. (2015). On calculation of European and American type options. *Proceedings of the International Conference on Probability Theory, Random Processes, Mathematical Statistics and their Applications.* Minsk: RIHS, pp. 47-49.

Section 5

SURVEY ANALYSIS AND OFFICIAL STATISTICS

TECHNIQUE OF CALCULATING THE GENDER AND AGE SCALES OF REAL CONSUMER EXPENDITURE

Nina Agabekova
Belarus State Economic University
Minsk, BELARUS
e-mail: agabnin@mail.ru

Abstract

The article deals with approach to the construction of the gender and age scales of real consumer expenditure. We propose a regression model for calculating the numerical values of the equivalence scale and information base for the calculation. The results of testing technique are given.
Keywords: equivalence scale, consumer expenditure, household sample surveys, regression model

1 Introduction

The evaluation and comparison of household expenditure is necessary to solve the problems of social and economic policy. The comparison of expenditure of various groups of households basis reflect the economic stratification of society, including its aspects, as the distribution of the population in terms of income and consumption, and the boundaries of poverty.

Vast scientific and practical experience in the measurement of expenditure of households of different sizes and composition have accumulated abroad. The founder of this direction is the German statistician E. Engel, who proposed a method of estimating the welfare of households using such a statistical tool as the equivalence scale (see [1]). E. Rothbart (see [2]), M. Orsha (see [3]), J. Nicholson (see [4]), A Atkinson (see [5]) and others continued studies. At present, the equivalence scales are an integral part of the official methodology for poverty calculations and instrument of social policy. In the Republic of Belarus is used an expert scale which does not reflect real differences in consumption depending on gender and age.

2 Approach to the construction of the gender and age scales of real consumer expenditure

The author proposed a regression model for calculating the numerical values of the scale, where the resultant variable accepted consumption expenditure of the household, as well as signs of factors are used the number of household members in each age and gender group.

$$CE = \sum_{i=0}^{100} \sum_{j=1}^{2} \alpha_{ij} x_{ij} + \varepsilon, \tag{1}$$

where CE is consumption expenditure of the household (the total cost of food, the cost of food outside the home, pet food, drink, tobacco, clothing, footwear, textiles, goods cultural and household goods, fuels for home heating, utility costs, education services, early childhood education, health care, the costs of public transport services and associated with the operation of personal transport, expenditure on culture, recreation and sports, services, personal care items and other goods and services); i is the age of household member (0-4 years, 5-9 years, 10-14 years, 15-19 years, 20-24 years, 25-29 years, 30-34 years, 35-39 years, 40-44 years, 45-49 years, 50-54 years, 55-59 years, 60-64 years, 65-69 years, 70-74 years, 75-79 years, 80+years); j is the gender of the household member (1 - male, 2 - female); x_{ij} is the number of representatives of different gender and age groups in the household.

Regression estimates are used to determine individual consumption expenditure for each member i of the household j:

$$IE_{ij} = \frac{\alpha_{ij}}{\sum_{i,j} \alpha_{ij}} \cdot CE, \tag{2}$$

where IE_{ij} is individual consumption expenditure household member.

Technique of calculation is based on actual consumer expenditure of households and takes into account the gender, the age structure and the size of the household.

The proposed model requires a combination of annual sample surveys files by members of households and households in general. To this end, the annual data on the members of households participated in the sample survey are grouped according to the serial number of the household (table rows) and selected age groups (table columns) to the layout depending on the gender. The data are transferred in the annual image for households as a whole, where the variable is created CE ($\sum_{i \in I} \mathrm{EXPEND}_i$, $I = \{1, \ldots, 9, 11, 13, 14, 15, 21, 22, 23, 27, 28, 31, 32\}$).

The models are designed according to the household sample surveys, which include more than 5000 households, combining more than 14000 members, it is quite a large collection. Regression coefficients in the models are named number (rubles per person at age x), which is shown as the average change consumer expenditure is changing the age of the individual.

3 The results of testing technique

The author obtained estimates of age and gender ratios of consumption, taking into account the effect of cohabitation on the basis of sample surveys of households in the Republic of Belarus for 2008. The results give a non-significant regression coefficients in the age groups of men and women to 14 years, since the equation takes into account not only the age and composition, but also the number of members of the household. It is logical that the majority of consumer expenditure in the household is related adult persons. To obtain more reliable estimates of possible testing factors, specifying the age groups for equality between itself and the union of these age categories. However, in this study, the scale is calculated at five-year groups, which ensures equal age intervals. Quality assessment parameters confirmed the normal distribution model residues.

On the basis of the formula (2) regression estimates define individual consumer expenditure household member and find the average consumer expenditure in each gender and age group and population (see [6]).

4 Findings

The highest consumption rates in the household in both men and women of working age are marked. In addition, the need for women's consumption of almost all ages are superior to the consumption needs of men, which may be due to the need of women to spend more money on clothes, shoes, cosmetics. It is also no secret that women make up the majority of visitors to exhibitions, theaters and sports clubs. Higher food costs men in middle age are neutralized decrease in the proportion of expenditure on food in total. As already noted coefficients of consumption of persons under 14 years are underestimated. To overcome this drawback, in the framework of improving the techniques necessary to assess consumer groups spending data of the population only on set of households, having in its composition of this age, and refine the estimates obtained.

References

[1] Engel E. (1898). Human Value. *Library of general interest, for self*. Vol. **14**.

[2] Rothbarth E. (1943). Note on a method of determining equivalent income for families of different composition. *War time pattern of saving and spending*.

[3] Orshansky M. (1965). Who is among the poor: a demographic view of poverty. *Social security bulletin*. Vol. **28**, pp. 3–39.

[4] Nickolson J. (1980). Appraisal of different methods of estimating equivalence scales and their results. *The Review of income and wealth*. Vol. **70**(1).

[5] Atkinson A. (1989). *Poverty and social security*. Vol. **4**, pp. 12–57.

[6] Agabekova N. (2015). *Methodology of economic-statistical evaluation of the effectiveness of human activity: monograph*. Minsk: BSATU.

MICRO-ENTITIES AND SMALL ENTERPRISES SURVEYS IN BELARUS

Natalia Bokun
Belarus State Economic University
Minsk, BELARUS
e-mail: nataliabokun@rambler.ru

Abstract

The paper briefly describes the sampling methodology of micro-entities and small enterprises, problems of introduction of the micro-entities sample survey in practice of Belarusian official statistics. The sampling frame, sampling design and precision estimation are considered.
Keywords: micro-entities, sample survey, sampling frame, weighting, small enterprises

1 Introduction

In recent years, the growing number of small enterprises has motivated the development of specialized methodology and software for micro-entities and small enterprises sample surveys.

Since 2005 and until 2008 sample surveys of small enterprises spent quarterly. Survey objects were artificial persons of small business, i.e. SE. According to the legislation this was the organization with number of employees 100 persons and less. Sample frame was the file of SE. The territorial one-stage stratified sample was used. But in 2008 quarter survey was cancelled. Due to this reason only annual continuous small enterprises survey is conducted.

Nowadays, the National Statistical Committee of the Republic of Belarus together with Department of Statistics (BSEU) makes the preparatory work on implementation of the micro-entities and small enterprises sample surveys. In November 2014 a test sample survey was conducted; since 2015 Micro-entities Sample Survey (MS) is provided on a regular basis. The first results of Micro-entities Sample Survey indicated the appearance of significant organizational and methodological problems: non-responses, the need for localization of the sample, the presence of atypical units, using a combination of statistical weighting methods, samples in small domains.

This paper on small business sampling has the next parts:

- history of development of branch sample surveys;
- small enterprises sample survey;
- micro-entities sampling frames that incorporate two files of economic units: micro-entities and private farms;
- micro-entities sample design; territorial stratified univariate and multivariate (multidimensional) samples are used. The algorithm to receive optimal sample size for i-th kind of activity and j-th region is presented;

- statistical weighting that includes three methods: traditional Horvitz-Thomson estimator and calibration (GREG- and SYN-estimators).

2 History of development of branch sample surveys

In the conditions of command economy in the national statistics of Belarus, as well as in other countries of FSU Region, a priority it was given to methods of continuous survey with the exception of 3.5 thousand family budgets survey of workers, employees and collective farmers. Then the two-level stratified sample was used: at the first step the enterprises was selected within branches, than hired workers was selected. Such principle of selection ensured wages data representativeness.

In consequence of disintegration of the USSR and occurrence of market relations the economic situation has changed. Notably restrictions on individual labor activity have been removed, the structure of sources of revenue has changed, the number of small state and private enterprises has sharply increased in all economic branches. So, the total number of the small enterprises (SE) in republic has came to 28310 in 2000, 33094 in 2005, 111792 in 2014. From each of them was inexpedient to demand of statistic registration. Full coverage of population has become economically unjustified and almost unrealizable. As a result process stage-by-stage introduction of enterprises sampling in the practical statistics has begun:

1. 1997–2005 Theoretical workings out and pilot sample surveys (retail trade, services, small business);

2. Since 2006 until now. Theoretical workings out and selection of the enterprises on a regular basis (retail trade, small business, labor statistics).

At the first stage of introduction sampling in statistical practice (1997–2005) Statistics research institute provided with methodology and software of branch survey of the enterprises, based on using of group of methods of univariate selection: systematic sampling, random selection without allocation, simple random sample, stratified sample with proportional and optimal allocation. Pilot surveys of SE in retail trade were carried out in 1998–1999, survey of enterprises in services — in 2002, survey of small enterprises in economic branches — in 2003. In 2005–2006 problems of building of multivariate sample are investigated, the first version of the program is developed, trial multivariate samples of SE are spent.

At the second stage (since the end of 2006) researches of multivariate sampling and improvement data extrapolation are hold on. State statistics began to carry out quarter samples of SE in area of labor statistics on a regular basis. Since 2008 has added sample surveys in retail trade and in catering. Special quarter sample surveys of SE concerning employment and unemployment, and also personal subsidiary plots is predicted. Since 2015 micro-entities sample survey spent early.

Despite such advantages of sample, as enough low expenses, efficiency of and high reliability, statisticians was confronted with a number of problems:

Non-responses. The population of micro- and small enterprises is extremely dynamical: the creation of new entities, liquidations, changes in kinds of activity and size of the enterprises are taken place constantly. Sampling frame is based on the data of the previous year of complete survey, and not responded enterprises can be included into the sample (liquidated, changed the kind of activity or not presented the questionnaire).

Atypical units (outliers), i.e. presence in the frame of atypical units, inclusion (or not inclusion) of which in a sample strongly influences the estimates of parameters. Atypical units are the units, which have extreme values of variables, large sample weights, complex structure.

Samples in small domains. Construction samples of small enterprises by economic activity and regions, in some cases is connected with partition of survey population into the small groups and sample fractions become unacceptably high (50–60%). As a consequence, possibilities of control an admissible sampling error are problematic.

Problems of compromise between the accuracy requirement for various groups caused by stratification and restrictions on sample size.

Estimation. The problem of estimation still persists when the univariate stratified sample with admissible standard error and sampling fraction is built. Weights, raising factors allow to estimate precisely enough values of the parameter which was used for sample selection, but other estimates which number can rich 10–30 are of a low quality. In the case of multivariate sample, the error for some group of indicators will be in admissible limits (to 10%), but will be considerably above comparing with the case of univariate sample.

The problems of the software are caused by the complexity of mathematical apparatus of sample survey and the necessity of integration of the sample survey programs in the general system of collection and processing of statistical data.

Specific problems are met designing the *multivariate sample* (stratified by several variables): complexity of a choice of an optimal way of multivariate selection, complexity of a choice of a leading indicator (variable), technical difficulties of construction of multivariate general population (over 500 units), absence of the standard estimation methods.

The problems of non-responses and atypical units may be solved within traditional univariate sample; the solution is connected with the change of general population structure, allocation in separate files of atypical enterprises, use of weighting or replacement procedures. Multivariate sampling and different weighting schemes are used to handle remaining problems, it allows to receive the samples of small size, which are representative for many different parameters.

It is offered by the author to apply a combination of univariate and multivariate sampling methods in order to receive representative small business samples [1–3].

3 Small enterprises sample surveys

Survey objects are artificial persons of small business, i.e. small enterprises. According to the legislation this was the organization with number of the working from 100

persons in industry and transport branches till 25–30 persons in services, nowadays — 100 persons and less. Sampling carried out at each regions and Minsk by branches. The used sample design provides possibility of a choice to use a sampling method depending on population, number and character of survey variables (the program "Multivariate sampling"). It should be done several steps for searching optimal sample size for i-th branch and j-th region: allocation of observed variables, applying multivariate or univariate sample, selection is carried out by the cluster analysis.

Extrapolation of total value of variables on all population is carried out by traditional group raising factors (ratio of number of units in i-cluster of total population and corresponding cluster of sample) and simple errors.

Sampling frame is 20-30% from all number of small enterprises. As to branch sampling fraction is depending on number of SE and the degree of accuracy on a leading variable: a relative sampling error on regions less than 2%, on branches less than 5%, and on small branches less than 8–9%.

4 Micro-entities sampling frames and sample design

Sampling Frames are two files of economic units: 1) micro-entities, represented the state statistical reports on the financial results for basic years (report 1-MP (micro)); 2) set of the private farms. The first file is high — 80 thousands units, sample fraction depends on a character of the initial information, namely: the size of total population, kind of economic activity, region. The second array includes more than 2 thousands farms; it is observed completely (sample fraction is 100%). Predicted non-responses rate for republic is 12–13%, for regions — 6–12%, for Minsk is higher (18–20%). The combination of univariate and multivariate (multidimensional) sample is used.

To receive optimal sample size for i-th kind of activity and j-th region the author together with the colleagues-statisticians have developed the next algorithm:

1. The set of observed variables is allocated (for example, the wages fund, average number of employees, volume of production, revenues, profitability). Average, total values, variability of indicators are calculated.

2. Statistician chooses sampling method: univariate or multivariate. Univariate stratified samples with simple, proportional and optimal allocation are most often used.

3. It should be executed one of three conditions for applying multidimensional sampling: variation coefficient is more than 100%; survey objects are non-uniform on many variables; the small size of total population (top limit — 30–40 units). Otherwise univariate sampling should be used: random selection without allocation, simple random sample, proportional and optimal allocation.

4. It is expediently to use univariate stratified sample, total population is divided by rather homogenous groups. Then different variants of the sample size are ex-

ecuted (minimal is $0.05N$, maximal is $0.08N$). Predicted sample size is allocated by received groups. The choice of an optimal sample size and optimal kind of univariate stratified sample depends on a standard error. So, it (the minimal error) is a main criteria of the determination of sample size.

5. It is expediently to use multidimensional sample, selection is carried out by cluster analysis: total population is partitioned using cluster analysis (agglomerative hierarchical, iterative method of k-means) on homogenous groups to k-variables, i.e. clustering; in each received group the leading (basic) variable is determined and subsequent random selection of units is performed.

Optimal sample population is chosen for each cluster, where standard errors of k-variables are criteria of productivity. If the error exceeds admissible bounds, three methods of its reduction may be applied: a) increasing sample population in cluster; b) additional stratification of the enterprises in cluster to a leading variable; c) repetition of clustering, but with larger number of steps, or using an iterative method with the preliminary number of clusters $r > 1$.

Sample population is formed once in three-four years, i.e. fixed sample (yearly) is used.

5 Statistical weighting

To extrapolate sample data on the total population traditional group raising factors (weights) and standard errors have been used [2,3].

The methodology of weighting for univariate stratified sample is based on the assignment for each enterprise corresponding statistical weight (k_{ijl}):

$$k_{ijl} = N_{ijl}/n_{ijl}, \tag{1}$$

where k_{ijl} is individual weight for each enterprise of l-th group of i-th kind of activity (3 digit for NACE) in j-th region; N_{ijl} is the size of l-th group of i-th kind of activity in j-th region in total population; n_{ijl} is the size of sample group; l is the number of groups by observed variable value ($l = 1, \ldots, m$).

Individual weights are equal within each group of micro-entities, calculated by region, kind of activity, observed indicator (output, employees or others). Individual weights, determined for multidimensional sample, are:

$$k_{ijr} = N_{ijr}/n_{ijr},\ k_{ijrh} = N_{ijrh}/n_{ijrh}, \tag{2}$$

where k_{ijr} is the weight of r-th cluster of enterprises; k_{ijrh} is the weight for h-th group of r-th cluster; r is the number of clusters in i-th branch of j-th region ($r = 1, \ldots, \alpha$); h is the number of groups in r-th cluster ($h = 1, \ldots, \gamma$).

To improve the representativeness by region weighting procedure can be complicated. It is possible to use GREG-estimators and calibration [2–4]. The results of trial calculations testing the first version of methodological and software sampling have shown that the main difficulties are associated with the use of different weighting

schemes, necessary estimation of the whole variables, splitting of the same population on the smaller groups, little subsamples. Sampling fraction — 10–15%. As to branch sampling fraction is depending on the number of enterprises and the degree of accuracy on a leading variable: a relative sampling error on regions less than 2–4%, on branches (kinds of activity) less than 5–6%, and on small branches less than 8–12%.

6 Concluding remarks

The use of combination of univariate and multidimensional samples, different weighting methods will provide very reliable information over larger number of variables: employment, wages fund, revenues and others. However, standard errors, calculated by separate indicators in the context of different kinds of activity at regional level are rather high. To improve the representativeness by region weighting procedure can be complicated by usage of auxiliary calibration estimators. Besides, it is important to take into account the necessity of annual sample updating. The creation of new entities, liquidations, changes in kinds of activity and size of enterprises are taken place constantly.

References

[1] Bokun N., Chernyshova T. (1997). *Methods of sample surveys*. Minsk (in Russian).

[2] Bokun N. (2010). Problems of multidimensional samples in retail trade. *Questions of statistics*. Vol. **3**, pp. 52–60 (in Russian).

[3] Bokun N. (2014). Micro-entities sample survey: problems of design, formation and usage. *Workshop of Baltic-Nordic-Ukrainian network on Survey Statistics, Tallinn, Estonia, August 25-28, 2014*. pp. 25–30.

[4] Sarndal C. E., Swensson B., Wretman J. (2003). *Model assisted survey sampling*. Springer-Verlag.

THE R&D INTENSITY FACTORS OF GDP

Irina Kolesnikova
Belarusian State Economic University
Minsk, BELARUS
e-mail: KLSNK_A@tut.by

Abstract

The object of research is the level and dynamic of the R&D intensity of GDP as a factor of sustainable development and social, economic, scientific, technological and environmental security of the country.
Keywords: R&D intensity, innovative production, effective national innovation system, domestic expenditure on technological innovation, investment climate

1 Introduction

In accordance to the Concept of National Security of the Republic of Belarus (see [1]), scientific and technological safety of the country — a state of national scien-tific and technological and educational potential. It is provided the possibility of implementing the national interests of Belarus in the sphere of science and technology. Technological evolution is the source of fundamentally new threats. It is provided previously inaccessible possible negative impact on the individual, society and the state.

The basic element of the national innovation system is formed in the scientific and technological sphere. Scientific, technological and innovative developments reoriented to the specific needs of economic, social and other spheres, increasing their effectiveness.

Research and development intensity (R&D intensity) of GDP remains low and the share of innovative production in total industrial production is so. An effective national innovation system has not been created as a whole. An innovative infrastructure isn't developed, there is a high level of depreciation of equipment.

Internal sources of threats of the national security in the scientific and technological sphere are:

- R&D intensity of GDP is below the critical level which is necessary for the reproduction of scientific and technological capacity;
- level of the innovation activity and the susceptibility of the Belarusian economy are low;
- the national innovation system (including legislation, infrastructure, technology transfer from science to manufacturing, material and technical base of scientific institutions, the financing system, the industry science) is ineffective-ness.

The annual increase in GDP and knowledge-intensive approach will increase innovative activity and the susceptibility of the Belarusian economy will streng-then the industry and science.

An effective system of incentives is established for the development of high-tech industries and of cross-flow mechanism of financial, human and material resources from declining sectors of the economy in the long-term, comprehensive computerization of the economy and society. It is ensured the formation of a qualitatively new technological order in the Belarus, the expansion of exports of high technology products, the attract foreign investment and integration of national innovation system in the global innovation system in the world.

2 Main results

Indicators of innovative capacity and the activity can identify the strengths and weaknesses of the innovative development of the country and its regions, to find the barriers to innovation. The analysis of indicators can be used as management tools, economic systems at different levels.

The level of R&D intensity of GDP is the general indicator which characterizes the impact of scientific and innovative activity.

R&D intensity of GDP (on technological innovation) is calculated as the ratio of the size of domestic expenditure on technological innovation in terms of GDP (CTI/GDP).

Multiplicative model was investigated. It identifies the factors which was influenced the research intensity. It characterizes the relationship between the size of R&D intensity of GDP (CTI/GDP) and science intensity of shipped innovative products (CTI/IPS), share of shipped innovative products in shipped products (IPS/PS) and share of products shipped to the GDP (PS/GDP):

$$\frac{\text{CTI}}{\text{GDP}} = \frac{\text{CTI}}{\text{IPS}} \cdot \frac{\text{IPS}}{\text{PS}} \cdot \frac{\text{PS}}{\text{GDP}}, \tag{1}$$

where CTI is the cost of technological innovation, GDP is a gross domestic product, PS is the products shipped, IPS is the innovative products shipped.

Factors for solving the model are presented in Table 1 (source: own elaboration based on data value from [2]).

Indicators, %	2005	2014	Growth rate, %
CTI/GPD	3.63	1.69	46.56
CTI/IPS	33.73	12.94	38.36
IPS/PS	15.20	17.85	117.43
PS/GDP	70.79	72.95	103.05

Table 1: Dynamics of GDP R&D intensity factors in Belarus in 2005-2014 years.

As can be seen from Table 1, the level of research intensity of GDP by technological innovation has decreased by more than 2 times, from 3.63% to 1.69% for the period from 2005 to 2014. In addition, the level of research intensity of shipped innovative products has decreased from 33.73% to 12.94%.

Reducing R&D intensity of GDP due to the fact that the rate of growth of domestic spending levels on technological innovation, equal to 4.5 times, significantly lags behind GDP growth and IPR, which are, respectively, 9.8 and 11.8 times.

At the same time there are significant differences in the level of research in-tensity of gross regional product by regions. In 2014, the highest in comparison with the national level of research intensity was recorded in Mogilev (5.7%), Vitebsk (3.8%) and Gomel regions (3.2%). The rest of the research intensity level is lower than the national, and the smallest — in the Grodno region (0.45%) [9, p.15].

The contribution of each factor in the model change in research intensity of GDP in 2005–2014 is presented in Table 2.

Factors	Change in the research intensity of GDP by each factor, %	The share of production growth at the expense of each factor, %
High technology of shipped innovative products	-2.71	-139.0
Share of shipped innovative products in the products shipped	0.65	33.3
Share of products shipped in GDP	0.11	5.7
Total change in research intensity of GDP	-1.95	-100.0

Table 2: Contribution of factors in the change of R&D intensity of GDP in the years 2005–2014.

As can be seen from the table, model factors have different effects on the direction of change in the research intensity of the GDP.

Reduced-tech innovative products shipped by 20.8 percentage points had a strong negative impact, the magnitude of this impact was 139% with a minus sign. However, two other factors have had a positive effect, reducing the decrease in R&D intensity.

Namely, due to the growth level of the most important indicators of the im-pact of scientific innovation, as the share of shipped innovative products shipped products (up 2.6 pp) research intensity of GDP has increased by 0.11 percentage points, which is 33.3% of its total change; and by increasing the share of prod-ucts shipped in the GDP (2.1 percentage points) and an increase research intensity of GDP by 0.11 percentage points, which corresponds to 5.7% of its total change.

3 Conclusion

The innovative new-technological structure, its formation and growth will deter-mine the economic dynamics in the coming years. It is characterized by increasing instability, causing the need to transform the economy. It is occurs in an envi-ronment where information sphere is transformed into a system factor of society.

To move businesses to a higher technological level of innovative activity, it is necessary to make a quantitative leap in the volume of unit costs for technological innovations. It will be contribute to the competitiveness of products and organizations in the country.

According to the national strategy for sustainable socio-economic development in 2030 is expected to increase the share of domestic spending on research and development in GDP (R&D intensity of GDP) to 2-2.5%. Index the share of innovative products in the total volume of industrial production is one of the criteria for effective use of scientific and technological capacity. It is forecasted at 28-30%.

It is necessary to ensure adequate funding of technological innovations carried out at the enterprises upon reaching the competitive scientific groundwork. Currently, equity in the enterprises in most cases is not enough.

Improving the investment climate in Belarus, required for the perception of innovation and creation of high-tech jobs is a necessary step in ensuring the at-traction of budgetary funds and investors to support and implement innovative projects.

References

[1] National Strategy for Sustainable Socio-Economic Development of the Republic of Belarus for the period up to 2030 [electronic resource] / Access:// www.economy.gov.by/nfiles/001708_663161_Proekt_21_11.docx

[2] National Statistical Committee of the Republic of Belarus (2015). *Science and Innovative Activities 2015: stat. compilation*. Minsk.

STATISTICAL ASSESSMENT OF GENDER ISSUES IN SOCIAL AND LABOR SPHERE

Alla Kulak[1], Yauheniya Sharilova[2]
Belarus State Economic University
Minsk, BELARUS
e-mail: [1]alla_kulak@mail.ru, [2]sharilovaee@mail.ru

Abstract

The article presents the evaluation and statistical analysis of the gender-sensitive development of the social and labor sphere of the Republic of Belarus.
Keywords: gender, labor market, gender segregation, disaggregated statistics, equality

1 Introduction

Belarus commitment to and targeting at sustainable social and economic development dictates the buildup of human potential impossible without gender equality in society and economy.

Today, the importance of gender statistics reflecting the situation of men and women has become ever more evident. Its improvement is of great importance, as in the same socio-economic environment men and women have different needs and opportunities and face different problems. The possibility and reliability of solving these problems, development of an efficient national socio-economic policy (both at the national and regional levels), public information, key managerial decision making, elimination of existing gender stereotypes all depend on the scope and adequacy of statistical data.

2 The main indicators of gender statistics in the Republic of Belarus

Official gender disaggregated statistics in the Republic of Belarus is developed on national and subnational levels in the following areas: analysis of natural population movement and migration; education; labor and employment; health; crime study; public administration; family relations; child protection.

Gender disaggregated data are an efficient and indispensable tool for research into the causes and assessment of gender inequality in the country, reflection of existing gender asymmetry, analysis of possible effects of gender problems, development and introduction of necessary changes to the existing social and economic policy, etc.

Lack of statistical information on obstacles and difficulties faced by men and women in Belarus is one of the reasons for inadequate study of many gender issues. First and foremost, this is true for gender differentiation in wages, access to key resources, division of household work and professional segregation. This indicates the need to extend the national system of gender statistics indicators.

In the Republic of Belarus, gender disaggregated data are generally collected and processed by statistical authorities (based on results of population censuses, sample surveys, current records, special surveys, etc.) and state administration bodies authorized to maintain statistics on issues within their competence or competence of their subordinate organizations (for example, the Ministry of Health of the Republic of Belarus).

3 Statistical evaluation of the gender imbalance of social development of the Republic of Belarus

More or less, gender segregation is present in the labor market of any coun-try. Accurate assessment of this phenomenon requires a clear picture of its dynamics and processes in the labor market and in society generally that contribute to increasing or decreasing gender asymmetry in distribution of workers by sectors and occupations.

In this sphere gender-specific indicators include economically active population, female/male proportions by personnel categories, occupational groups, economic sectors, education level, age groups, and gross average monthly wage for separate industries.

Also, gender disaggregated data on registered unemployment (by unemployment time, education level, age, causes of dismissal, latest employment) are systematically reflected, as well as age- and sector-specific statistics on proportion of employed in hazardous industries and victims of industrial injuries. However, age-specific differentiation in wages depending on education level is not monitored.

Of total employed population at the beginning of 2015, women accounted for 49.6% (at the beginning of 1999 — 52.0%). As a result of reforms in the national economy, the number of public sector workers decreased from 57% in 1999 to 39% in 2015.

In recent years, proportion of women in traditionally "female" sectors has been increasing: in education — from 78.0% at the beginning of 1999 to 78.8% at the beginning of 2015, in culture and arts — from 70.7% to 70.9%. Proportion of women continues to be high in health care, physical culture and social welfare organizations (83.0% at the beginning of 1999 and 82.4% at the beginning of 2015), trade and public catering (75.9% and 66.5%).

At the same time, despite an insignificant reduction of proportion males continue to predominate in the forest and transport sector (more than 70%). In the largest sector — production — proportion of males increased from 50.8% at the beginning of 1999 to 58.6% at the beginning of 2015. A similar increase is observed in construction (from 76.6% to 84.0%).

In terms of personnel categories, as before men predominate in the total number of blue-collar workers (more than 56%), and women - among white-collar workers (about 70%). In recent years, the number of women in executive positions has increased from 44.2% of the total number of white-collar workers at the beginning of 1999 to 47% at the beginning of 2009.

Education level of women is generally higher than that of men, the situation being the same over a long period of time. About 60% of working women have tertiary and

secondary special education (about 40% of working men) (in 1999, 46.7% and 31.3%, respectively).

It's interesting to know, that in tertiary education there is a substantial gender asymmetry. In accordance with the International Standard Classification of Education, tertiary education includes specialized secondary education.

Women are more inclined to obtain complete general education and more often wish to obtain higher professional qualifications. Over the past ten years, tertiary education in the Republic of Belarus has become more feminized: at the beginning of 2015 females accounted more than 57% of total student population as against 56.24% in 1999.

High education level is a competitive advantage in the labor market; however, women face objective difficulties due to their greater involvement in child care.

In the Republic of Belarus, the right to equal remuneration for equal work for women and men is envisaged in the current Constitution of the Republic of Belarus (Article 42) and Labor Code. However, wages of females are on average less than wages of males. Gender-based difference in wages is usually explained by unequal distribution of males and females by occupations and sectors (horizontal segregation), inequality in wages within occupations and types of activity (vertical segregation), and the fact that women usually hold posts with less remuneration. Wage disbalance to the disfavor of women is in fact typical for all sectors. Thus, in December 2015, the ratio of female to male average accrued wages was more than 70%.

The gender wage gap is largely due to gender specifics of employment. More women than men are traditionally employed in economic sectors such as trade and public catering, education, health care, physical culture and social welfare, culture and arts, where wages are 7-30% less than average in the Republic, while in sectors such as industry, construction, transport where wages are quite high there are much more men than women.

It should also be mentioned that wages in the private sector are usually considerably higher than in the public sector (in December 2009 — by 12.1%), while a considerable number of women are employed in education and child care where state-run organizations prevail.

4 Conclusion

Analysis of gender issues in social and labor sphere makes it possible to identify measures that would help mitigate trends outlined above and identify key areas of socio-economic policy that would reduce gender gaps in some indicators under review.

APPLICATION OF METHOD LAGS MODEL FOR THE ANALYSIS OF THE IMPACT ECOLOGY ON HEALTH

Oksana Matkovskaya
Belarus State Economic University
Minsk, BELARUS

Abstract

In this direction there is identification of problems the effects of long-term exposure of air pollutants on human health.
Keywords: multivariate statistical analysis, component analysis, ecological indicators, correlation analysis lag, Almon method

The relevance of the statistical analysis of the environmental impact on the demographic processes in the context of the territories of Belarus is due, on the one hand, the increase of anthropogenic load on the air, and all the more apparent relationship of demographic and ecological processes, and on the other — insufficient development of the mutual influence of ecological and demographic processes. The present study is intended to fill the gaps in quantitative analysis in this area, to create a system of indicators and models to describe and analyze the environmental characteristics in relation to the demographic development of certain areas. Our study is a response to the increased attention of all countries with developed market economies to environmental issues, including a debate on the adoption of the Kyoto Protocol, as declared by the international community of the Millennium goals for sustainable development. The selected object - the state of the atmospheric air is vital to ensure the normal life of the people, for the demographic reproduction, achieving the goals of economic and social development.

The impact of environmental factors on human activity is a complex and multifaceted socio-economic process, which covers all aspects of society. This raises the problem of assessing the changes in the environment and public health. For the said task, you must use the aggregate, to absorb all information necessary to analyze the influence of the environment on human health. Moreover, in practice there are different variants of general indicators, making it difficult to achieve the goal.

To construct the integral index (regardless of the method of its calculation) you need to define an initial set of features and the degree of influence of each of them on the outcome. Included in the comprehensive indicator parameters and their weights should be chosen so that the composite indicator best reflects the true picture of changes in the environment and public health. Two different approaches to the solution of the issue can be used.

The first way — is to use expert estimates. In this case it is necessary to collect a large enough group of experts, each of whom will have to rank the proposed indicators on him the degree of public health impact, based on his personal judgment. The disadvantage of this method is the subjectivity of expert evaluations, so that the results

do not reflect the actual situation and its representation, understanding of specific people.

The second possible way to solve this problem — is the use of multivariate statistical analysis. These methods allow us to determine the hidden, implicit laws objectively existing in the study of social and economic processes, but not directly measurable. The most promising in economic studies are a factor or component analysis. It enables the reduction of extensive numerical material to several independent and simple factors.

The choice of system baselines largely determines the results of the analysis and therefore is a very responsible stage of the study. In our study we propose to distinguish four groups of indicators that describe the basic directions of changes in environmental conditions impact on population health. The study examined the effect of 10 characteristics of the environment on human health, and in the initial set of indicators included in the Republic of Belarus for 15 years, from 2000 to 2014.

As a result of two factors were obtained:

the first factor (F_1) — indicator of anthropogenic load on the environment;

the second (F_2) — integral index of resources and development of society.

In the future, we assessed the relationship of the processes on the basis of the time series (F_1 and F_2). Evaluation was performed using correlation analysis lag (Almon method).

As a result it was found the presence of certain speakers depending on the values of the factor F_2 (building society development) on the dynamic changes of anthropogenic load level on the environment (F_1).

Selecting the maximum lag length and degree of the polynomial carried out empirically. The proposed 3-year lag polynomial first degree possible to obtain cost-interpretable model parameters, namely, the influence of one orientation of the time factor on the exogenous variable. Model relationships of society's development potential (F_2) from the change of anthropogenic load on the environment (F_1) is as follows:

$$F_2^t = -1.15F_1^t - 0.475F_2^{t-1} - 0.2F_2^{t-2} - 0.88F_2^{t-3},$$
$$R^2 = 0.735,\ n = 12,\ F_{\text{calculated}}(4,8) = 5.548 > F_{\text{tabular}} = 3.838.$$

Relative regression coefficients in the model are:

$$\beta_0 = \frac{1.15}{2.703} = 0.425 \text{ or } 42.5\%$$
$$\beta_1 = \frac{0.475}{2.703} = 0.176 \text{ or } 17.6\%$$
$$\beta_2 = \frac{0.2}{2.703} = 0.074 \text{ or } 7.4\%$$
$$\beta_3 = \frac{0.88}{2.703} = 0.324 \text{ or } 32.4\%.$$

Thus, 42.5% (F_1) general decline in society's development potential (F_2), due to the increased anthropogenic load on the environment, there is in the current time; 17.6% — at time $t + 1$; 7.4% — at the moment $t + 2$ and 32.4% of this decrease occurs at time $t + 3$.

Average lag in this model is:

$$\bar{l} = \sum_j j\beta_j = 0 \cdot 0.425 + 1 \cdot 0.176 + 2 \cdot 0.074 + 3 \cdot 0.324 = 1.296.$$

On average, most of the effect of anthropogenic load growth on the natural environment manifests itself almost immediately (more precisely through 1,296 years) on reducing the development of society. As a result of constructing distributed lag model it was established and demonstrated statistically significant presence of feedback between the factors F_1 and F_2. It should be noted that 43% of the overall society development potential (F_2) in the current period due to increased anthropogenic load on the environment (F_1) in the same period. However, increased anthropogenic load on the environment in the current period and has a deterrent nature of the impact, causing 32,4% of the overall society development potential reduction after only three years. Given the dynamic relationship established as a result of the lag analysis, short-term positive changes will not have a significant positive impact on the stabilization and even more to improve the health of children conditional aged 0-14 years. In this case, the stabilization of (maintaining the current level) of healthy children aged 0–14 years is a long-term nature, as shaped by the degree of anthropogenic load both current and previous periods.

Achieving sustainable improvement of healthy children aged 0–14 years, perhaps through an annual sustained reduction of anthropogenic load on the environment by increasing the level of dust and gas cleaning equipment manufacturing equipment; activation and diffusion of energy- and resource-saving technologies; ensure environmentally optimal spatial planning in the implementation of economic activities which are harmful to the environment.

References

[1] Republics scientific-practical health center (2010). *Health and the Environment (2010). Sat. scientific. tr.*. Minsk.

[2] Matkovskaya O. G. (2009). Technique of construction of integrated indicators of air condition. *Statistics of Ukraine*. Vol. **2**(45), pp. 12–17.

[3] Soshnikova L.A. (1999). *Multivariate statistical analysis in economics*. M.: UNITY-DANA.

[4] National Statistical Committee of the Republic of Belarus (2015). *Protection of the environment in Belarus (2015): stat. sb.*. Minsk.

[5] National Statistical Committee of the Republic of Belarus (2015). *Statistical Yearbook (2015): stat. sb.*. Minsk.

THE RELEVANT LEADING INDICATOR OF MACROECONOMIC DYNAMICS

M. M. Novikov
Belarusian State Economic University
Minsk, BELARUS
e-mail: mm_novikov@rambler.ru

Abstract

The dynamic cross-correlation-regressive model of gross internal product is worked out on explaining to the variables consumer and investment functions of gross internal profit. It is set that a gross internal receipt is a passing ahead index, laying the "waterway" of subsequent motion in time of physical volume of GDP with the two-year passing, that allowed to get the expected values of gross internal product with horizon of prediction on three years.

1 Introduction

Conditions of foreign economic activity are characterized by the mobility level and the prices of exports and imports. Exporters, given the pricing situation on the world markets, are trying to direct their export and import activities in the riverbed advanced dynamics of export prices in comparison with prices of imports of goods and services. If export prices rise faster than prices to import, it shows the improvement of foreign trade conditions. In these circumstances, to pay a specified amount of imports requires a smaller volume of export sales. A similar situation exists in conditions when the prices for exports decline more slowly compared with the decrease in import prices.

2 Indicator description

The advanced dynamics of prices for exports of goods and services has a direct impact on the balance of foreign trade and participates in the formation of gross domestic product (GDP) by the final use method. While studying the dynamics of physical volume of GDP, its components are converted into constant prices by deflation (divide) to the relevant indexes-deflators. As a result, the deflation thereby benefit from advanced dynamics of export prices becomes negligible, while producers of goods and services in such economic policies get increasing the benefits from its export-import activities.

The author constructed a macroeconomic indicator, sensitive to changes in conditions on external markets. Multidirectional dynamics of prices on exports and imports of goods and services reflected in the index of physical volume and dynamics of gross domestic income (GNI). Gross domestic income as an analytical indicator contained in the system of national accounts 2008 [3, p. 371-373). However he is not developed in statistical practice of the Republic of Belarus .Meanwhile, in an open economy of the Republic of Belarus its development becomes a high priority. GDP and GNI at current market prices differ. The difference between them is detected due to the fact

that the export-import balance as a component of physical volume indices of GNP and GDP in different ways converted to constant prices. Because of this differ the only real indicators physical volume of GDP and GNI.

In GDP balance indicator export and import activities is recalculated in the prices of the base period using the two deflators. Exports at current prices deflated in the export price index and import used when deflating the price index for imports of goods and services. Thus, the export-import balance as a component of GDP physical volume becomes meaningful expression of pure export-import product. Balance component of gross domestic income is filled with economic content in the real volume of export and import in net income [1, p. 52]. Calculation algorithm of this indicator is presented in a single operation deflation: current export-import balance deflate to the standardized deflator. In terms of advanced dynamics of export prices as the deflator standardized should recognize the composite price index for imports. In the opposite case, the standardized deflator will be the composite index of export prices. The difference between the real net export-import revenue and net export-import product - volume profit (losses) from changes in the terms of export and import activities. Real gross domestic income is represented by the sum of GDP in the estimate by the method of end use and the amount of profit (loss) from changes in the terms of export and import activities.

3 Conclusion

As a result of a comparative assessment of the trajectories of the physical volume of GDP and GNI for the economy of the Republic of Belarus in 2000-2013 was found that the dynamics of real gross domestic income has outpaced the dynamics of gross domestic product. The author developed a statistically significant autoregressive model gross domestic income, then its explanatory variables the resulting regression equation of the gross domestic product has predictive power for two time periods. Based on its evaluations, the following conclusions are formulated.

First, it is analytically proven that in the dynamics of the gross domestic product does not reflect the change of conditions of export-import activities.

Secondly, the change of conditions of export-import activities measured by the indicators of volume of profit (loss). Thus the profit is formed under the influence of strong performance in export prices of goods and services in comparison with dynamics of the prices for import purchases. Otherwise, the image of losses.

Third, to change the terms of export and import activity is sensitive gross domestic income (GNI). Common components of GDP and GNI are final consumption expenditure and gross. The index of physical volume of WSC additionally describes characteristics of the volume of profits (losses) from changes in the terms of export and import activities.

Fourth, at the macroeconomic level, the generated indicators of volume and dynamics of: a) gross domestic product b) gross domestic income of the Republic of Belarus in 2000-2013 and methodological basis for the development of autoregressive models of the 2nd and 3rd orders [2, p. 70-95] the regularities of their behavior in time.

The close interaction between the dynamics of GDP and GNI.The indicator of gross domestic income is a leading indicator that determines the subsequent change of the physical volume of GDP lagged two years ahead of schedule. It is established that the lead lag effect gross domestic income is determined by the duration of the period of redistribution of profit from improved terms of export-import activities on consumption and accumulation.

Property advanced dynamics of gross domestic income compared to dynamics of GDP used to obtain expected values of GDP with the horizon of the prediction three years. The predictive power of recommended models tested on actual data 2014. On the predicted estimates of the likely growth of GDP in 2014 compared with the previous year represented a growth rate equal to 0.7% when the actual rate of increase of 1.6%. By the author's estimation in 2015 is expected to reduce the physical volume of GDP by 0.56 percent, and in 2016 - 1.5 percent. Gross domestic income, being expressed in constant prices that represents the initial information base for the development of indicators of the dynamics of real gross national income and gross disposable income. As shown in the SNA 2008, the real gross national income is equal to the sum real gross domestic income and net primary income from abroad in real terms. In turn, real gross national disposable income is formed as the algebraic sum of real gross national income and net current transfers from abroad [2, p. 372]. In this regard, formulated and solved the question of the choice of the deflator index external traffic streams of income from abroad and abroad.

According to accounts of external transactions with the rest of the world external flows of primary incomes, current and capital transfers, foreign lending and borrowing in the pure additive measurement in the form of a functionally balanced with a balance indicator the external transactions of goods and services. Hence, a definite conclusion as to the choice of the deflator index external traffic streams of income from abroad and abroad. Assessment of balance indicators of primary incomes and current transfers from abroad in constant prices as components of real gross national income and gross disposable income can be accomplished by deflation of the respective external income streams in a pure measurement on the same standardized deflator, which was used in the assessment of profit (loss) from changes in the terms of export and import activities.

Thus, under changing conditions of foreign economic activity after real gross domestic income real development of indicators of dynamics of gross national income and gross disposable income for the study of macroeconomic dynamics becomes equally relevant.

References

[1] Novikov M.M. (2012). The Indicators of net export product and net export income. *Accounting and analysis*. Vol. **3**, pp. 49–53. (in Russian)

[2] Novikov M.M. (2005). Autoregressive Model of gross domestic product. *Statistics: indicators and methods of analysis*. pp. 70–95. (in Russian)

[3] European Commission et al. (2012). *The system of national accounts 2008*. NY.

THE STATISTICAL VALIDITY OF THE INCREASE IN RETIREMENT AGE IN THE REPUBLIC OF BELARUS

Yauheniya Sharilova[1], Alla Kulak[2]
Belarus State Economic University
Minsk, BELARUS
e-mail: [1]sharilovaee@mail.ru, [2]alla_kulak@mail.ru

Abstract

There is a statistical justification of the need of increase the retirement age in the Republic of Belarus. In particular, through the use of exponential smoothing procedure, was found a trend toward life expectancy improvement in Belarus.
Keywords: pension reform, life expectancy at birth, prediction

1 Introduction

In determining the retirement age take into account life expectancy and health status, the economic situation in the country, the needs of society in the labor force, the ability of the pension system perform its mandated functions. In general, low bound of the retirement age is the precise point beyond which the performance of the average worker is lowered so that the preservation of previous earnings is economically impractical or require excessive effort and health costs. Of course, much of this depends on the age of the changes in the human body caused by the aging process of an individual. Features of aging individuals create obstacles in establishing the average age of retirement (later retirement age), as researchers face a very heterogeneous set of statistics. Thus, one group of persons qualitative changes in the condition of the body associated with aging, already observed in the 45, and the other — only 70 years.

2 Factors determining the increase in the retirement age in the Republic of Belarus

In the Republic of Belarus adopted by the Presidential Decree 137 "On improvement of pension" April 11, 2016. Until the end of 2016, the country will remain "Soviet" threshold of retirement: 55 years for women and 60 for men. From 1 January 2017 the retirement age will increase annually for 6 months before reaching the age of men 63, women — 58 years (see [1]).

The following facts should be noted as a prerequisite to increase the retirement age:

1) in Belarus since the 1930s. preserved rather low retirement age in comparison with many countries in the world. Lower values are recorded only in the poorest countries of Africa and the Pacific (Swaziland, Nigeria, Kiribati — 50 years for men and women), the population of which is characterized by low levels of life expectancy at birth;

2) from 1990 to 2015, the ratio of the number of employed persons and the retirement age, so significant for the PAYG pension system of the country, one of the principles which the operation is a "means of distribution of able-bodied citizens to the disabled, from working to non-working" was reduced from 2.56 to 1.93;

3) for the 1993-2012 biennium percentage ratio of total public spending on pensions and the gross domestic product increased from 5.8 to 8.4%;

4) part of the retirement age workers is continue working. Thus, in the period between the censuses of population of Belarus 1999 and 2009 the level of employment of the population in the age older than able to work rose from 9.1 to 12.9%, while the share of employed persons of retirement age in the total employment — rose from 4.4 to 6%. According to census of population 2009, 12.7% of women of retirement age are employed in the economy (according to the 1999 census — 7.7%).

3 Obstacles to increase the retirement age in the Republic of Belarus

In Belarus there are "contra" to raise the retirement age (increase in morbidity and disability; possible increase in unemployment due to the significant and sudden increase in the number of people of working age, the need for employment, "the former senior citizens", their training and retraining, qualification and educational level and etc.).

Intercountry differentiation of retirement age values retired due to including the differences in the levels of life expectancy at birth. In 2014, the value of this index ranged from 45.6 years (Sierra Leone) to 83.6 years (Japan). In Belarus, the value of life expectancy at birth is almost corresponds to the world average. For the 1958-2014 biennium the value of this indicator in the country has increased by 2.9 years, and in 2014 was 73.2 years. Although a certain gap from the levels of life expectancy in some countries (Switzerland — 82.6 years, Italy — 82.4 years, Canada — 81.5 years, Germany — 80.7 years, and others.), 2002 (point minimum - 68 years) in the Republic of Belarus has been a fairly stable positive dynamics of this value. So, for the 2002-2014 biennium. it increased to 5.2 years in an absolute, or 7.6% in relative terms.

There is an of interest in calculation promising levels of life expectancy at birth in Belarus. For these purposes it can be used a model based on the use of exponential smoothing procedure. As a general rule, using the method of least squares on the first point of the time series are estimated values of linear trend model parameters to zero time (see [2]).

As a starting point of the calculations are the parameters of a linear trend, which was built according to the in 2002-2006. In accordance with the results forecast for the 2015-2017 biennium. in Belarus by the end of the period of anticipation value of life expectancy at birth will reach 74.4 years, which should be regarded as a positive development in the field of health and longevity of the population. The average relative error of approximation was 4.5%, which indicates the high quality of the model.

4 Conclusion

The aging of the population — an objective, natural and long-term process that is associated with the development of society. Therefore, as far as strengthening the position of the Belarusian economy and further improve people's living standards is quite expect western script development process of demographic aging, based on the increase in life expectancy at older ages. The proof of this assumption is the outlined Belarus increase in life expectancy, which is in line with the forecast for the future will continue. Therefore, using the advocacy role of such statistical indicators as the ratio of old-age dependency ratio, the average age of the population, the ratio of number of employees and retirees, the percentage of spending on pensions and GDP and others, should always inform the public about the necessity and inevitability of the measure. In the information society to achieve this goal it is advisable to use traditional and electronic media, the Internet.

References

[1] On the improvement of the pension system. Presidential Decree 137 of April 11, 2016. http://president.gov.by/ru/official_documents_ru/view/ukaz-137-ot-11-aprelja-2016-g-13449/

[2] Lukashin Y. (2003). *Adaptive methods of short-term time series forecasting.*. Finance and Statistics, Moscow.

METHODOLOGICAL APPROACHES TO THE REFLECTION OF ENVIRONMENTAL ASSETS IN SEEA[1] AND NAMEA[2]

L. A. Soshnikova
Belarus State Economic University
Minsk, BELARUS
e-mail: ludmila_sosh@mail.ru

Abstract

The article deals with possible approaches to the construction of environmental-economic accounting of the national system using a variety of methodological approaches. We consider two systems of environmental-economic accounting: SEEA and NAMEA. Comparative characteristic is given.
Keywords: environmental-economic accounting

1 Introduction

In modern conditions of managing the exploitation of natural resources should be carried out within the framework of long-term preservation of environment concepts for the needs of the person. Based on this reflection process involving natural resources in economic activity also must undergo a very significant change.

The traditional system of national accounts (SNA) does not have the necessary methodological tools and analytical capabilities for the valuation of the total volume of consumption and stocks of natural assets, environmental protection industries and sectors of the economy. This article describes the basic principles of the national environmental accounting, the advantages and disadvantages of existing systems and proposed approaches to the integration of environmental factors in the statistics. Before you begin to develop a national system of environmental-economic accounting, it is necessary, in our view a detailed analysis of options offered by international organizations, to assess their strengths, weaknesses, labor and cost of implementation. You should start with the simplest options that do not require any additional large-scale statistical observations.

2 Approaches to Environmental-Economic Accounting

Between 1970 and 1995 a number of approaches to the accounting involved natural resources and to the evaluation of environmental damage have been developed. In particular it has been proposed two approaches to the construction of environmental accounts to account for the natural assets on the one hand payment of so-called “green

[1]The System of Environmental-Economic Accounting
[2]National Accounting Matrix including Environmental Accounts

national income" and on the other hand the construction of "physical accounts" for certain types of natural resources.

Dutch economist Roofie Hurting was probably the first to propose the measure "sustainable national income" (SNI), which should be fully taken into account the consumption of natural resources by deducting from the gross national income of the amount of damage to the environment. Proponents of "green national income" offered to give each of its natural resource valuation, monetization damage. Dutch economist Stephen King and his colleague de Haan of Statistics Netherlands (CBS) offered to link economic performance to damage to the environment, but measured in physical units, that is, to develop a hybrid system of accounting, which will be connected to the cost parameters and physical quantities.

Consequently, the development of national statistical methodology in the direction of its greening can go two ways. You can develop a hybrid system of environmental-economic accounting, which along with the traditional accounts of the SNA will be built satellite accounts for each type of natural resources in physical terms. This option is available in the Central bases SEEA, adopted in 2012, the UN Statistical Commission as an international statistical standard [1]. The basis for the development of natural assets in the SEEA are accounts SNA non-financial assets, which also include non-produced and natural assets. SEEA provides the above SNA account in part in aggregate form and partly in a disaggregated form. Disaggregation helps identify environmental protection measures to prevent or attenuate the deterioration of environmental quality or reducing the damage caused by environmental degradation.

As for non-financial assets, it is proposed to further disaggregate data on stocks and changes in the volume of natural assets to improve and extend the registration of consumption of natural resources in the production process, taking into account changes in the value of natural assets under the influence of production and consumption.

On the accounting principles damage builds another hybrid accounting system — National accounting matrix including environmental accounts (NANEA), developed by Dutch researchers [2]. It uses economic indicators, measured in monetary units, and linked to their environmental indicators presented in physical units. According to the developers of the system, to get a clear understanding of the relationship between the natural environment and the economy, it is necessary to use a physical representation of environmental resources in order to avoid the problems of valuation and revaluation.

The basic idea is to expand NANEA traditional national accounting SNA due to two additional accounts. The developers of this system offer to keep records on key environmental areas: the greenhouse effect; depletion of the ozone layer of the earth; Soil oxidation; waste, discharge of polluted wastewater and others.

Second additional expense to the environment by such substances as carbon dioxide or sulfur dioxide, in which these substances should be expressed in physical quantities (kilotons, tons, etc.). We can say that NAMEA create the summary indicators for the environmental issues, which are considered the most relevant at the international level. This system is based on a set of tables that provide an overview of relevant relationships between accounts and data streams on environmental change [2]. The indicators in this table characterize the contribution of each activity in the economic

performance and environmental burden as a percentage. The tables gives an idea about the total rejection of pollution per unit of final demand for specific activities in relation to the average for all industries. Development of this type of table is the complexity of data sources, according to the method of calculation of indicators for individual environmental themes (destruction of the ozone layer, the greenhouse effect, eutrophication).

3 Comparative characteristics of the two systems of integrated environmental and economic accounting

In fact, NAME has much in common with the SEEA. Both systems are similar to the format used by the accounting matrices. However, there are some differences:

1. The SEEA focuses on expanding the standard accounts of assets due account of environmental assets such as water, air and others. In contrast, NAMEA starts with expansion to a full national accounting system to account polluting substances and environmental topics.

2. NAMEA does not involve calculation of environmentally adjusted 'green' GDP, as it makes SEEA.

3. NAMEA links pollutants c environmental themes (for example, the destruction of the ozone layer), and the SEEA system does not contain such an aggregation.

4. NAME system can be used for analytical applications based on Leontief model. It can be used to determine the amount of pollution induced by one unit of final demand for each type of activity. This type of accounting is unfortunately not provided for in the SEEA system.

5. SEEA methodology allows for the degradation of natural resources as consumption of fixed capital in the traditional SNA. This is not provided for in the NAMEA system.

It can be concluded that NAMEA is a multipurpose information system that is able to generate information for the public and governments about the status of environmental assets and environmental pollution. The choice of environmental problems depends on political decisions, rather than on the decisions of scientists. This is the reason why the NAMEA of different countries are different[3]. Without a doubt, it would be useful to standardize sets of pollutants and the list of environmental topics for all countries,

[3]British NAMEA contains 15 environmental substances and only 3 environmental issues (Vaze 1999), Japan has 16 agents and 6 environmental themes (IKE 1999), the German has 8 pollutants and 2 environmental issues (Tjahjadi, Schaefer, Radermacher & Hoh 1999) and Swedish NAMEA are 5 pollutants (Hellsten, Ribacke & Wickbom 1999).

since environmental problems are global in nature. NAMEA is a tool for the integration of environmental concerns and combines environmental data with the economic data of the main SNA accounts. There is no specific economic constraints to select a specific nationally adapted version NAMEA. Developers are free to decide which environmental themes and some substances that may pollute the environment should be controlled to solve environmental problems. In addition NAMEA provide data in the required format for all kinds of in-depth environmental and economic analysis.

4 Findings

Any of the above systems can begin to put into practice the work of statistical bodies of the Republic of Belarus, but first need to select priorities in environmental-economic accounting, identify the most important environmental issues, the integration of all available information on the state of the environment, stocks and consumption of natural assets within unified statistical methodology aimed at the greening of macroeconomic indicators and assessment of sustainable development. It is possible to start with the development of water resources and forests accounts in physical units on the SEEA methodology, parallel to the table to develop a relationship of economic performance and contamination by NAMEA sample.

References

[1] United Nations et al. (2014). *System of Environmental-Economic Accounting 2012 – Central Framework*. New York.

[2] Stauvermann P., van der Veen A. (1999). National accounting matrix including environmental accounts (NAMEA).
`http://www.ivm.vu.nl/en/Images/AT4_tcm234-161575.pdf`

MODELING COMPETITIVE ADVANTAGE OF TERRITORIES

Stanislav Visotski
Belarus State Economic University
Minsk, BELARUS
e-mail: visozkij@yandex.ru

Abstract

The article gives a theoretical and economic substantiation category of "regional competitive advantages". The essence and the novelty lies in the positioning of the author's determination not only as a definition of the theory of competition, but also as an object of statistics. The proposed interpretation of the category within the statistical science to allow formal assessment of the impact nationwide, industry and regional incentives for growth of industrial activities on the dynamics of the major indicators of the territories.
Keywords: regional competitive advantages, industrial activities, indicators of the territories

1 Introduction

Scarcity of resources in the regional industrial complex provokes the infusion of the republican and local budgets, which induces dependency in industrial business. Such a path of economic development should be recognized as a dead end because of its inefficiency.

Today more than ever, it became clear that the modernization of industrial enterprises is not just the replacement of fixed assets and a reduction in the number of redundant employees. The transformation of the regional industrial complex, in-first — is the conversion based on the philosophy of innovation of industrial activity in the synthesis of the development of entrepreneurial competence of senior management.

In this respect, on the part of economic science and the business community with an interest in the quantification of the contribution of competitive advantages in the dynamics of the key indicators of industrial production regions. Therefore, the search and evaluation of the statistical regularities of influence of regional competitive advantages in key indicators of the economic development of the region are now extremely relevant.

2 The system attributes (criteria) statistical measurement "regional competitive advantages" category

The statistics are no methodological development of the assessment of regional competitive advantages. Not the theoretical substantiation categories of regional competitive

advantages of industrial activity, which complicates the process of analytical reasoning to assess its impact on the dynamics of the key indicators.

The author of the study highlighted a set of properties inherent theoretical meaningful economic interpretation of the definition of regional competitive advantages. The generalization of these properties allows to uniquely identify the competitive advantages of regional development category of industrial activity in the framework of statistical science.

Learning the term is relatively new to economics. Domestic and foreign researchers carried out an analysis of this category in relation to the competitiveness of the term.

The basis of competition and competitiveness theory was laid in the work [1] A. Smith. According to the scientist, the country exported goods, the production of which has an absolute advantage. Later, Ricardo Smith improved teaching and developed the theory of comparative costs. According to this theory, the country benefits from trade in goods, the production of which have a higher relative efficiency.

The greatest recognition in the modern world received theory of competition American economist Michael Porter [2]. At the heart of Porter's teachings [2] is the concept of "value chain". Underneath scientist understands a set of interrelated activities, which allows you to create value (cost) for the end customer (consumer). Porter's value chain in companies of one branch may vary. These differences, in his opinion, arise due to variations in the company's strategy, buying groups, the organization's history, geographical location, etc. A comparison of the value chains of competitors brings out their differences, the underlying competitive advantages.

Symptom of comparison allows you to compare competitors in the same market, ie to relate the subjects of competition involved a uniform industrial production. The grouping of statistical data of economic entities on the criterion of homogeneity of their products is made possible by using statistical classifiers.

From 1 January 2011, statistical surveys practice in the Republic of Belarus started using "General Classification of Economic Activities". According to the new classifier economic activity of the country divided by economic activity. Thus under economic activity is meant a process where material resources, equipment, labor and technology are combined in such a way that it produces a similar set of outputs.

Using the modern classification of activities, allows you to group objects of statistical observation in terms of homogeneity of products. Research category "competitive advantage" in relation to industrial production should be carried out by directly comparable products. Relatively high security of the territory of a particular resource may have a positive impact on some economic activities and to be neutral in relation to the other. Therefore, analysis of regional competitive advantages of industrial activities is made possible by using the classifier.

Theory L. N. Chainikova [3] reveal the dynamism of the property. The use of this property in the statistical description of the key category will take into account the volatile nature of the outcome of competitive action.

Belarusian scientist A. S. Golovachev [4] emphasizes the importance of the effectiveness of the control of local authorities in addressing the problems of regional competitiveness. Increased levels of competitiveness of the territorial entities it connects

with the creation and management of the competitive advantages of the region. In this context it reveals another property "regional competitive advantages" category — conditioning. Symptom conditionality will allow competition to compare entities engaged homogeneous production, based on the impact of their activities, caused by scientific and technological innovation and transformation.

It is proposed to supplement the three identified properties that should be reflected in the definition of the category, another — the target determination. It will allow to characterize the key study object with the subject position estimates and the main purpose of the meeting.

National statistical offices form the necessary information support to national and regional authorities in order to make timely management decisions on the economic development of the country as a whole and its separate territories.

The role of statistical science in the knowledge of the socio-economic processes and phenomena is determined by the subject of its study. Statistics as a science studying the massive socio-economic processes and identify regularities inherent in them. Therefore, the regional competitive advantages in the context of the statistical interpretation is necessary to present a quantitative assessment of the development of industrial activities areas. The ability to quantify factors "regional competitive advantages" will assess its impact on the dynamics of the major indicators of industrial production in the region as a source of growth of the industry and the country as a whole.

Under the regional competitive advantages of industrial activity in the statistical science are invited to understand the comparative assessment of the dynamics of quantitative and qualitative indicators that determine economic growth of industrial activity of the administrative-territorial unit in comparison with the country as a whole (with the maximum of the observed values of other units / reference value).

The theoretical justification of the category "regional competitive advantages" allows mathematically formalized system of algorithms for the isolation of republican, branch and regional incentives in the region of the key indicators of industrial activity in the region. In the framework developed by stimulating the region's industrialization initiatives put multilevel target setting regional economic development industry: macro-level targets, industry and mezo-level targets (regional competitive advantages).

According to the author, republic-wide incentives for industrial activity growth in the region is a projection macro-level targets economic development areas. Macro-level targets regional development involves stimulating the growth of efficiency and competitiveness of industrial activities in the region through regulatory and legislative activity, subsidies and others. A formal expression of republican stimulus industrial activity development in the regions proposed to express the volume index of the key criterion of the study in the industry of the country.

The main purpose of the regional branch of industry growth is the formation of a competitive innovative industrial complex. One way to designated targets proposed definition of points of growth, the development of promising economic activities. The mathematical description of the industry factor is represented by the degree of lead (lag) of the dynamics of the physical volume of the key criteria for the type of industrial

activity to the same indicator for the country's industry as a whole.

Achieving targets meso-level intended to increase the intensity of use and (or) to create a new regional competitive advantages for the development of industrial activities in the areas of the Republic of Belarus. Impact of regional competitive advantages in key criterion for the study proposed to estimate the degree of lead (lag) of the dynamics of the key criterion of activity in the region compared to its dynamics in a similar type of industry in the country as a whole.

3 Conclusion

Based on the properties of the system, substantially inherent in the concept of "regional competitive advantages" (comparative, agility, conditioning, target determinism, the ability to quantitatively measure), in statistical science categories formulated its definition.

Theoretically grounded and mathematically formalized system of analytical algorithms isolation of republican, branch and regional industrial development incentives for key indicators of the region's industrial activities in the region.

References

[1] Smith A. (2009). *The Wealth of Nations: [trans. from English]*. M.: Eksmo (in Russian).

[2] Porter M. (1993). *International competition: [trans. from English]*. M.: Internat. relationship (in Russian).

[3] Chainikova L. N. (2008). *Methodological and practical aspects of the assessment of regional competitiveness: a monograph*. Tambov: Publishing House of the Thumb. state. tehn. University Press (in Russian).

[4] Golovachev A. S. (2009). Improving competitiveness of the region — the main task of the state and local authorities. *Problems of management*. Vol. **4**(33), pp. 119–125 (in Russian).

Section 6

COMPUTER DATA ANALYSIS AND MODELING IN APPLICATIONS

MULTIVARIATE ANALYSIS FOR IMAGE RECOGNITION SYSTEM TO ASSESS THE QUALITY OF THE MINERAL SPECIES

Olga Baklanova[1], Alexander Baklanov[2], Olga Shvets[3]
D.Serikbayev East-Kazakhstan State Technical University
Ust-Kamenogorsk, KAZAKHSTAN
e-mail: [1]OEBaklanova@mail.ru, [2]ABaklanov@ektu.kz,
[3]Olga.Shvets75@gmail.com

Abstract

This paper contains development of methods and algorithms of image recognition for mineral rocks. It is described algorithms of the cluster and morphological analysis for definition of rocks composition on colors and shapes. This approach is explained by the fact of the possible presence of objects with similar color-brightness characteristics, but with different shapes and there are objects with similar color-brightness characteristics also. Preliminary definition of group membership allows reducing the computational complexity of classification. It is determined the sorting into groups according to color of the object at the stage of segmentation. It is described and discussed the example using multivariate analysis for mineral rocks recognition.

1 Introduction

Minerals are homogeneous in composition and structure of the rocks and ores. They are natural chemical compounds resulting from various geological processes. Historically minerals initially determined by color and shape [1].

The development of computer vision system for mineral rocks is discussed in offered work in order to assess the qualitative composition of mineral rocks, in particular some problems of a technique and image recognition technology.

2 Materials and Methods

2.1 Methods of Identification of Mineral Rocks Images

Let us consider a sample of anode copper slag as an example (Figure 1. Micrographs of this sample were kindly provided by Eastern Research Institute of Mining and Metallurgy of Non-ferrous Metals (Kazakhstan, Ust-Kamenogorsk).

According to experts on microscopy of minerals from Eastern Research Institute of Mining and Metallurgy of Non-ferrous Metals at this picture there is no minerals having dependent on the direction of the plane of polarization of light. In this picture you can detect metallic copper and the following minerals: cuprite Cu_2O, magnetite Fe_3O_4, Delafosse $CuFeO_2$, silicate glass.

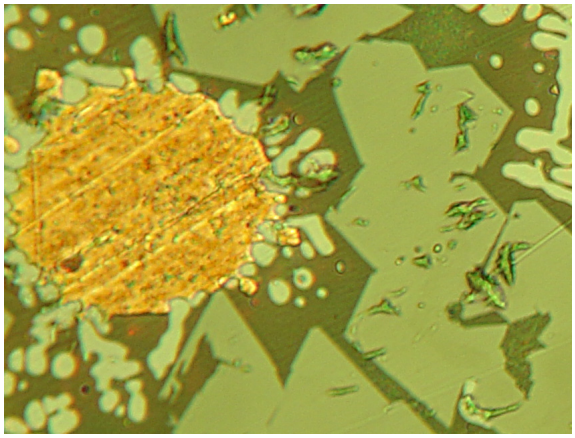


Figure 1: Micrograph of a sample of slag copper anode, increasing in 500 times.

Cuprite Cu_2O can be identified as follows: it is characterized by the shape of a round shape, color - it is light gray (sometimes with a slight bluish tint). Fe_3O_4 magnetite on micrographs may also be detected by color and shape. Color of magnetite on micrographs is dark gray. Shape is angular, as expressed by technologists, "octahedral".

Delafossite $CuFeO_2$ micrographs can allocate to the needle shape and gray (with a brownish tint) color.

Metallic copper on the micrographs can be found on the following criteria: color - yellow, shape - round, without flat faces.

Silicate glass - is a dark gray mass fills the rest of the space that is left of the other minerals. These data indicate that for real micrographs slag samples (and some other minerals) it is possible to use automated qualitative assessment of the mineral composition. After receiving the full image it is often needed to treat it, mainly to simplify further analysis.

2.2 Methods of Cluster Analysis for Mineral Rocks Images

Clustering - is the automatic partitioning of a set of elements into groups according to their similarity. Elements of the set can be anything, for example, data or characteristics vectors. Themselves groups are also called clusters [2].

In our case, using algorithms of cluster analysis will be the identification of ore minerals by color and texture characteristics of color-coded minerals identified in images taken in reflected light using a microscope [3].

In general, the K-means method segments the image on K different clusters (areas) located far away from each other based on certain criteria [4].

Segmentation method "K-means" is implemented through a two-step algorithm that minimizes the sum of distances "point-to-centroid" obtained by summing over all K clusters. Another words, the purpose of the algorithm is to minimize variability within clusters and maximize variability between clusters [5].

The purpose of cluster analysis - to implement such a partition of the n-dimensional feature space for k-clusters, in which the length between centroids of the resulting

clusters would be greatest, it is shown in the expression (1).

$$d_{i,j} \to max, \tag{1}$$

where $d_{i,j}$ is the distance between centroids of i-th and j-th clusters, $i, j = 0, \ldots, k$.

In this case, the most appropriate method of solving the problem of clustering is classic algorithm of unsupervised learning - a method of k-means (k-means method). Clustering incrementally in this case is as follows:

1. Lets specifie the number of clusters K, you want to find.
2. It is randomly selected K vectors $'$ from the set of vectors in selected space. These vectors are centroids of the clusters on the initial calculation stage.
3. Lets calculate the distance from each vector space used to each of the obtained centroids in step 2. It can be used metric (2)-(3) to determine the distance.

$$D_{(x,y)k} = \sqrt{\sum_{p=1}^{n} (P_{x,y}^{p} - P_{k}^{p})^2}, \tag{2}$$

$$D_{(x,y)k} = \sum_{p=1}^{n} |(P_{x,y}^{p} - P_{k}^{p})|, \tag{3}$$

   where:

   - (x, y) – coordinates of the observation,
   - $k \in [1, K]$ - cluster index,
   - n – dimensionality of the used feature space,
   - $p \in [1, n]$ – index of the feature observations.

4. Than we determine the centroid of the cluster to which the distance from the observation is the smallest. This cluster matched the observation.
5. Going through all available vectors we can recalculate centroids for each resulting cluster according (4).

$$P\prime_{(x,y)k}^{n} = \frac{1}{S(k)} \sum_{s=1}^{S_k} (P_{(x,y)s}^{n}), \tag{4}$$

   where:

   - k - cluster index,
   - $S(k)$ – number of observations related to the cluster index k,
   - s - indexes of the observations,
   - $P_{k}^{\prime n}$ - new value n-th feature of centroid cluster k.

6. Iterative process stops on steps 3-5 when the process of centroids changes stops or centroids will be fluctuate around some stable values. If the step of centroids change reaches a predetermined value it is possible to stop iterations.

Algorithm of the program can provide additional information after completion of the segmentation such as:

- a sum of distances "point-to-centroid";
- coordinates of centroid as well as some other data.

Algorithm of K-method can converge to a local optimum, when the separation points move any point to another cluster it increases the resultant sum of the distances. This problem can be solved only by a reasonable (successful) choice of initial points [7].

2.3 Methods of the Morphological Analysis of Mineral Shapes

Identification of the classification parameters is one of the primary task in pattern recognition [6].

It is offered the following description of the basic model of the object on the basis of morphological features (5–8):

$$M = \langle C, F, G \rangle, \tag{5}$$
$$C = \langle H, Sc, V \rangle, \tag{6}$$
$$F = \langle A \rangle, \tag{7}$$
$$G = \langle S, \beta \rangle, \tag{8}$$

where:

- C - cortege of metrics color of the object;
- F – cortege of morphological metrics of the object;
- G – geometrical metrics of the object;
- H - tone, Sc - saturation; V - value;
- A - number of allocated erosion circles;
- S - area of the object , β - the ratio of the long axis to the short one.

Proposed formalized description is focused on the entire spectrum of morphologically recognizable object parameters.

3 Results and Discussion

Nowadays developed automated image recognition system for assessing the qualitative composition of mineral rocks consists of 7 main subsystems [3]:

1. Research and getting micrograph rock.
2. Input and identification micrograph rock.
3. Pre-processing: improving the quality.
4. Definition of image reduction threshold [8].
5. Select the feature vector for cluster analysis.
6. Cluster analysis of color image to determine the mineralogical composition of rocks.
7. Morphological analysis of mineral shape to determine the mineralogical composition of rocks.

Each cluster includes a certain number of points. Given the ratio of the number of points allocated in each cluster with a number of common points can be displayed relative rates of minerals in rock samples. Various minerals marked in different colors. In this case, the metallic copper is red, magnetite - blue cuprite - orange. The result of the segmentation is shown in Figure 2.

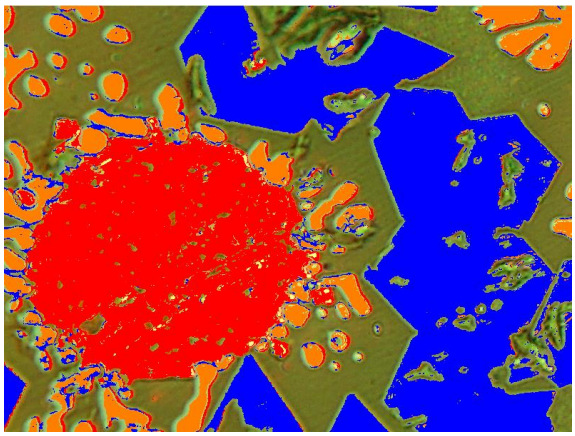


Figure 2: DFD - diagram decomposition subsystem “The result of cluster analysis”.

Considered sample has the following content of useful elements:

- Magnetite - 28.45%;
- Metallic copper - 18.45%;
- Cuprite - 7.92%.

4 Conclusion

In this article we have considered the development of segmentation algorithms for solving tasks of geological material analysis. We have proposed two different methods of ensuring the stability of results, based on pre-selection of centroids according to a few established principles in order to increase the stability of the segmentation pattern. The method of single-component searching consists of preliminary segmentation of the image based on a single variable vector. Every cluster has to get a matching segment assigned to it after the stage of segmentation. We have taken the average of every component of given variable space inside the assigned segment as initial values of the centroids. The basis of another method includes defining a point situated inside clusters defined by variables in N-dimensional space. We have proposed non-uniform partition of analyzed variable space followed by selection of initial values of centroids with maximum difference of color brightness characteristics in order to ensure stability. The program complex has been written in the language C# Visual Studio 2015. It was developed for results of research checking.

References

[1] Farndon J. (2006). *The practical encyclopedia of rocks and minerals.* Lorenz Books, London.

[2] Mandel J. (1988). *Cluster analysis.* Moscow: Finance and statistics.

[3] Baklanova O.E., Shvets O.Ya., Uzdenbaev Zh.Sh. (2014). Automation System Development For Micrograph Recognition For Mineral Ore Composition Evaluation In Mining Industry. *IFIP*. Vol. **436**, pp. 604–613.

[4] Odell P.L., Duran B.S. (1974). *Cluster Analysis: A Survey*. Springer-Verlag, NY.

[5] Huang Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *D. Mining and Knowledge Disc.* Vol. **2**, pp. 283–304.

[6] Gonsalez R.C., Woods R.E. (2011). *Digital image processing*. Pearson Education.

[7] Baklanova O.E., Shvets O.Ya. (2014). Methods and Algorithms of Cluster Analysis in the Mining Industry. Solution of Tasks for Mineral Rocks Recognition. *Proc. 11th Int. Conf. Sig. Proc. and Multimedia App. (SIGMAP'2014)*, pp. 165–171.

[8] Baklanova O.E., Shvets O.Ya. (2014). Development of Methods and Algorithms of Reduction for Image Recognition to Assess the Quality of the Mineral Species in the Mining Industry. *LNCS*. Vol. **8671**, pp. 75–83.

ON MODE JUMPING IN MCMC FOR BAYESIAN VARIABLE SELECTION WITHIN GLMM

A. A. Hubin[1], G. O. Storvik[2]
University of Oslo
Oslo, NORWAY
e-mail: [1]aliaksah@math.uio.no, [2]geirs@math.uio.no

Abstract

Generalized linear mixed models (GLMM) are addressed for inference and prediction in a wide range of different applications providing a powerful scientific tool for the researchers and analysts coming from different fields. At the same time more sources of data are becoming available introducing a variety of hypothetical explanatory variables for these models to be considered. Estimation of posterior model probabilities and selection of an optimal model is thus becoming crucial. We suggest a novel mode jumping MCMC procedure for Bayesian model averaging and model selection in GLMM.

1 Introduction

In this paper we study variable selection in generalized linear mixed models (GLMM) addressed in the Bayesian setting. These models allow to carry out detailed modeling in terms of both linking reasonably chosen responses and explanatory variables via a proper link function and incorporating the unexplained variability and dependence structure between the observations via random effects. Being one of the most powerful modeling tools in modern statistical science GLMM models have proven to be efficient in numerous applications from banking to astrophysics and genetics [2, 3]. The posterior distribution of the models can be viewed as a relevant measure for the model evidence, based on the observed data. The number of models to select from is exponential in the number of candidate variables, moreover the search space in this context is often extremely non-concave. Hence efficient search algorithms have to be adopted for evaluating the posterior distribution of models within a reasonable amount of time. In this paper we introduce efficient mode jumping MCMC algorithms for calculating and maximizing posterior probabilities of the GLMM models.

2 The generalized linear mixed regression model

Generalized linear mixed models consist of a response Y_t coming from the exponential family distribution, a vector of P variables X_{ti} for observations $t \in \{1, ..., T\}$ and latent indicators $\gamma_i \in \{0, 1\}, i \in \{1, ..., P\}$ defining if variable X_{ti} is included into the model ($\gamma_i = 1$) or not ($\gamma_i = 0$). We are also addressing the unexplained variability of the responses and the correlation structure between them through random effects δ_t with a specified parametric and sparse covariance matrix structure. Conditioning on the

random effect we model the dependence of the responses on the explanatory variables via a proper link function $g(\cdot)$:

$$Y_t|\mu_t \sim \mathfrak{f}(y|\mu_t),\ g(\mu_t) = \beta_0 + \sum_{i=1}^{P} \gamma_i \beta_i X_{ti} + \delta_t,\ \boldsymbol{\delta} = (\delta_1, ..., \delta_T) \sim N_T(\mathbf{0}, \boldsymbol{\Sigma}_b).$$

Here $\beta_i \in \mathbb{R}$, $i \in \{0, ..., P\}$, are regression coefficients showing in which way variables influence the linear predictor and $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_b(\boldsymbol{\psi}) \in \mathbb{R}^T \times \mathbb{R}^T$ is the covariance structure of the random effect. We then put relevant priors for the parameters of the model in order to make a fully Bayesian inference:

$$\gamma_i \sim Binom(1, q),\ \beta_i|\gamma_i \sim \mathbf{1}(\gamma_i = 1)N(\mu_\beta, \sigma_\beta^2),\ \boldsymbol{\psi} \sim \varphi(\boldsymbol{\psi}),$$

where q is the prior probability of including a covariate into the model.

Let $\boldsymbol{\gamma} = (\gamma_1, ...\gamma_P)$, which uniquely defines a specific model. Then there are 2^P different fixed models in the space of models Ω_γ. We would like to find a set of the best models of this sort with respect to a certain model selection criterion - namely marginal posterior model probabilities (PMP) - $p(\boldsymbol{\gamma}|\mathbf{y})$, where $\mathbf{y}$ is the observed data. For the class of models addressed marginal likelihoods (MLIK) - $p(\mathbf{y}|\boldsymbol{\gamma})$ are obtained by the INLA approach [5]. Then PMP can be found using Bayes formula and estimated by iterating through the reasonable set of models $\mathbb{V}$ in the space of models Ω_γ.

$$p(\boldsymbol{\gamma}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \Omega_\gamma} p(\mathbf{y}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')} \approx \frac{\mathbf{1}(\gamma \in \mathbb{V})p(\mathbf{y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \mathbb{V}} p(\mathbf{y}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}. \quad (1)$$

In (1) only models with high MLIK give significant contributions and thus iterating through them when constructing $\mathbb{V}$ is vital. The problem seems to be pretty challenging, because of both the cardinality of the discrete space $\Omega_{\boldsymbol{\gamma}}$ growing exponentially fast with respect to the number of variables and the fact that $\Omega_{\boldsymbol{\gamma}}$ is multimodal in terms of MLIK. Furthermore, the modes are often sparsely located [3]. [3] also report and discuss properties of the obtained in (1) estimator.

3 Mode jumping MCMC

In the MCMC approach as described by [4], Metropolis-Hastings algorithms are addressed as a class of methods for drawing from a complicated target distribution. [6] describes high potential flexibility in choices of proposals by means of generating additional auxiliary states allowing cases where the proposal densities are not directly available. The auxiliary states can be chains generated by some local optimizers chosen randomly from a mixture and allowing for jumps to alternative modes. [6] shows that the detailed balance equations is satisfied for this general case. Assume the current state to be $\boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma})$. Generate $(\boldsymbol{\chi}^*, \boldsymbol{\gamma}^*) \sim q(\boldsymbol{\chi}^*, \boldsymbol{\gamma}^*|\boldsymbol{\gamma})$ and consider $\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\chi}^*, \boldsymbol{\gamma}^* \sim h(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\chi}^*, \boldsymbol{\gamma}^*)$ as some auxiliary variables for some arbitrary chosen $h(\cdot|\cdot)$. Accept $\boldsymbol{\gamma}' = \boldsymbol{\gamma}^*$ with the following acceptance probability

$$r_m(\boldsymbol{\chi}, \boldsymbol{\gamma}; \boldsymbol{\chi}^*, \boldsymbol{\gamma}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\gamma}^*)h(\boldsymbol{\chi}^*|\boldsymbol{\gamma}^*, \boldsymbol{\chi}, \boldsymbol{\gamma})q(\boldsymbol{\chi}, \boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma})h(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\chi}^*, \boldsymbol{\gamma}^*)q(\boldsymbol{\chi}^*, \boldsymbol{\gamma}^*|\boldsymbol{\gamma})}\right\}, \quad (2)$$

or remain in the previous state otherwise. Then an ergodic Markov chain is generated and $\boldsymbol{\gamma}' \sim \pi(\boldsymbol{\gamma}')$. In a typical setting $\boldsymbol{\chi}^*$ is generated first, followed by $\boldsymbol{\gamma}^*$. The extra $\boldsymbol{\chi}$ is needed in order to calculate a legal acceptance probability, relating to a symmetric reverse move.

For generating the locally optimized proposals we first make a big jump to a new region of interest with respect to kernel $\mathsf{q}_\mathsf{l}(\boldsymbol{\chi}_0^*|\boldsymbol{\gamma})$, followed by some local optimization of $\pi(\boldsymbol{\gamma})$ with the chosen transition kernels $\mathsf{Q}_\mathsf{o}(\boldsymbol{\chi}_i^*|\boldsymbol{\chi}_{i-1}^*)$, $i \in \{1, ..., k\}$, which can be either stochastic or deterministic, and finally make randomization $\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}^*|\boldsymbol{\chi}_k^*)$ with a kernel based on a small neighborhood. For the reverse move we correspondingly first make a big jump $\mathsf{q}_\mathsf{l}(\boldsymbol{\chi}_0|\boldsymbol{\gamma}^*)$, followed by the same type of local optimization $\mathsf{Q}_\mathsf{o}(\boldsymbol{\chi}_i|\boldsymbol{\chi}_{i-1})$, $i \in \{1, ..., k\}$, and finally the probability of transition from the point at the end of optimization to the initial solution γ is calculated with respect to the randomizing kernel $\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}|\boldsymbol{\chi}_k)$. Then acceptance probabilities with respect to (2) are calculated and the move to a new state is either accepted or rejected. A convenient choice of $h(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^*, \boldsymbol{\chi}^*)$ function allowing to store very little of the information from the local optimization routine is to consider it of a form $h(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^*, \boldsymbol{\chi}^*) = \mathsf{h}(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$:

$$\mathsf{h}(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \mathsf{q}_\mathsf{l}(\boldsymbol{\chi}_0|\boldsymbol{\gamma}^*) \left[\prod_{i=1}^{k} \mathsf{Q}_\mathsf{o} \left(\boldsymbol{\chi}_i | \boldsymbol{\chi}_{i-1} \right) \right].$$

Then (2) reduces to

$$r_m(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \min \left\{ 1, \frac{\pi(\boldsymbol{\gamma}^*)\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}|\boldsymbol{\chi}_k)}{\pi(\boldsymbol{\gamma})\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}^*|\boldsymbol{\chi}_k^*)} \right\}.$$

We recommend that in not less than 95% of the proposals no mode jumping is performed. This provides the global Markov chain with both good mixing between the modes and accurate exploration of the regions around them. As described by [3] we address *accept the first improving neighbor*, *accept the best neighbor*, *simulated annealing*, and *local MCMC* approaches for performing local combinatorial optimization, whilst transitions in these routines are based on random change or deterministic swaps of a fixed or randomized number of components of $\boldsymbol{\gamma}$, or by uniform addition or deletion of a positive component in $\boldsymbol{\gamma}$. Notice that tuning of the probabilities of addressing local optimizers with particular proposal kernels in a mixture is often beneficial and we can carry it out during the burn in of the mode jumping MCMC without violating the desired ergodicity of the chain [3]. Also notice that both local optimizers and the global MCMC procedures are extensively parallelizible [3]. Finally, all of the unique models visited during the procedure are then appended to $\mathbf{V} \subseteq \Omega_\gamma$ and used to estimate (1). Alternative MCMC estimators for (1) as described in [1, 3, 4] are also available.

4 Results and discussion

We apply and compare the described algorithm further addressed as MJMCMC on the famous U.S. Crime Data and compare its performance to some popular algorithms such as BAS and competing MCMC methods (MC^3, RS, and thinned RS) with no

mode jumping [1, 3]. We apply the Bayesian linear regression with a *g-prior* [1] to the aforementioned data set with $T = 47$ observations and $P = 15$ explanatory variables. We carry out 100 replications of each algorithm on 10% of cardinality of Ω_γ, which in the best case scenario contains 86% of the total posterior model mass. As can be seen

Parameter	**Truth**	**MJMCMC**		**BAS**	**MC^3**	**RS**	**RS-thin**
BIAS$\times 10^5$	0.00	15.49	9.28	10.94	27.33	27.15	27.3
RMSE$\times 10^5$	0.00	16.83	10.00	11.65	34.39	34.03	28.99
Explored mass	1.00	0.58	0.71	0.67	0.10	0.10	0.13
Unique models	32768	1909	3237	3276	829	1071	1722
Total models	32768	3276	5936	3276	3276	3276	3276

Table 1: BIAS, RMSE of posterior model probabilities, explored masses, total and efficient numbers of iterations from the 100 replications of the involved algorithms.

from Table 1, our approach by far outperforms simpler MCMC methods in terms of the total posterior mass captured [1, 3] as well as the RMSE and BIAS [1, 3] of the model posterior probabilities (1); moreover, unlike the latter, it does not get stuck in the local modes and estimates a greater number of the unique models within the same amount of proposals. On the same amount of estimated models MJMCMC outperforms BAS in terms of all parameters, however for the same amount of proposals BAS is slightly better. More examples with various GLMM addressed and description of the developed R package *EMJMCMC* can be found in [3]. In general, we claim that MJMCMC is not only a very competitive novel algorithm, but also that it addresses a much wider class of models (GLMM) than all of the competing approaches. In future it would be of an interest to extend the procedure to level of the choice of link functions, priors and response distributions.

References

[1] Clyde M., Ghosh J., Littman M. (2011). Bayesian adaptive sampling for variable selection and model averaging. *J. Comp. Graph. Stat.* Vol. **20**(1), pp. 80–101.

[2] Cressie N., Wikle C.K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, NJ.

[3] Hubin A., Storvik G. (2016). Efficient mode jumping MCMC for Bayesian variable selection in GLMM. `arXiv:1604.06398v1`

[4] Robert C., Casella G. (2005). *Monte Carlo statistical methods*. Springer, NY.

[5] Rue H., Martino S., Chopin N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Royal Statistical Sosciety*. Vol. **71**(2), pp. 319–392.

[6] Storvik G. (2011). On the flexibility of Metropolis-Hastings acceptance probabilities in auxiliary variable proposal generation. *Scand. J. Stat.* Vol. **38**, pp. 342–358.

DEVELOPMENT OF THE MASTER PROGRAM ON APPLIED COMPUTER DATA ANALYSIS WITHIN THE TEMPUS PROJECT "APPLIED COMPUTING IN ENGINEERING AND SCIENCE"

Alexey Kharin[1], Peter Filzmoser[2], Peter Gabko[3]
[1]*Belarusian State University*
Minsk, BELARUS
[2,3]*TU Wien*
Vienna, AUSTRIA
e-mails: [1]KharinAY@bsu.by, [2]P.Filzmoser@tuwien.ac.at, [3]gabko@tuwien.ac.at

Abstract

Development of the new master program on Applied Computer Data Analysis within the TEMPUS Project "Applied Computer Data Analysis" is discussed. The program has been started in 2015 at the Belarusian State University.
Keywords: master program, TEMPUS, ACES project, computer data analysis, statistical modeling

The successfull research co-operation in the area of Statistical Data Analysis between scientists from the Vienna University of Technology (TU Wien) and Belarusian State University (BSU) was the base for the TEMPUS project "Applied Computing in Engineering and Science" (ACES) that has gathered in one team the experts in 5 areas benefiting from each other:

- Scientific Computing;
- Mathematical Modeling;
- Numerical Analysis and Optimization;
- Statistical Modeling;
- Statistical Computing.

The main goal of the ACES Project [1] is starting of a new master program in the area of Applied Computing at three Universities from Belarus (BSU, Belarusian National Technical University, Yanka Kupala State University in Grodno) and two Universities from Russian Federation (Siberian Federal University in Krasnoyarsk, Tomsk Polytechnic University) under the methodical support from five EU Universities: TU Wien (Austria), KU Leuven (Belgium), University of Wuppertal (Germany), Technical University of Lisbon (Portugal), and Palacky University in Olomouc (Czech Republic). In Belarus the mentioned Universities will run the master program "Applied Computer Data Analysis" in accordance with the Educational Standard approved in 2015 by the Ministry of Education of the Republic of Belarus.

The BSU has started the master program in 2015 with students who have successfully completed the 5-years programs with the diploma of a specialist. In June of 2016 the nine graduates from the master program were acknowledged with the Master of Science diploma in applied mathematics and infromation technologies. The two other Belarusian Universities plan to start the master program in 2016.

Two major activities were performed in the Project to reach the goal:

- Training at the EU partner Universities of the staff from the eastern partner Universites to get the optimal structure and contents of the curriculum;
- Preparing of the teaching materials books for students in 5 areas mentioned above to provide students with the information agreed by the consortium of partner Universities as the obligatory knowledge.

Some of the teaching materials books cover more than just the obligatory topics as, for expamle, the teaching materials on Statistical Modeling. The main parts of the contents are: Introduction; Basics of Multivariate Statistical Analysis; Principal Component Analysis; Factor Analysis; Regression Analysis; Discriminant Analysis; Advanced Methods for Classification; Cluster Analysis; Advanced Methods for Statistical Inference; Conclusions; References. The mentioned book is strongly related to the teaching materials on Statistical Computing.

At the moment we can mention the fact that the International Conference on Computer Data Analysis and Modeling in Minsk granted a lot into the research co-operation between TU Wien and BSU, and later into the ACES Project, is now benefiting from contributions of the developed master program graduates.

References

[1] `http://www.ai.tuwien.ac.at/aces`

UNRELIABLE QUEUEING SYSTEM WITH BACKUP SERVER

V. I. Klimenok
Belarusian State University
Minsk, BELARUS
e-mail: klimenok@bsu.by

Abstract

In this paper, we analyze a queueing system with two main unreliable servers and backup reliable server. The input flow is a *BMAP* (Batch Markovian Arrival Process). Heterogeneous breakdowns arrive to the main servers according to a *MMAP* (Marked Markovian Arrival Process). Service times and repair time have *PH* (Phase Type) distribution. The queue under consideration can be applied for modeling of hybrid communication system. We derive a condition for stable operation of the system, calculate its stationary distribution and base performance measures.

1 Introduction

As it is mentioned in [6], one of the main directions of creating the ultra-high speed (up to 10 Gbit/s) and reliable wireless means of communication is the development of hybrid communication systems based on laser and radio-wave technologies. Because of the high practical need for hybrid communication systems, a considerable amount of studies of this class of systems have appeared recently. Some results of these studies are presented in [1], [5]- [7]. As we know, all research on hybrid communication systems are devoted to study single-server queues with backup server. The present work is a further development of these studies to the case of a queueing system with two main unreliable servers and backup reliable server. This system is suitable to model a hybrid communication system consisting of two main communication channels - FSO (Free Space Optics) channel and millimeter-wave radio channel - and radio-wave IEEE802.11n channel which is used as a backup channel. FSO channel can not transmit data in conditions of poor visibility (fog or overcast weather) and millimeter radio-wave channel can not transmit during precipitation (rain, snow, etc.). In case when FSO-channel and millimeter radio-wave channel break down, information is transmitted via backup radio-wave IEEE802.11n channel which is absolutely reliable but has a much slower rate compared with the main transmission channels. Thus, a hybrid communication system is able to transmit data at practically all weather conditions.

2 Mathematical model

We consider a queueing system with waiting room and two unreliable heterogeneous servers which model FSO and mm-wave channels and backup reliable server which

models radio-wave IEEE802 channel. In the following, FSO cannel will be named as server 1, mm-wave channel as server 2 and radio-wave IEEE802 channel as server 3.

Customers arrive into the system in accordance with Batch Markovian Arrival Process ($BMAP$). The $BMAP$ is very general arrival process which is able to capture any correlation and burstiness that are commonly seen in the traffic of modern communication networks. The $BMAP$ is defined by the underlying process ν_t, $t \geq 0$, which is an irreducible continuous-time Markov chain with finite state space $\{0, \ldots, W\}$, and the matrix generating function $D(z) = \sum_{k=0}^{\infty} D_k z^k$, $|z| \leq 1$. The batches of customers enter the system only at the epochs of the chain ν_t, $t \geq 0$, transitions. The $(W+1) \times (W+1)$ matrices D_k, $k \geq 1$, (non-diagonal entries of the matrix D_0) define the intensities of the process ν_t, $t \geq 0$, transitions which are accompanied by generating the k-size batch of customers. The intensity (fundamental rate) of the $BMAP$ is defined as $\lambda = \boldsymbol{\theta} D'(1)\mathbf{e}$ where the vector $\boldsymbol{\theta}$ is the unique solution of the system $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Hereinafter $\mathbf{e}$ is a column vector of units. For more information about the $BMAP$ see, e.g. [3].

The service time of a customer by the jth server, $j = 1, 2, 3$, has PH type distribution with irreducible representation $(\boldsymbol{\beta}_j, S_j)$. The service process on the jth server is directed by the Markov chain $m_t^{(j)}$, $t \geq 0$, with state space $\{1, \ldots, M_j, M_j + 1\}$ where $M_j + 1$ is an absorbing state. The intensities of transitions into the absorbing state are defined by the vector $\mathbf{S}_0^{(j)} = -S_j\mathbf{e}$. For more information about PH type distribution see, e.g. [4].

Breakdowns arrive to the servers 1,2 according to a $MMAP$ which is defined by the underlying process η_t, $t \geq 0$, with state space $\{0, \ldots, V\}$ and by the matrices H_0, H_1, H_2. The matrix H_0 defines the intensities of the process η_t, $t \geq 0$, transitions which does not lead to generation of a breakdown. The matrix H_j defines the intensities of the η_t, $t \geq 0$, transitions which are accompanied by generating a breakdown which is directed to the server $j, j = 1, 2$.

When a breakdown attacks one of the main server, the repair period at this server starts immediately and the other main server, if it is available, begins the service of the interrupted customer anew. If the latter server is busy or under repair, the customer goes to the server 3 and starts its service anew. However, if during the service time of the customer at the server 3 one of the main servers becomes fault-free, the customer restarts its service on this server.

The repair period at the jth main server, $j = 1, 2$, has PH type distribution with an irreducible representation $(\boldsymbol{\tau}_j, T_j)$. The repair process at the jth server is directed by the Markov chain $r_t^{(j)}$, $t \geq 0$, with state space $\{1, \ldots, R_j, R_j + 1\}$ where $R_j + 1$ is an absorbing state.

3 Process of the system states

Let at the moment t

i_t be the number of customers in the system, $i_t \geq 0$,

$$n_t = \begin{cases} 0, & \text{if both main servers are fault-free (both ones are busy or idle);} \\ 0_j, & \text{if both main servers are fault-free, the } j\text{th server is busy and} \\ & \text{the other one is idle, } j = 1, 2; \\ 1, & \text{if the server 1 is under repair;} \quad 2, \text{ if the server 2 is under repair;} \\ 3, & \text{if both servers are under repair;} \end{cases}$$

$m_t^{(j)}$ be the state of the directing process of the service at the j-th busy server, $j = 1, 2, 3, m_t^{(j)} = \overline{1, M_j}$; $r_t^{(j)}$ be the state of the directing process of the repair time at the j-th busy server, $j = 1, 2$, $r_t^{(j)} = \overline{1, R_j}$; ν_t and η_t be the states of the directing processes of the $BMAP$ and the $MMAP$ respectively, $\nu_t = \overline{0, W}$, $\eta_t = \overline{0, V}$.

The process of the system states is described by the regular irreducible continuous time Markov chain, $\xi_t, t \geq 0$, with state space

$$X = \{(0, n, \nu, \eta),\ i = 0, n = \overline{0, 3},\ \nu = \overline{0, W},\ \eta = \overline{0, V}\} \bigcup$$

$$\{(i, 0_j, \nu, \eta, m^{(j)}),\ i = 1, j = 1, 2, n = 0_j, \nu = \overline{0, W}, \eta = \overline{0, V},\ m^{(j)} = \overline{1, M_j}\} \bigcup$$

$$\{(i, 0, \nu, \eta, m^{(1)}, m^{(2)}),\ i > 1, n = 0, \nu = \overline{0, W}, \eta = \overline{0, V},\ m^{(1)} = \overline{1, M_1}, m^{(2)} = \overline{1, M_2}\} \bigcup$$

$$\{(i, 1, \nu, \eta, m^{(2)}, r^{(1)}),\ i \geq 1, n = 1,\ \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(2)} = \overline{1, M_2}, r^{(1)} = \overline{1, R_1}\} \bigcup$$

$$\{(i, 2, \nu, \eta, m^{(1)}, r^{(2)}),\ i \geq 1, n = 2,\ \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(1)} = \overline{1, M_1}, r^{(1)} = \overline{1, R_2}\} \bigcup$$

$$\{(i, 3, \nu, \eta, m^{(3)}, r^{(1)}, r^{(2)}),\ i > 0, n = 3,\ \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(3)} = \overline{1, M_3},$$
$$r^{(j)} = \overline{1, R_j}, j = 1, 2\}.$$

Lemma 1. *Infinitesimal generator of the Markov chain ξ_t, $t \geq 0$, has the following block structure*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & Q_{0,4} \cdots & \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & Q_{1,4} \cdots & \\ O & Q_{2,1} & Q_1 & Q_2 & Q_3 \cdots & \\ O & O & Q_0 & Q_1 & Q_2 \cdots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where blocks $Q_{i,j}$, $i, j \geq 0$, are the matrices formed by intensities of the chain transition from the states corresponding to the value i of the denumerable component i_n to the states corresponding to the value j of this component.

Corollary 1. *The Markov chain $\xi_t, t \geq 0$, belongs to the class of continuous time quasi-Toeplitz Markov chains, see [2].*

4 Stationary distribution. Performance measures

Theorem 1. *The necessary and sufficient condition for existence of the stationary distribution of the Markov chain* ξ_t, $t \geq 0$, *is the fulfillment of the inequality*

$$\lambda < -\boldsymbol{\pi}_0(S_1 \oplus S_2)\mathbf{e} + \boldsymbol{\pi}_1\mathbf{S}_0^{(2)} + \boldsymbol{\pi}_2\mathbf{S}_0^{(1)} + \boldsymbol{\pi}_3\mathbf{S}_0^{(3)}, \tag{1}$$

where $\boldsymbol{\pi}_0 = \mathbf{x}_0(\mathbf{e}_{V+1} \otimes I_{M_1M_2})$, $\boldsymbol{\pi}_1 = \mathbf{x}_1(\mathbf{e}_{V+1} \otimes I_{M_2} \otimes \mathbf{e}_{R_1})$, $\boldsymbol{\pi}_2 = \mathbf{x}_2(\mathbf{e}_{V+1} \otimes I_{M_1} \otimes \mathbf{e}_{R_2})$, $\boldsymbol{\pi}_3 = \mathbf{x}_3(\mathbf{e}_{V+1} \otimes I_{M_3} \otimes \mathbf{e}_{R_1R_2})$ *and the vectors* $\mathbf{x}_1$, $\mathbf{x}_3$ *are sub-vectors of the vector* $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$, *which is the unique solution of the system of linear algebraic equations.*

In what follows we assume inequality (1) be fulfilled. Denote by $\boldsymbol{p}_i$ the row vector of steady state probabilities corresponding the value i of the first component of the chain ξ_t, $t \geq 0$, $i \geq 0$. To calculate the vectors $\boldsymbol{p}_i$, $i \geq 0$, we use the numerically stable algorithm, see [2], which has been elaborated for calculating the stationary distribution of multi-dimensional continuous time quasi-Toeplitz Markov chains. Having the stationary distribution $\boldsymbol{p}_i$, $i \geq 0$, been calculated we find a number of important stationary performance measures of the system and examine their behavior through the numerical experiments.

References

[1] Arnon S., Barry J., Karagiannidis G., Schober R., Uysal M. (2012). *Advanced Optical Wireless Communication Systems*. Cambridge University Press.

[2] Klimenok V.I., Dudin A.N. (2006). Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*. Vol. **54**, pp. 245–259.

[3] Lucantoni D.M. (1991). New results on the single server queue with a batch Markovian arrival process. *Comm. Stat.-Stoch. Models*. Vol. **7**, pp. 1–46.

[4] Neuts M. (1981). *Matrix-geometric Solutions in Stochastic Models – An Algorithmic Approach*. Johns Hopkins University Press, Baltimore.

[5] Sharov S.Yu., Semenova O.V. (2010). Simulation model of wireless channel based on FSO and RF technologies. *Distributed Computer and Communication Networks. Theory and Applications*. pp. 368-374.

[6] Vishnevsky V., Kozyrev D., Semenova O. (2014). Redundant queueing system with unreliable servers. *Proc. 6th Int. Cong. Ultra Modern Telecommunications and Control Systems and Workshops*. pp. 383–386.

[7] Vishnevsky V.M., Semenova O.V., Sharov S.Yu. (2013). Modeling and analysis of a hybrid communication channel based on free-space optical and radio-frequency technologies. *Automation and Remote Control*. Vol. **72**, pp. 345–352.

ON SOME ASPECTS IN ACQUISITION OF BRAIN ELECTRICAL ACTIVITY

A. V. Kolchin[1], H. G. Ionkina[2]
[1]*Moscow Automobile and Road Construction State Technical University*
[2]*The First Sechenov Moscow State Medical University*
Moscow, RUSSIA
e-mail: [1]andrei.kolchin@gmail.com

Abstract

We give a description of a portable system for acquisition of the brain electrical activity and discuss some problems which arise while developing, implementing, and fine-tuning the system.

We developed, implemented, and fine-tuned a portable system for acquisition of the electrical activity of a brain, which was successfully utilised to acquire the electroencephalogram and nociceptive evoked potentials (EPs) in the somatosensory S_1HL and the anterior cingulate Cg areas of cerebral cortex in the right hemisphere in rats.

The heart of the system is either an Intel Pentium IV-based portable computer (an IBM ThinkPad G40 in our study) or a Raspberry Pi 2 ARM Cortex-A7-based microcomputer loaded with Linux. So the analogue-to-digital converter is selected from amongst those supported by the `COMEDI` project [7] which develops open-source drivers, tools, and libraries for data acquisition implemented as a core Linux kernel module suitable for real-time tasks. We choose the 16-channel analogue-to-digital converter `usbdux-fast` coupled with 4-channel amplifier modules assembled to the open specifications provided by Incite Technology Ltd., Computing & Maths Dept., University of Stirling, United Kingdom (see [5, 6]); the amplifier was originally developed for teaching ECG at the Medical Faculty of the Ruhr University Bochum. The full schematic diagrams of the converter and the amplifier can be found in [3, 5, 6]. We make use of readily-available electronic components which inhabit custom printed circuit boards. Since the libraries and the firmwares source codes are in public domain, in our experiments we succeeded in implementing necessary corrections and revisions of the software in minimal time. The generation of the stimulus routed to the tail of an experimental animal (a male Wistar rat) via a constant current isolator unit (we used the isolator unit A365 produced by World Precision Instruments, Inc.), as well as that of the synchronising stimulus which triggered the start of acquisition, were carried out with the use of either the IEEE-1284 parallel port of the Intel-based computer or the general purpose input-output ports of the Raspberry box. Both of the electrophysiological data acquisition and stimuli generation tasks can also be executed concurrently on dedicated computers of the above architectures.

The key features of the system consist of the following: high sensitivity (μV); high-resolution measurement (discretisation up to a hundred kHz per channel); presence of no filters of the input signal in both the analogue-to-digital converter and amplifier modules; this results in the near absence of analogue data loss while acquiring the real

time brain bioelectrical activity. The system can perform a software filtering of the input data flow when needed.

The complex problem to prevent garbling of the input data due to intense electromagnetic pollution of the environment was solved by making use of a multilayer shielding of the analogue part of the system (the laboratory animal, cables, and the amplifier) and by using an autonomous direct current source to feed the whole system.

If one takes the laboratory animal as a 'black box' whose input is some external stimulus while the output yields a high-volume data flow, then the goal of the experiment consists of separating the response to the input stimulus in the output data flow. The start of acquisition of the electrical activity of the rat brain is triggered by the synchronising impulse issued at a fixed (maybe zero) time interval before the leading front of the stimulating impulse.

The input electroencephalogram is conveniently observed in the real time with the use of `xoscope 1.12` [8]. In order to capture the data, we use `ktimetrace 0.2.37` [2]; it permits to capture samples from desired channels of our data acquisition device in a given time interval starting either from an arbitrary time instant or from that governed by the external synchronising signal and to save it to a file while providing a real-time graphing display. The data thus obtained form a text file whose each row consists of numerical values captured from the channels at the corresponding time instants. The size of the file can grow to a very large value, so we decide to use the appropriate file system (`ext4` in our study).

We investigate the role which the brain cortex plays in formation of the nociceptive reactions by means of analysis of the evoked potentials acquired in the somatosensory S_1HL and the anterior cingulate Cg areas of cerebral cortex in the right hemisphere in immobilised Wistar male rats before the intraperitoneal injection of a lipopolysaccharide (LPS) and at the 1st, 3rd and 7th days after it upon an electrocutaneous stimulation of the tail. The stimulation of the rat tail is by single rectangular current impulses of 80% of the initial vocalisation threshold. The EPs are averaged over 36 trials. The changes of the late components of the EPs, which reflect the emotional component of the nociceptive reaction, were analysed by their peak-to-peak amplitudes (A) and the areas of the secondary negative responses (S). In [4, 3], we give an example of dynamics of nociceptive evoked potentials registered in the somatosensory area of the rat's cerebral cortex before the intraperitoneal injection of the lipopolysaccharide and at the first and the seventh days after it. We thus came to the classical biostatistics problem to find whether there was an effect of a single administration of a drug or not (see, e.g., [1]); to solve it, we made use of the non-parametric Wilcoxon test; this test uses only the information on the differences between values of the parameters and their signs, and there is no need to make assumptions concerning the laws of distribution of the differences of the parameters under investigation upon the action of the drug. The parametric tests based on the normal approximations appear to be of little use in our case.

The battery of solutions we have used while developing and setting up this system are pioneering and allow us to deal with a wide range of problems of electrophysiology including electromyography, electrocardiography, electroencephalography, and record-

ing of neuronal activity in the brain.

All investigations on the laboratory animals were carried out in full compliance with the GLP principles.

References

[1] Glantz S.A. (2013). *Primer of Biostatistics.* McGraw–Hill, NY.

[2] Hess F.M. (2005). *KTimeTrace.*
`http://ktimetrace.sourceforge.net`

[3] Ionkina H.G., Kolchin A.V. (2015). Acquisition of the electrical activity of rat cerebral cortex. *Proc. Karelian Sci. Centre Russian Acad. Sci.* Vol. **15**, pp. 62–66.

[4] Kolchin A. V., Ionkina E. G. (2013). On acquisition of nociceptive evoked potentials in rats cerebral cortex. *Proc. 10th International Conf. Computer Data Analysis and Modeling: Theoretical and Applied Stochastics.* Vol. **1**, pp. 72–73.

[5] Porr B. (2007). *`USBDUX-fast`: Product description.*
`http://linux-usb-daq.co.uk/prod2_duxfast`

[6] Porr B. (2012). *ECG preamplifier for the* `USBDUX-D`.
`http://linux-usb-daq.co.uk/howto2/ecg`

[7] Schleef D., Hess F.M., Abbott I. (2014). *Comedi: Linux control and measurement device interface.*
`http://www.comedi.org`

[8] Witham T. (2005). *Xoscope for Linux.*
`http://xoscope.sourceforge.net`

ANALYSIS AND APPLICATION OF G-NETWORK WITH INCOMES AND RANDOM WAITING TIME OF NEGATIVE CUSTOMERS

M. Matalytski, V. Naumenko[1], D. Kopats
Grodno State University
Grodno, BELARUS
e-mail: [1]victornn86@gmail.com

Abstract

In the article an open Markov queueing G-network with incomes and random waiting time of negative customers has been considered. Negative customers destroy positive customers on the expiration of a random time. Queueing system (QS) receives a certain random income when positive customer arrives to it and loss when negative customer arrives to it. A technique for finding the expected incomes of the network QS has been proposed. In information systems and networks negative customers may describe behavior of requests for service, at which request is a command to stop the operation being performed or the behavior of computer viruses, the effects of which on the information (positive customer) occurs through a random time and has a damaging effect.

1 Introduction

Consider an open G-network [1] with n single-queues QS: $S_1,\ S_2,\ ...,\ S_n$. Lets introduce system S_0, from which Poisson flow of customers arrive to the network. The network state at time t described by the vector $k\left(t\right) = \left(\left(k_1(t), l_1(t)\right), ..., \left(k_n(t), l_n(t)\right)\right)$, which forms a homogeneous Markov process with a countable number of states, where the state $(k_i(t), l_i(t))$ means, that at time t in QS S_i there are k_i positive customers and l_i negative customers, $i = \overline{1, n}$. We introduce the vectors $k\left(t\right) = \left(k_1\left(t\right), k_2(t), ...\ , k_n(t)\right)$ and $l\left(t\right) = \left(l_1\left(t\right), l_2(t), ...\ , l_n(t)\right)$.

External arrivals to the network, service times of rates and probabilities of customer transitions between QS depend on time, [2]. In QS S_i from the outside (from the system S_0) is coming a Poisson stream of the positive customers with the intensity $\lambda_{0i}^{+}(t)$ and Poisson stream of negative customers with the intensity $\lambda_{0i}^{-}(t)$, $i = \overline{1,\ n}$. All flows of customers incoming to the network are independent. Lets $\mu_i^{+}\left(k_i(t)\right)$ – service rate of positive customers in QS S_i at time t, depend on count of customers at it system, $i = \overline{1,\ n}$. If in QS S_i at time t there are $k_i\left(t\right)$ customers, then the probability, that the positive customer serviced in QS S_i during time $[t, t + \Delta t)$, are equals $\mu_i^{+}\left(k_i(t)\right)\Delta t + o\left(\Delta t\right)$. Positive customer, get serviced in S_i at time t with probability $p_{ij}^{+}(t)$ move to QS S_j as a positive customer, and with probability $p_{ij}^{-}(t)$ – as a negative customer, and with probability $p_{i0}(t) = 1 - \sum_{j=1}^{n}\left[p_{ij}^{+}(t) + p_{ij}^{-}(t)\right]$ come out from the network to external environment, $i,\ j = \overline{1,\ n}$.

Negative customer is arrived to QS increases the length of the queue of negative customers for one, and requires no service. Each negative customer, located in i-th

QS, stay in the queue random time according to a Poisson process of rate $\mu_i^-(l_i)$, $i = \overline{1,\ n}$. By the end this time, negative customer destroy one positive customer in the QS S_i and leave the network. If after this random time in the system there are no positive customers, then given negative customer leave network, without exerting any influence on the operation of the network as a whole. Wherein probability that, in QS S_i negative customer leave queue during $[t, t+\Delta t)$, on condition that, in this QS at time t there are l_i negative customers, equals $\mu_i^-(l_i)\,\Delta t + o(\Delta t)$.

2 Finding expected incomes of network systems

Consider the dynamics of income changes of a network system S_i. Denote by the $V_i(t)$ its income at moment time t. Let the initial moment time income of the system equal $V_i(0) = v_{i0}$. The income of its QS at moment time $t + \Delta t$ can be represented in the form

$$V_i(t + \Delta t) = V_i(t) + \Delta V_i(t, \Delta t), \tag{1}$$

where $\Delta V_i(t, \Delta t)$ – income changes of the systemS_i at the time interval $[t, t+\Delta t)$, $i = \overline{1,n}$. To find its value we write down the value of the conditional probabilities of events that may occur during Δt and the income changes of its QS, associated with these events:

1. With probability $p_i^{(1)}(t, \Delta t) = \lambda_{0i}^+(t)\Delta t + o(\Delta t)$ at moment time t to the system S_i from the external environment will come positive customer, which will bring an income to the amount of r_{0i}, where r_{0i} – random variable (RV), expectation (E) which is equals $E\ \{r_{0i}\} = a_{0i}$, $i = \overline{1,n}$.

2. With probability $p_i^{(2)}(t, \Delta t) = \lambda_{0i}^-(t)\Delta t + o(\Delta t)$ in the QS S_i at moment time t from the external environment will come a negative customer, $i = \overline{1,n}$; income change of this system this case will not occur.

3. If at the moment time t at the system S_i is located $k_i(t)$of positive customers, then with probability $p_i^{(3)}(t, \Delta t) = \mu_i^+(k_i(t))\,u(k_i(t))\,p_{i0}(t)\Delta t + o(\Delta t)$, where $u(x)-$ Heaviside function, positive customer comes out from the network to the external environment, while the total amount of income of QS S_i is reduced by an amount which is equal to $-R_{i0}$, where $E\{R_{i0}\} = b_{i0}$, $i = \overline{1,\ n}$, $\mu_i^+(0) = 0$.

4. With probability $p_i^{(4)}(t, \Delta t) = \mu_i^-(l_i(t))\,u(l_i(t))\,\Delta t + o(\Delta t)$ in QS S_i at the moment time t negative customer, destroying positive customer in QS S_i, will leave the network, $i = \overline{1,n}$; In this case income for the system S_i decreases by an amount $-R_{i0}^{neg}$, $E\{R_{i0}^{neg}\} = b_{i0}^{neg}$, $\mu_i^-(0) = 0$, $i = \overline{1,n}$.

5. If at the moment time t in QS S_i there were $l_i(t)$ negative customers and there were not positive customers, then negative customer leaves this QS with probability $p_i^{(5)}(t, \Delta t) = \mu_i^-(l_i(t))\,(1 - u(k_i(t)))\,\Delta t + o\ (\Delta t)$, $i = \overline{1,n}$. In this case income change of QS S_i will not occur, $i = \overline{1,n}$.

6. If at the moment time in the system S_i there is positive customer, then after finishing it servicing in QS S_i it move to QS S_j as a positive customer With probability $p_i^{(6)}(t, \Delta t) = \mu_i^+(k_i(t))\,u(k_i(t))\,p_{ij}^+(t)\Delta t + o(\Delta t)$, $i, j = \overline{1,\ n}$, $i \neq j$; in such a transition

income of system S_i decreases by an amount R_{ij}, and income of system S_j will increase by this amount, where $E\{R_{ij}\} = a_{ij},\ i,j = \overline{1,\ n},\ i \neq j$.

7. If at the moment time t in system S_j there is positive customer, then serviced in S_j, it will move to system S_i with probability $p_i^{(7)}(t,\Delta t) = \mu_j^+(k_j(t))\,u(k_j(t))\,p_{ji}^+(t)\Delta t + o(\Delta t)$, at the same time income of QS S_i decreases by an amount r_{ji}, and income of QS S_j it will increase by the same amount, where $E\{r_{ji}\} = b_{ji},\ i,j = \overline{1,\ n},\ i \neq j$.

8. With probability $p_i^{(8)}(t,\Delta t) = \mu_i^+(k_i(t))\,u(l_i(t))\,p_{ij}^-(t)\Delta t + o(\Delta t)$ positive customer, serviced in QS S_i, at the moment time t forwarded to QS S_j as negative customer $i,j = \overline{1,\ n},\ i \neq j$; in such a transition system income of S_i decreases by an amount R_{ij}^{neg}, and income of system S_j will not change, where $E\left\{R_{ij}^{neg}\right\} = a_{ij}^{neg},\ i,j = \overline{1,\ n},\ i \neq j$.

9. With probability

$$p^{(9)}(t,\Delta t) = 1 - \left\{\lambda_{0i}^+(t) + \lambda_{0i}^-(t) + \mu_i^+(t)p_{i0}^+(t) + \mu_i^-(t) + \sum_{j=1}^{n}\mu_i^+(t)p_{ij}^+(t) + \right.$$

$$\left. + \sum_{j=1}^{n}\mu_j^+(t)p_{ji}^+(t) + \sum_{j=1}^{n}\mu_i^+(t)p_{ij}^-(t)\right\}\Delta t + o(\Delta t)$$

on time interval $[t, t+\Delta t)$ there will be no change of system S_i nothing is going to happen (not a positive customer or a negative customer is received and no customer is serviced), in this case, the total income of S_i may increase (decrease) to the amount of $r_i\Delta t$, where $E\{r_i\} = c_i,\ i = \overline{1,\ n}$.

It's obvious that $r_{ji}(\xi_j) = R_{ji}(\xi_j)$ with probability 1, i.e. $b_{ji} = a_{ji}, i,j = \overline{1,\ n}$. Suppose that at any instant of time RV r_{0i}, R_{i0}, R_{i0}^{neg}, R_{ij}, r_{ji}, R_{ij}^{neg} does not depend on RV r_i.

We will assume, that all network systems are single-queues and customers service rates in QS S_i has an exponential distribution with rate $\mu_i^+(t)$; let also negative customer, arrives to QS S_i, will leave it queue after a random time, that has an exponential distribution with rate $\mu_i^-(t)$. Consequently, in these cases, we obtain, that $\mu_i^+(k_i(t)) = u(k_i(t))\,\mu_i^+(t),\ \mu_i^-(l_i(t)) = u(l_i(t))\,\mu_i^-(t),\ i = \overline{1,n}$.

In addition suppose, that all systems operating under heavy-traffic regime, i.e. $k_i(t) > 0\ \forall t > 0,\ i = \overline{1,\ n}$. In the simulation, the effect of virus penetration into a computer network or during a computer attack on it occurs just such a situation. Also, we assume, that $l_i(t) > 0\ \forall t > 0,\ i = \overline{1,\ n}$. Then it follows from the foregoing

$$\Delta V_i(t,\ \Delta t) = \begin{cases} r_{0i} + r_i\Delta t \textit{ with probability } \lambda_{0i}^+(t)\Delta t + o(\Delta t), \\ -R_{i0} + r_i\Delta t \textit{ with probability } \mu_i^+(t)p_{i0}^+(t)\Delta t + o(\Delta t), \\ -R_{i0}^{neg} + r_i\Delta t \textit{ with probability } \mu_i^-(t)q_{i0}^-(t)\Delta t + o(\Delta t), \\ -R_{ij} + r_{\,i}\Delta t \textit{ with probability } \mu_i^+(t)p_{ij}^+(t)\Delta t + o(\Delta t), \\ r_{ji} + r_{\,i}\Delta t \textit{ with probability } \mu_j^+(t)p_{ji}^+(t)\Delta t + o(\Delta t), \\ -R_{ij}^{neg} + r_{\,i}\Delta t \textit{ with probability } \mu_i^+(t)p_{ij}^-(t)\Delta t + o(\Delta t), \\ r_i\Delta t \textit{ with probability } 1 - \left\{\lambda_{0i}^+(t) + \lambda_{0i}^-(t) + \right. \\ +\mu_i^+(t)p_{i0}^+(t) + \mu_i^-(t) + \sum_{j=1}^{n}\mu_i^+(t)p_{ij}^+(t) + \\ \left. + \sum_{j=1}^{n}\mu_j^+(t)p_{ji}^+(t) + \sum_{j=1}^{n}\mu_i^+(t)p_{ij}^-(t)\right\}\Delta t, j = \overline{1,n},\ j \neq i\ . \end{cases} \tag{2}$$

Therefore, taking into account (2) $E\left\{\Delta V_i\left(t,\,\Delta t\right)\right\} = f_i(t)\Delta t + o\left(\Delta t\right)$, where

$$f_i(t) = \left(\lambda_{0i}^{+}(t) + \lambda_{0i}^{-}(t)\right) a_{0i} + \sum_{j=1}^{n}(\mu_j^{+}(t)p_{ji}^{+}(t)b_{ji}) +$$

$$+ \sum_{j=1}^{n} p_{ij}^{-}(t)\left[a_{ij}^{neg}\mu_i^{+}(t) + c_i\left(\mu_i^{+}(t) + \mu_i^{-}(t)\right)\right] +$$

$$+ \sum_{j=1}^{n}\left[p_{ij}^{+}(t)\mu_i^{+}(t)(2c_i - a_{ij})\right] - b_{i0}\mu_i^{+}(t)p_{ij}^{+}(t) + b_{i0}^{neg}\mu_i^{-}(t) + c_i.$$

For $v_i\left(t\right) = M\left\{V_i\left(t\right)\right\}$ from (1) we have $v_i\left(t+\Delta t\right) = v_i(t) + E\left\{\Delta V_i\left(t,\,\Delta t\right)\right\}$, where, passing to the limit $\Delta t \to 0$, we get linear inhomogeneous differential equations of first order $\frac{dv_i(t)}{dt} = f_i\left(t\right)$, $i = \overline{1,\,n}$, i.e.

$$\frac{dv_i\left(t\right)}{dt} = \left(\lambda_{0i}^{+}(t) + \lambda_{0i}^{-}(t)\right) a_{0i} + \sum_{j=1}^{n}(\mu_j^{+}(t)p_{ji}^{+}(t)b_{ji}) +$$

$$+ \sum_{j=1}^{n} p_{ij}^{-}(t)\left[a_{ij}^{neg}\mu_i^{+}(t) + c_i\left(\mu_i^{+}(t) + \mu_i^{-}(t)\right)\right] +$$

$$+ \sum_{j=1}^{n}\left[p_{ij}^{+}(t)\mu_i^{+}(t)(2c_i - a_{ij})\right] - b_{i0}\mu_i^{+}(t)p_{ij}^{+}(t) + b_{i0}^{neg}\mu_i^{-}(t) + c_i.$$

By setting the initial conditions $v_i(0) = v_{i0}$, $i = \overline{1,\,n}$, we can find the expected incomes of the network systems. In this way

$$v_i\left(t\right) = v_{i0}\left(0\right) + \int_0^t f_i\left(\tau\right)d\tau,\ \ i = 1, \ldots, n.$$

References

[1] Gelenbe E. (1991). Product form queueing networks with negative and positive customers. *Appl. Prob.*. Vol. **28**, pp. 656-663.

[2] Naumenko V., Matalytski M. (2015). Investigation of networks with positive and negative messages, many-lines queueing systems and incomes. *J. Applied Mathematics and Computations Mechanics*. Vol. **14**(1), pp. 79-90.

ON THE INEQUALITY IN OPEN MULTISERVER QUEUEING NETWORKS

Saulius Minkevičius[1], Edvinas Greičius[2]
[1,2]*Faculty of Mathematics and Informatics, Vilnius University*
[1]*Institute for Mathematics and Informatics, Vilnius University*
Vilnius, LITHUANIA
e-mail: [1]minkevicius.saulius@gmail.com, [2]edvinas.greicius@gmail.com

Abstract

The paper is devoted to the analysis of queueing systems in the context of the network and communications theory. We investigate the inequality in an open multiserver queueing network and its applications to the theorems in heavy traffic conditions (fluid approximation, functional limit theorem, and law of the iterated logarithm) for a queue of jobs in an open multiserver queueing network.

1 Statement of the problem and the network model

This paper is devoted to the analysis of queueing systems in the context of the network and communications theory. We investigate the inequality in an open multiserver queueing network and its applications to the theorems in heavy traffic conditions for a queue of jobs in an open multiserver queueing network. Familiar researches made by [1], [2], and others, were used in this paper.

Consider a network of j stations, indexed by $j = 1, 2, \ldots, J$, and the station j has c_j servers, indexed by $(j, 1), \ldots, (j, c_j)$. A description of the primitive data and construction of processes of interest are the focus of this section. No probability space will be mentioned in this section, and of course, one can always think that all the variables and processes are defined on the same probability space.

First, $\{u_j(e), e \geq 1\}, j = 1, 2, \ldots, J$, are J sequences of exogenous interarrival times, where $u_j(e) \geq 0$ is the interarrival time between the $(e-1)$-st job and the e-st job that arrive at the station j exogenously (from the outside of the network). Define $U_j(0) = 0, \quad U_j(n) = \sum\limits_{e=1}^{n} u_j(e),\ n \geq 1$ and $A_j(t) = \sup\{n \geq 0 : U_j(n) \leq t\}$, where $A_j = \{A_j(t), t \geq 0\}$ is called an exogenous arrival process of the station j, i.e., $A_j(t)$ counts the number of jobs that arrived at the station j from the outside of the network.

Second, $\{v_{jk_j}(e), e \geq 1\},\ j = 1, 2, \ldots, J,\ k_j = 1, 2, \ldots, c_j$, are $c_1 + \ldots + c_J$ sequences of service times, where $v_{jk_j}(e) \geq 0$ is the service time for the e-th customer served by the server k_j of the station j. Define $V_{jk_j}(0) = 0, \quad V_{jk_j}(n) = \sum\limits_{e=1}^{n} v_{jk_j}(e),\ n \geq 1$ and $x_{jk_j}(t) = \sup\{n \geq 0 : V_{jk_j}(n) \leq t\}$, where $x_{jk_j} = \{x_{jk_j}(t), t \geq 0\}$ is called a service process for the server k_j at the station j, i.e., $x_{jk_j}(t)$ counts the number of services completed by the server k_j at the station j during the server's busy time. We define $\mu_{jk_j} = \left(M\left[v_{jk_j}(e)\right]\right)^{-1} > 0,\ \sigma_{jk_j} = D\left(v_{jk_j}(e)\right) > 0$ and $\lambda_j = \left(M\left[u_j(e)\right]\right)^{-1} > 0$, $a_j = D\left(u_j(e)\right) > 0,\ j = 1, 2, ..., k$; with all of these terms assumed finite.

Also, let $\tilde{\tau}_j(t)$ be the total number of jobs routed to the jth station of the network in the interval $[0,t]$, $\tau_j(t)$ be the total number of jobs after service departure from the jth station of the network in the interval $[0,t]$, $\tilde{\tau}_{jk_j}(t)$ be the total number of jobs routed to the k_j server of the jth station of the network in the interval $[0,t]$, let $\tau_{jk_j}(t)$ be the total number of jobs after service departure from the k_j server of the jth station of the network in the interval $[0,t]$, and $\tau_{ijk_i}(t)$ be the total number of jobs after service departure from the k_i server of the ith station of the network and routed to the k_j server of the jth station of the network in the interval $[0,t]$. Let p_{ij} be a probability of the job after service at the ith station of the network routed to the jth station of the network. Denote $p^t_{ijk_i} = \dfrac{\tau_{ijk_i}(t)}{\tau_{ik_i}(t)}$ as part of the total number of jobs which, after service at the k_i server of the ith station of the network, are routed to the jth station of the network in the interval $[0,t]$, $i,j = 1,2,\ldots,J$, $k_i = 1,\ldots,c_i$ and $t > 0$.

The processes of primary interest are the queue length process $Q = (Q_j)$ with $Q_j = \{Q_j(t), t \geq 0\}$, where $Q_j(t)$ indicates the number of jobs at the station j at time t. Now we introduce the following processes $Q_{jk_j} = \{Q_{jk_j}(t),\ t \geq 0\}$, where $Q_{jk_j}(t)$ indicates the number of customers waiting to be served by the server k_j of the station j at time t; clearly, we have $Q_j(t) = \sum\limits_{k_i=1}^{c_j} Q_{jk_i}(t),\ j = 1,2,\ldots,J$.

The dynamics of the queueing system (to be specified) depends on the service discipline at each service station. To be more precise, "first come, first served" (FCFS) service discipline is assumed for all J stations. When a customer arrives at a station and finds more than one server available, it will join one of the servers with the smallest index. We assume that the service station is work-conserving; namely, not all servers at a station can be idle when there are customers waiting for service at that station. In particular, we assume that a station must serve at its full capacity when the number of jobs waiting is equal to or exceeds the number of servers at that station. Suppose that the queue of jobs in each station of the open queueing network is unlimited. All random variables are defined on one common probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

2 The main results

First, denote $\beta_j = \sum\limits_{i=1}^{J}\sum\limits_{k_i=1}^{c_i} \mu_{ik_i}\cdot p_{ij} + \lambda_j - \sum\limits_{k_j=1}^{c_j} \mu_{jk_j} > 0,\ \hat{\sigma}_j^2 = \sum\limits_{i=1}^{J}\sum\limits_{k_i=1}^{c_i} \mu_{ik_i}^3 \cdot \sigma_{ik_i}\cdot p_{ij}^2 + \lambda_j^3 \cdot a_j + \sum\limits_{k_j=1}^{c_j} \mu_{jk_j}^3 \cdot \sigma_{jk_j} > 0,\ j = 1,2,\ldots,J.$

We assume that the following conditions are fulfilled:

$$\sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \mu_{ik_i}\cdot p_{ij} + \lambda_j > \sum_{k_j=1}^{c_j} \mu_{jk_j},\ j = 1,2,\ldots,J. \tag{1}$$

Theorem 1. *If $Q_j(0) = 0$, then*

$$|Q_j(t) - \hat{x}_j(t)| \le \sum_{i=1}^{k} w_i(t) + \sum_{i=1}^{k} \gamma_i(t), \quad where \;\; \hat{x}_j(t) =$$

$$\sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t)\cdot p_{ij} + A_j(t) - \sum_{k_j=1}^{c_j} x_{jk_j}(t), \quad w(t) = \sum_{j=1}^{J}\sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t)\cdot |p^t_{ijk_i} - p_{ij}|,$$

$$\gamma(t) = \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \sup_{0\le s\le t} (x_{ik_i}(s) - \tau_{ik_i}(s)), \; j = 1, 2, \ldots, J.$$

Proof. By definition of the queue of customers at the stations of the network, we get that, for $j = 1, 2, \ldots, J, \; k_j = 1, 2, \ldots, c_j$

$$\begin{aligned}
Q_j(t) &= \tilde{\tau}_j(t) - \tau_j(t) = \sum_{k_i=1}^{c_j} Q_{ik_i}(t) = \sum_{k_i=1}^{c_j} \tilde{\tau}_{ik_i}(t) - \sum_{k_i=1}^{c_j} \tau_{ik_i}(t) \\
&= \sum_{k_i=1}^{c_j} \tilde{\tau}_{ik_i}(t) - \sum_{k_i=1}^{c_j} x_{ik_i}(t) + \sum_{k_i=1}^{c_j} x_{ik_i}(t) - \sum_{k_i=1}^{c_j} \tau_{ik_i}(t) \\
&\le \sum_{k_i=1}^{c_j} \tilde{\tau}_{ik_i}(t) - \sum_{k_i=1}^{c_j} x_{ik_i}(t) + \sum_{k_i=1}^{c_j} \sup_{0\le s\le t} (x_{ik_i}(s) - \tau_{ik_i}(s)) \\
&= \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \tau_{ijk_i}(t) + A_j(t) - \sum_{k_i=1}^{c_j} x_{ik_i}(t) + \sum_{k_i=1}^{c_j} \sup_{0\le s\le t} (x_{ik_i}(s) - \tau_{ik_i}(s)) \\
&\le \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \tau_{ik_i}(t)\cdot \frac{\tau_{ijk_i}(t)}{\tau_{ik_i}(t)} + A_j(t) - \sum_{k_j=1}^{c_j} x_{jk_j}(t) + \sum_{k_i=1}^{c_j} \sup_{0\le s\le t} (x_{ik_i}(s) - \tau_{ik_i}(s))
\end{aligned}$$

$$\begin{aligned}
&\le \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t)\cdot p^t_{ijk_i} + A_j(t) - \sum_{k_j=1}^{c_j} x_{jk_j}(t) + \sup_{0\le s\le t} (x_{jk_j}(s) - \tau_{jk_j}(s)) \\
&= \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t)\cdot (p^t_{ijk_i} - p_{ij} + p_{ij}) + A_j(t) - \sum_{k_i=1}^{c_j} x_{ik_i}(t) \\
&\le \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t)\cdot p_{ij} + A_j(t) - \sum_{k_i=1}^{c_j} x_{ik_i}(t) + \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t)\cdot |p^t_{ijk_i} - p_{ij}| \\
&+ \sum_{k_i=1}^{c_j} \sup_{0\le s\le t} (x_{ik_i}(s) - \tau_{ik_i}(s)) = \hat{x}_j(t) + w(t) + \gamma(t), j = 1, 2, \ldots, J \text{ and } t > 0.
\end{aligned}$$

Hence it follows that

$$Q_j(t) \le \hat{x}_j(t) + w(t) + \gamma(t), \; j = 1, 2, \ldots, J \text{ and } t > 0. \tag{2}$$

Also, note that

$$Q_j(t) \geq \tilde{\tau}_j(t) - \sum_{k_i=1}^{c_j} x_{ik_i}(t) = \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \tau_{ik_i}(t) \cdot p_{ijk_i}^t + A_j(t) - \sum_{k_i=1}^{c_j} x_{ik_i}(t)$$

$$= \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} (x_{ik_i}(t) + \tau_{ik_i}(t) - x_{ik_i}(t)) \cdot p_{ijk_i}^t + A_j(t) - \sum_{k_i=1}^{c_j} x_{ik_i}(t)$$

$$= \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot p_{ijk_i}^t + \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} (\tau_{ik_i}(t) - x_{ik_i}(t)) \cdot p_{ijk_i}^t + A_j(t)$$

$$- \sum_{k_i=1}^{c_j} x_{ik_i}(t) = \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot p_{ijk_i}^t - \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} (x_{ik_i}(t) - \tau_{ik_i}(t)) \cdot p_{ijk_i}^t$$

$$+ A_j(t) - \sum_{k_j=1}^{c_j} x_{jk_j}(t) \geq \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot p_{ijk_i}^t + A_j(t) - \sum_{k_j=1}^{c_j} x_{jk_j}(t)$$

$$- \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} (x_{ik_i}(t) - \tau_{ik_i}(t)) \geq \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot p_{ijk_i}^t + A_j(t) - \sum_{k_j=1}^{c_j} x_{jk_j}(t)$$

$$- \sup_{0\leq s\leq t} \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} (x_{ik_i}(s) - \tau_{ik_i}(s)) \geq \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot p_{ijk_i}^t + A_j(t)$$

$$\begin{aligned}
&- \sum_{k_j=1}^{c_j} x_{jk_j}(t) - \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \sup_{0\leq s\leq t} (x_{ik_i}(s) - \tau_{ik_i}(s)) \\
&= \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot (p_{ijk_i}^t - p_{ij} + p_{ij}) + A_j(t) - \sum_{k_j=1}^{c_j} x_{jk_j}(t) \\
&- \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \sup_{0\leq s\leq t} (x_{ik_i}(s) - \tau_{ik_i}(s)) = \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot p_{ij} + A_j(t) \\
&- \sum_{k_i=1}^{c_j} x_{ik_i}(t) + \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot (p_{ijk_i}^t - p_{ij}) \\
&- \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \sup_{0\leq s\leq t} (x_{ik_i}(s) - \tau_{ik_i}(s)) \geq \hat{x}_j(t) - \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} x_{ik_i}(t) \cdot |p_{ijk_i}^t - p_{ij}| \\
&- \sum_{i=1}^{J}\sum_{k_i=1}^{c_i} \sup_{0\leq s\leq t} (x_{ik_i}(s) - \tau_{ik_i}(s)) = \hat{x}_j(t) - w(t) - \gamma(t),\ j = 1, 2, \ldots, J \text{ and } t > 0.
\end{aligned} \tag{3}$$

Hence it follows that

$$Q_j(t) \geq \hat{x}_j(t) - \sum_{i=1}^{k} w_i(t) - \sum_{i=1}^{k} \gamma_i(t), \tag{4}$$

$j = 1, 2, \ldots, J$ and $t > 0$.

By combining (2) and (4), we can prove an inequality. □

3 Application of the inequality

Note that the inequality is the key inequality to prove several laws (fluid approximation, functional limit theorem, and law of the iterated logarithm) for a queue of jobs in open multiserver queueing networks in heavy traffic conditions. At first we present a theorem on the fluid approximation for a queue of jobs in open multiserver queueing networks under heavy traffic conditions.

Theorem 2. *Under conditions (1) the weak convergence holds:*

$$t^{-1}(Q_j(t))_{j=1}^J \Rightarrow (\beta_j)_{j=1}^J,\ 0 \le t \le 1.$$

Next, we present a theorem on the functional limit for a queue of jobs in open multiserver queueing networks in heavy traffic conditions.

Theorem 3. *Under conditions (1) the following CLT holds:*

$$n^{-1/2}\left(\frac{Q_j(nt) - nt\beta_j}{\hat{\sigma}_j}\right)_{j=1}^J \Rightarrow (W_j(t))_{j=1}^J,\ 0 \le t \le 1,$$

for independent standard Wiener processes $W_j(t)$, $j = 1, \ldots, J$.

One of the results of the paper is the following theorem on the law of the iterated logarithm for a queue of jobs in an open multiserver queueing network.

Theorem 4. *Under conditions (1) the following law of the iterated logarithm holds:*

$$P\left(\overline{\lim_{t\to\infty}} \frac{Q_j(t) - t\beta_j}{\hat{\sigma}_j\sqrt{2t\ln\ln t}} = 1\right) = 1,\ j = 1, \ldots, J.$$

The proof of these theorems is similar to that in [3].

References

[1] Dai J.G., Dai W. (1999). A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Systems*. Vol. **32**(1-3), pp. 5–40.

[2] Peterson W.P. (1991). A heavy traffic limit theorem for networks of queues with multiple customer types. *Math. Operations Research*. Vol. **16**(1), pp. 90–118.

[3] Sakalauskas L.L., Minkevičius S. (2000). On the law of the iterated logarithm in open queueing networks. *European J. Operational Research*. Vol. **120**, pp. 632–640.

COMPUTER ANALYSIS OF ESSENTIAL HYPERTENSION RISK ON THE BASE OF GENETIC AND ENVIRONMENTAL FACTORS

O. S. Pavlova[1], V. I. Malugin[2], S. E. Ogurtsova[3], A. Yu. Novopoltsev[2], I. S. Byk[2], T. V. Gorbat[1], M. M. Liventseva[1], A. G. Mrochek[1]
[1]*Republican Scientific and Practical Centre "Cardiology"*
[2]*Dep. Mathematical Modeling and Data Analysis, Belarusian State University*
[3]*Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus*
Minsk, BELARUS
e-mail: ol_skorochod@yahoo.com, malugin@bsu.by

Abstract

In this paper, the computer software implementing the model of the gene-gene and gene-environment interaction among the polymorphisms of nine candidate genes relating to essential hypertension (EH) and environmental cardiovascular risk factors, such as obesity, abdominal obesity, smoking and insufficient physical activity, is presented. Five significant genetic patterns for hypertensive patients were determined using APSampler software based on the Markov Chain Monte-Carlo method with a specially adapted Metropolis – Hastings algorithm. The binary logit model with high sensitivity demonstrated the cumulative effects for the multilocus combinations and environmental factors associated with EH. This model allows the classification of subjects into two classes: the healthy and the hypertensive patients. Performance evaluation of the model by means of statistical tests indicates an acceptable accuracy of classification and prediction.

1 Introduction

Essential hypertension (EH) is a complex disorder influenced by multiple genetic and environmental factors [1]. The renin-angiotensin-aldosterone system (RAAS), vascular endothelial and kallikrein-kinin systems have a vital role in the blood pressure (BP) regulation and the pathogenesis of EH. The single nucleotide polymorphisms (SNP) in genes encoding these systems are associated with EH that was showed in the previous studies [2]. But these associations are often not reproducible and depend on ethnicity, and environmental factors, such as smoking, overuse of alcohol, insufficient physical activity (PA), obesity, abdominal obesity (AO), psychological stress and depression also influence the development of EH. The interaction of the mutations candidate genes and environmental factors may substantially increase susceptibility to EH. The objectives of our study were to examine the gene-gene interaction among the polymorphisms of the nine candidate genes of RAAS, vascular endothelial and kallikrein-kinin systems — angiotensin-converting enzyme (ACE), angiotensinogen (AGT), angiotensin II type 1 and 2 receptor (AGTR1, AGTR2), aldosterone synthase gene (CYP11B2), renin (REN), 2-bradykinin receptor gene (BKR2), methylenetetrahydrofolate reductase (MTHFR), endothelial nitric oxide synthase (eNOS) and gene-environment interaction in patients (pts) with EH.

2 Methods and Software

A total of 532 subjects are included (356 hypertensive pts and 176 normotensives ones). Genotyping for ACE-I/D, AGT-235, AGTR1-A1166C, AGTR2-C3123A, CYP11B2-C344T, REN-19-83G/A, BKR2-T58C, eNOS-E298D, and MTHFR-C677T polymorphisms was performed by polymerase chain reactions and the restriction of digestion. The following environmental (biological, behavioral and psychosocial) factors were assessed: office BP, weight, AO, smoking, alcohol consumption, physical activity, as well as the level of psychological stress and depression using the international scales (Psychological Stress Measure (PSM-25) and Center for Epidemiologic Studies Depression scale (CES-D), accordingly). The study included a search of polygenic associations to determine the genetic pattern, i.e. combination of alleles or genotypes of different loci associated with a phenotypic trait (the gene-gene interaction analysis) using the APSampler software based on Markov chains (Markov Chain Monte-Carlo — MCMC method) with a specially adapted Metropolis – Hastings algorithm [3, 4]. Then a binary logit model was built to estimate the cumulative effects of the genetic and environmental factors associated with EH (the gene-environment interaction analysis). This model allows the classification of subjects into two classes: the healthy and the hypertensive patients. Performance evaluation of the model by means of statistical tests indicates an acceptable accuracy of classification and prediction.

A computer program was created as a result of all previous investigation. It allows to interrogate pts maintaining and accumulating their visit data, classify them using any applicable to a particular patient model. Binary logit models with different set of explanatory variables and accuracy are stored and could be added or updated with the accumulation of new pts data. This tool, on the one hand, helps collect data for future analysis of new combinations of genes, important factors to improve the classification models, and on the other, provides an opportunity to demonstrate to a patient his risk of developing EH in his current state and how it could change through adjusting lifestyle factors (Figure 1).

3 Results

The significant genetic patterns for the examined groups in patients with EH and healthy subjects were obtained (Table 1). The binary dependent variable logit model [4] was constructed to explore the cumulative effects of the significant multilocus combinations and environmental (biological, behavioral and psychosocial) factors. The results of estimation and testing the model (Table 2) indicate the statistical significance of all factors at a level near 0.05 and below, as well as the adequacy of the model as a whole.

Expectation-prediction evaluation of the model based on the classification tables gives the following estimates of accuracy of the classification: overall, the estimated model correctly predicts 74.08% of the observations; the percentage of correct decisions in the classification of hypertensive patients is 86.42%.

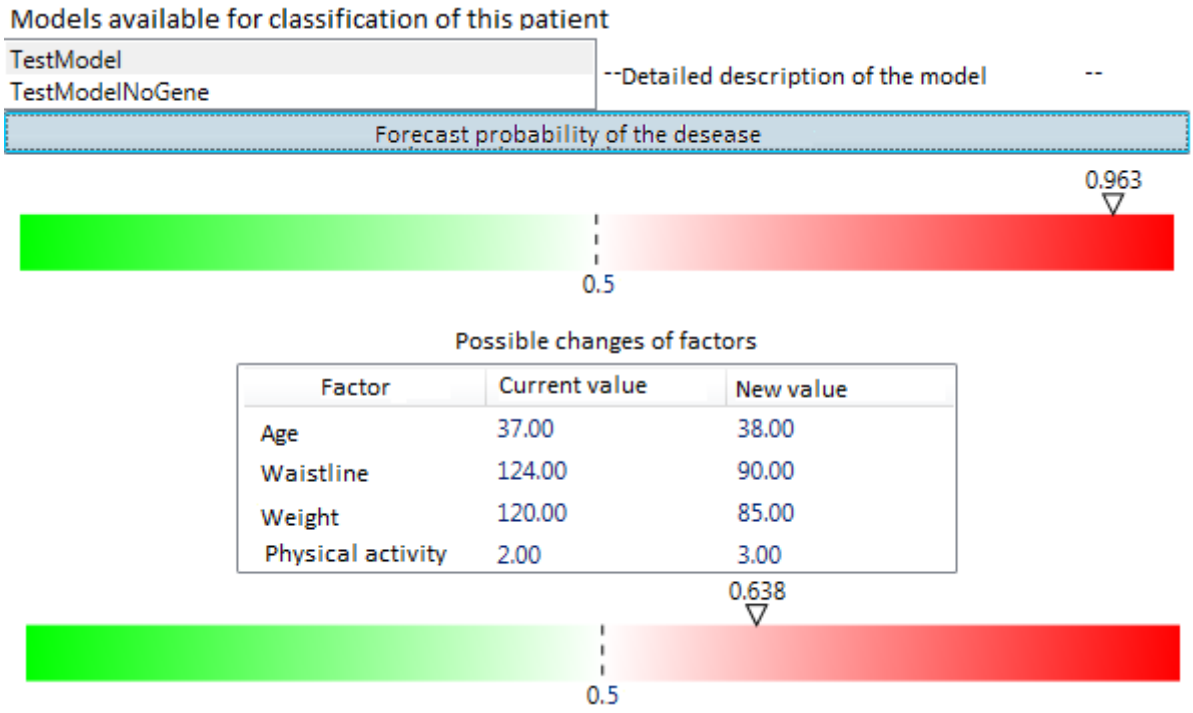


Figure 1: The current estimation results for a patient on the base of binary logit model and possible result after 1 year of therapy

Table 1: Significant genetic patterns for the hypertensive group

#	**Loci combinations**
1.	T allele AGT-235/AA genotype AGTR2-C3123A
2.	T allele AGT-235/T allele CYP11B2-C344T
3.	TT genotype AGT -235/D allele ACE-ID
4.	D allele eNOS-E298D/C allele BKR2-T58C
5.	D allele eNOS-E298D/ D allele ACE-ID

Table 2: Estimation results for the binary logit model

Variables	β**-coefficients**	**Std. Error**	z**-statistics**	p**-value**
Age	0.928626	0.224505	4.136318	0.0000
Smoking	0.438137	0.220897	1.983440	0.0473
AO	0.723835	0.273650	2.645118	0.0082
BMI	1.380139	0.310520	4.444605	0.0000
Insufficient PA	0.631120	0.218890	2.883269	0.0039
T allele AGT-235/AA genotype AGTR2-C3123A	0.491607	0.253880	1.936374	0.0528
D allele eNOS-E298D/ D allele ACE-ID	0.713285	0.223251	3.194985	0.0014
C	-1.295533	0.253726	-5.106024	0.0000
McFadden R-squared	0.185426	LR statistic (p-value)		121.6938 (0.0000)

4 Conclusions

A software that realizes the binary logit model which reflects the cumulative effects of the multilocus combinations and environmental factors associated with EH is developed. The software is intended for the classification of patients into two classes: the healthy and the hypertensive ones. Performance evaluation of the implemented model by means of statistical tests indicates an acceptable accuracy of classification and prediction.

References

[1] Lifton R.P., Gharavi A.G., Geller D.S. (2001). Molecular mechanisms of human hypertension. *Cell*. Vol. **104**(4), pp. 545–556.

[2] Abbate R., Sticchi E., Fatini C. (2008). Genetics of cardiovascular disease. *Clinical Cases Miner Bone Metab*. Vol. **5**(1), pp. 63-66.

[3] Favorov A.V., Andreewski T.V., Sudomoina M.A., Favorova O.O., Parmigiani G., Ochs M.F. (2005). A Markov Chain Monte Carlo technique for identification of combinations of allelic variants underlying complex diseases in Humans. *Genetics*. Vol. **171**(4), pp. 2113-2121.

[4] Gilks W.R., Richardson S., Spiegelhalter D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

BIOINFORMATICS ANALYSIS OF M.TUBERCULOSIS WHOLE-GENOME SEQUENCES

R. S. Sergeev[1], I. S. Kavaliou, A. V. Tuzikov, M. V. Sprindzuk
United Institute of Informatics Problems
Minsk, BELARUS
e-mail: [1]roma.sergeev@gmail.com

Abstract

Here we consider analysis of variations in *M.tuberculosis* genomes and discover methods for genome-wide association studies which allows identifying drug-resistance mutations in microorganisms. Alterations in genomes are among the main mechanisms by which pathogens exhibit drug resistance. Analysis of the reported cases and discovery of resistance-associated genetic markers may contribute greatly to the development of new drugs and effective therapy management.

1 Introduction

Two billion people worldwide are thought to be infected with latent form of tuberculosis (TB). Around 5-10% of these individuals will develop an active TB disease. There are big challenges associated with emergence and development of multi drug-resistant (MDR) and extensively drug-resistant (XDR) tuberculosis [1]. While MDR-TB is difficult and expensive to treat, XDR-TB is virtually an untreatable disease in most of the developing countries.

There are several lines of drugs applied in basic antituberculosis therapy. First-line drugs are used in standard course of treatment for newly diagnosed TB cases or fully sensitive organisms, while treatment of resistant TB requires many different drugs and therapeutic approaches. The most common causative agent of tuberculosis in humans is *Mycobacterium tuberculosis*. MDR-TB is defined as *M. tuberculosis* strains resistant at least to the two most important first-line drugs: rifampicin (RIF) and isoniazid (INH). XDR-TB is MDR-TB strains additionally resistant to a fluoroquinolone (FLQ) and a second-line anti-TB injectable agent such as kanamycin (KANA), amikacin (AMIK), or capreomycin (CAPR).

M. tuberculosis has devoted a large part of its genome towards functions that allow it to successfully establish progressive or latent infection in the majority of infected individuals. We analyzed 136 tuberculosis whole-genomes from Belarus for the known high-confidence drug-resistance mutations presented in TBDreamDB database [2] and investigated statistical methods for searching genetic variations that explain the observed resistance to the first-line and second-line drugs.

2 Materials and Methods

We included 17.7% of drug-sensitive, 10.3% of MDR, 22.1% of preXDR, 27.2% of XDR and 19.9% of totally drug-resistant *M. tuberculosis* strains in the analysis. Illumina HiSeq2000 instruments was used for sequencing at the Broad Institute (USA). Two sequencing libraries were created to capture genetic variations for each sample: fragment library with 180bp insert size and jumping library with 3-5kb insert size. Paired-end reads were mapped to H37Rv reference genome (GeneBank accession NC_018143.2) to identify variants. Pilon software [3] was used for variant detection to capture single-nucleotide polymorphisms (SNPs) and indels. We have repeated mapping and variant calling steps applying different reference genomes that represented most common genetic families.

We used RAxML software [4] to reconstruct the phylogeny under general time reversible model. We estimated single nucleotide polymorphisms arising in the phylogeny using maximum-likelihood ancestral site reconstruction.

For digital spoligotyping the reads were mapped against 43 spacer sequences and frequency was tallied. Background null model for the expected coverage was made from the total sequencing data using an exponential distribution under the assumption that more than 90% of the reads align. The Benjamini-Hochberg correction was applied to p-values calculated for frequency and when significant, the marker was reported as present.

Finally, we performed association analysis to check for significant differences between DNA sequences isolated from drug-resistant (cases) and drug-susceptible (controls) organisms. Here we first focused on 24 candidate genes and their promoter regions selected based on extensive literature mining. To characterize every reported SNP, we showed how many true positives, true negatives, false positives and false negatives occurred for each drug, assuming that the ideally discriminating SNP would be found in 100% of resistant genomes and 0% of sensitive genomes (or vice versa).

We applied a series of tests from multifactor statistical analysis to investigate significant SNPs in the whole genomes: regularized logistic regression [5], linear mixed model (LMM) [6] and mode-oriented stochastic search (MOSS) [7]. Elastic net regularization showed most relevant results in logistic regression approach encouraging a grouping effect, when strongly correlated predictors tend to occur together in a produced sparse model. LMM allowed correction for population structure due to the random effect of the linear mixed model that was calculated based on kinship/relatedness matrix. The MOSS algorithm, which is a Bayesian variable selection procedure, identified combinations of the best predictive SNPs associated with the response after searching for the best hierarchical log-linear models with the number of factors $k \in [2, 5]$. We used cross-validation (CV) procedure for parameter tuning and calculating generally accepted metrics to evaluate the quality of predictions: precision, recall, F-measure (the weighted harmonic mean of precision and recall) and accuracy.

However, results of drug-association tests may include pairs of correlated mutations with high and low association scores within the same pair. We used feature relevance network (FRN) [8] to take into account correlations between mutated sites so that any

highly correlated mutations should be either significant or not significant. Pearson coefficients of correlation were used to estimate linkages between loci. The algorithm searches for a minimum cut in a graph of a special structure that reflects relationships in data. Eventually this graph cut splits SNPs into subsets of significant and non-significant. We used F-measure within a validation procedure to measure accuracy of classification after FRN correction.

3 Results

Genotype/phenotype association tests have resulted in the lists of high-confidence mutations associated with drug resistance. Mutations were ordered according to their significance score for each drug and annotated using NCBI database. MOSS and LMM methods provided the smallest number of significant variations with sufficiently good classification quality. Regularized logistic regression showed the best results but produced much larger lists of drug-associated SNPs, even after the second run on the reduced SNP lists, which might indicate inclusion of noise characteristics in the resulting sets. However, most methods agreed in assigning the highest scores to the genetic markers probed by the Genotype MTBDRplus, MTBDRsl (HAIN Lifesciences, Germany) line-probe assays (Table 1). After FRN-based selection procedure, long mutation lists were halved for most genotype/phenotype association tests without serious loss in F-measure of classification.

Table 1: Top significant drug-resistance mutations according to each method

Drug	SNP position	RFN-supported	MOSS: inclusion probability	LMM: Bonferroni corrected p-value	Logistic regression: normalized coefficient	Annotation
INH	1673431	+	-	3.72×10^{-3}	0.09	inhA promoter
INH, RIF	2155175	+	1.00	5.84×10^{-15}	1.00	katG
RIF, INH	761158	+	-	1.28×10^{-18}	0.52	rpoB
INH, RIF	3842469, 3842475	+	-	1.29×10^{-2}	0.84	PPE family protein PPE57
FLQ	7570	+	1.00	3.73×10^{-1}	0.83	gyrA
FLQ	7582	+	1.00	4.0×10^{-3}	1.00	gyrA
AMIK	1473252	+	1.00	1.4×10^{-18}	1.00	rrs
CAPR	1473252	+	1.00	3.67×10^{-10}	1.00	rrs

We have run the association tests for the second time for each drug where known high-confidence resistance mutations were excluded from the analysis. This approach

was intended to determine more robust scores for mutations which appeared not very significant but correlated to the most significant SNPs. We showed that the exclusion of a few significant SNPs did not change the other scores dramatically. For example, the MOSS algorithm populated resulting sets of most promising log-linear models with a bulk of middle-quality models and lower SNP-inclusion probabilities. Correction of the significance scores using feature relevance network after the second run demonstrated a notable loss in predictive power of the remained SNPs in comparison to the first run.

4 Conclusion

We discovered methods for genome-wide association studies and developed an approach that allows identifying drug-resistance mutations in microorganisms. We analysed 136 tuberculosis whole-genomes from Belarus to investigate SNPs asociated with resistance to the most important drugs. This approach is used in the current research project to establish the Belarus tuberculosis portal (http://tuberculosis.by) and conduct comprehensive study of the obtained MDR and XDR TB strains. Prototype of the genomic portal for processing tuberculosis data is available at http://tb-portal.esy.es.

References

[1] World Health Organization (2015). *Global Tuberculosis Report 2015*. Geneva.

[2] Sandgren A. et al. (2009). Tuberculosis Drug Resistance Mutation Database. *PLoS Med*. Vol. **6**(2), pp. 132–136.

[3] Walker B.J. et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One*. Vol. **9**(11).

[4] Stamatakis A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*. Vol. **30**(9).

[5] Zou H., Hastie T. (2005). Regularization and Variable Selection via the Elastic Net. *J. Royal Statistical Society. Series B*. Vol. **67**(2). pp. 301–320.

[6] Zhou X., Stephens M. (2012). Genome-wide Efficient Mixed Model Analysis for Association Studies. *Nature Genetics*. Vol. **44**, pp. 821–824.

[7] Dobra A. et al. (2010). The Mode Oriented Stochastic Search (MOSS) for Log-Linear Models with Conjugate Priors. *Stat. Methodology*. Vol. **7**, pp. 240–253.

[8] Liu J. et al. (2012). High-Dimensional Structured Feature Screening Using Binary Markov Random Fields. *JMLR Workshop Conf Proc. 2012*. Vol. **22**, pp. 712–721.

FRACTAL DIMENSION AS A CHARACTERISTIC OF BIOLOGICAL CELL AFM IMAGES

I. E. Starodubtsev
IBA Gomel Park
Gomel, BELARUS

1 Introduction

Biological cells are open, dynamic self-organizing microsystems that exchange matter, energy and information with their environment. Performing their physiological functions, the biological cells interact with other cells, vessel walls and macromolecular complexes when as a rule they are subjected to mechanical stress.

The cell surface properties including its structural and mechanical properties are important parameters of cell state and functioning. Because the change of the cell surface properties occurs during the pathological processes, the qualitative and quantitative cell surface characteristics can be markers of cell health and pathology.

Atomic force microscopy (AFM) is one of the modern methods for studying solid surface structure and properties. AFM has tremendous advantages over electron microscopy (including scanning electron microscopy), as it allows working with objects directly both on air and in various fluids. Atomic force microscopes are widely used in many fields of science and technology: biophysics, biochemistry, materials science, pharmaceutics, surface physics, electronics and others. Nowadays AFM is used in studying the biological cells and tissues as well.

AFM provides the images of topography (topography scan mode) and spatial distribution of local physical and mechanical properties (torsion scan mode) of the studied surface with nanometer resolution (Figure 1).

AFM-image of a cell surface is a set of points with three coordinates (x, y, z) that represents either a topography map (in this case x, y and z are positions of the surface points) or map of local physical and mechanical properties (in this case x, y are positions of the surface points and z is a force value in the certain point).

The dimension is an important parameter of the surface of objects. Real surfaces are characterized by fractal (fractional) dimension. There are various methods to calculate the fractal dimension: box-counting method, power spectrum method, hand and drives method and others [1]. Each method has its own features that limit its usage.

The work aims at studying the relationship between the fractal dimension and geometrical parameters of AFM images of real biological cells and model surfaces.

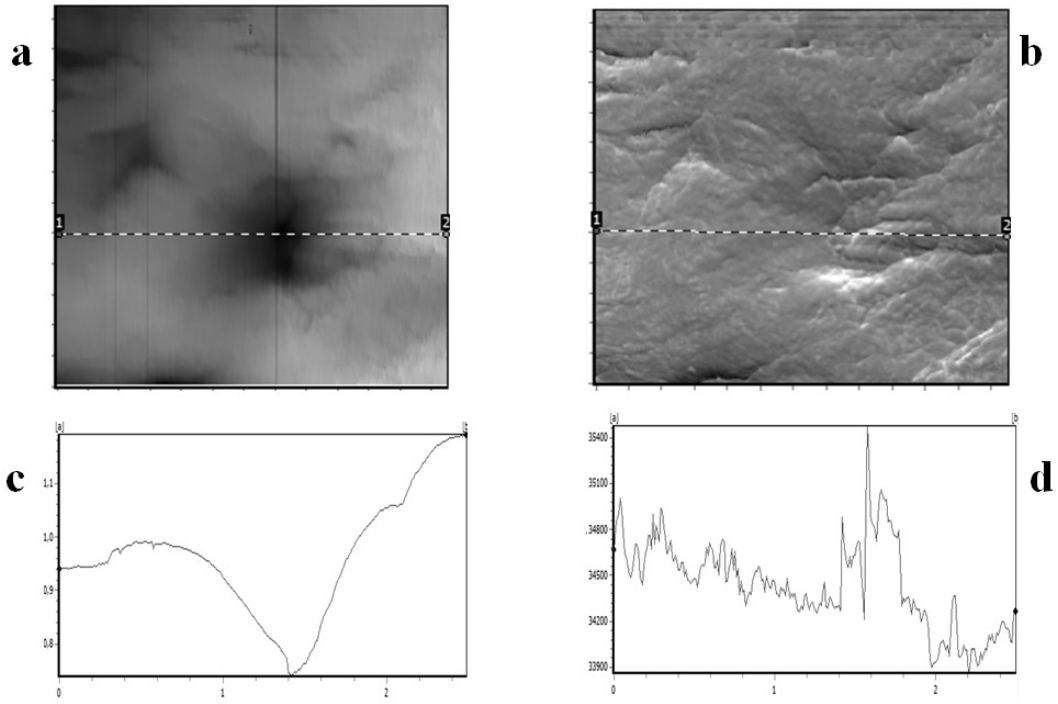


Figure 1: AFM images (a, b) and profiles (c, d) of 549 cancer cell surface (scan size is 2.5μm × 2.5μm).

2 Methods

The main method we used to calculate the fractal dimension is box-counting method [2,3]. It based on the following formula:

$$D_F = -\lim_{\varepsilon \to 0} \log_\varepsilon N(\varepsilon), \tag{1}$$

where $N(\varepsilon)$ is minimal number of cubes with edge ε that cover together the required surface.

To find the fractal dimension (D_F) the system of equations had to be solved:

$$\ln N(\varepsilon) = \ln C + D_F \ln \varepsilon, \tag{2}$$

where the number of equations is larger than the number of unknown variables. The system often has no exact solution and, therefore, is solved numerically.

In the present work, the mentioned above method was realized on C++ programming language using Borland C++ Builder IDE and STL library.

The implementation of the algorithm included the following steps. The spatial region with the studied surface was divided by the cubic lattice with cube edge ε (initially set as a half of the studied region size). Then the number of cubes $N(\varepsilon)$ that included at least one point of the surface was calculated. The cube edge ε was reduced by two and the process repeated in loop until cube edge became less than a constant

depending on AFM scanning step. At each step of loop pairs $\ln N(\varepsilon)$ and $\ln \varepsilon$ were added to resultant array. The plot $\ln N(\varepsilon)$ against $\ln \varepsilon$ was approximated with a line which slope was equal to surface fractal dimension D_F.

We used also the modified box-counting algorithm. The surface was divided into a few (from 2 to 8) equal fragments and the fractal dimensions were calculated for each fragment using box-counting algorithm. Then the fractal dimension for the whole surface was calculated using the sample of D_F and represented as the mean and confidence interval limits.

3 Results

We analyzed the change of fractal dimension with the change of the scale factor for axis Z (Z-scaling). The problem of the change of the object dimension during scaling has been recently reviewed by Simon Villerton in two-dimensional case [4]. In the present work, the analysis of the dependence of the fractal dimension on Z-scaling was performed in the following way: the data along axes X and Y were not changed but the data of axis Z was multiplied by factor t changed over a broad range of values. D_F of the whole surface was calculated for each value of factor t (scaling factor for axis Z):

$$D_F = \varphi(t). \tag{3}$$

Various modeling surfaces have been generated for the qualitative analysis of the dependencies: plane surface, plane surface with a finite number of Gaussian peaks, wave surfaces $Z = H \sin(\omega\sqrt{x^2+y^2})$ and $Z = H|\sin(\omega\sqrt{x^2+y^2})|$. The changes in $D_F = \varphi(t)$ with the changes of frequency, amplitude and other surface parameters were found and analyzed.

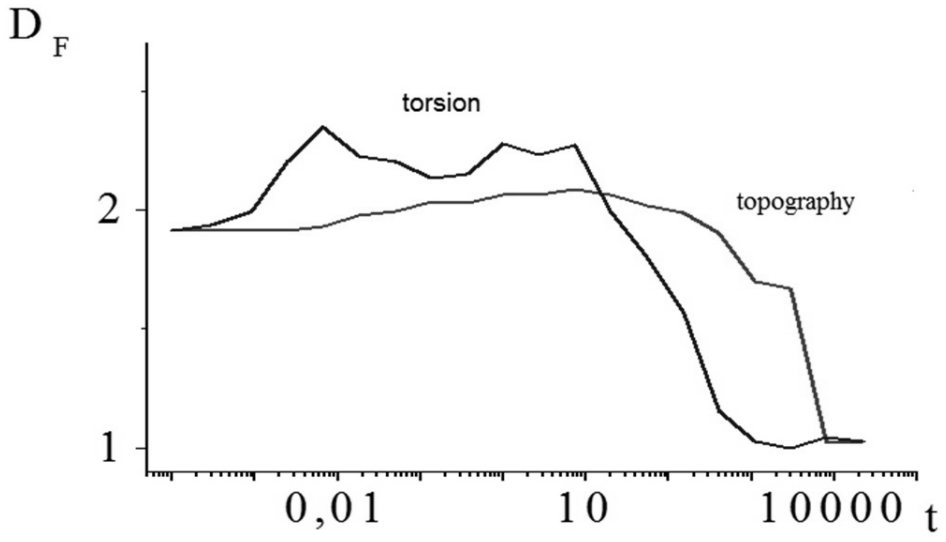


Figure 2: Dependence $D_F = \varphi(t)$ for torsion and topography scans of erythrocyte surface.

For the erythrocyte surface (Figure 2) D_F at the smaller values of t tends to 2 (plane surface) and at the larger values of t tends to 1 (line). In the intermediate range of t, function $D_F = \varphi(t)$ has some maxima. The results of the performed analysis has shown that the parameters of dependence $D_F = \varphi(t)$ was qualitatively related to the type of elements of the surface. For example, if the first peak in curve $D_F = \varphi(t)$ was higher than the second peak, the studied surface had the frequent small-scale heterogeneities, and if the second peak was higher than the first peak, the surface was relatively smooth with a few large-scale heterogeneities.

4 Conclusion

Dependence $D_F = \varphi(t)$ is a characteristic of AFM images of surfaces (including the surfaces of biological cells), which describes the surface features better than a single value D_F.

References

[1] Napolitano A., Ungania S., Cannata V. (2012). Fractal dimension estimation methods for biomedical images. *MATLAB - A fundamental tool for scientific computing and engineering applications*. Vol. **3**, pp. 161–178.

[2] Braverman B., Tambasco M. (2013). Scale-specific multifractal medical image analysis. *Computational and Mathematical Methods in Medicine*. ID 262931.

[3] Foroutan-pour K., Dutilleul P., Smith D. L. (1999). Advances in the implementation of the box-counting method of fractal dimension estimation. *Applied Mathematics and Computation*. Vol. **105**(2), pp. 195–210.

[4] Willerton S. (2015). Spread: a measure of the size of metric spaces. *Intern. J. Comput. Geom. Appl.* Vol. **25**(3), pp. 207–225.

KNOWLEDGE REPRESENTATION AND REASONING. MIVAR TECHNOLOGIES

O. O. Varlamov[1], I. A. Danilkin[2]
MIVAR LLC
Moscow, RUSSIA
e-mail: [1]o.varlamov@mivar.ru, [2]i.danilkin@mivar.ru

Abstract

The technology of knowledge representation based on mivar network and its matrix representation are considered. The mechanism that allows to design computational algorithms of solutions of assigned tasks on the basis of data from mivar networks is described.

1 Introduction

Development of artificial intelligence (AI) systems is an important and quite a topical task of today. Such system types as expert systems, text meaning understanding, image recognition, robotic systems aim to change the life of a modern man. However, it should be noted that currently available intelligent systems are designed to support regular user groups to solve highly specialized tasks. Thus, design of theory, methods and technologies of AI remains an urgent task for development of intelligent systems. Moreover, it becomes increasingly important [3].

As mentioned above, available intelligent systems aim to solve highly specialized tasks, since building and using knowledge bases that provide the foundation for such systems requires costly human and material resources. Thus, the developers of intelligent systems face a wide range of difficulties. Knowledge representation and search are the two fundamental problems that still occupy developers of intelligent systems.

Knowledge representation models should provide a simple mechanism of data description and development of the knowledge bases which are required to implement intelligent behavior of such systems. On the one hand, representation method should make the knowledge understandable for machine, on the other hand it should ensure easy description of knowledge structure for its developer. Therefore, in the process of development different knowledge representation models, two aims are pursued: expressiveness and efficiency. Moreover, such systems should ensure the most natural way of knowledge representation. Nowadays, there is a large number of approaches to data representation: predicate description, semantic networks, production rules, as well as neural-networking, evolutionary, agent-oriented and stochastic approaches to representation, and many others [2]. All these approaches aim to reach a fair compromise between efficiency and expressiveness of representation.

In this paper mivar-based technology of knowledge representation is considered. It is aimed to simplify knowledge acquisition and accumulation since there is no necessity in experts involvement and logical inference methods changing [1, 4]. More than that, mivar-based technology of data processing is proposed, which aims to increase the

speed and quality of acquiring results. Since the problems of knowledge representation and search are interconnected, an intelligent system should not only be aware of the subject, but also be able to solve tasks set in the subject domain.

Mivar technologies were used in development of such AI systems as: text meaning understanding, image recognition, robotic systems and expert systems.

2 Mivar knowledge representation technologies

The concept of mivar network is one of the basic concepts of the proposed mivar-based approach to data representation. It is mivar network that ensures formalization and representation of human knowledge. Let us consider a subject domain M.

Mivar network - the method of representing objects of the subject domain and rules of their processing in the form of a bipartite directed graph consisting of objects (P) and rules (R). These objects and rules together form the model of the subject domain.

Mivar network has the following significant properties:

1. The network consists of the elements of two types (two partitions of the graph): the nodes – objects (P) and the arcs – rules (R).

2. For each variable all the information is stored in explicit form about all the rules R, for which it is an input variable X or an output variable Y with the indication of that;

3. For each rule R information about all its input and output variables P is stored in explicit form, including the information about the number of input (X) and output (Y) variables;

4. The storage of all the necessary information of such a network is organized on the basis of database technologies adapted to operate with mivar networks;

5. In each element of the mivar network, being node or arc, all the adjacent arcs and nodes are determined coherently and completely. Being in any place of mivar network, it is always obvious from where can we move to it and to where we can move from it, which eliminates brute forcing while searching for logical inference on the mivar network.

Bipartite graph of mivar network can be represented in the form of two-dimensional matrix $(P) \times (R)$, where n is the number of parameters (objects) of the subject domain, m is the number of rules connecting objects (see Table 1).

Mivar network is constructed by binding aggregates of different types according to the following: "object-rule" and "rule-object" . Interconnection such as "object-object" and "rule-rule" are forbidden. In general form the interconnection is as follows: "object(s)-rule-object(s)" . The element from which the interconnection goes is indicated first. The element to which the interconnection moves is indicated second. It allows us to determine direction of the graph and exclude possibilities of misinterpreting or converting objects through backtracking. Designed in such a way, mivar network

Table 1: Matrix representation of mivar network

Rules \ Parameters	1	2	3	4	5	...	$n-2$	$n-1$	n
1	**X**	**X**	**X**					**Y**	**Y**
2			**X**	**Y**	**Y**			**X**	**X**
...						...			
m		**X**		**X**	**X**		**Y**		

is scalable, as at any time it is possible to add elements of aggregates of any types available without the necessity to change processing methods. Moreover, describing mivar network does not require the involvement of an expert. In most cases it is sufficient to move objectively existing objects and connections (rules) in the mivar form.

As an example let us consider the subject domain "Geometry. Triangles" . Here as variables are any sides, angles, segments of the triangle. Rules are different interconnections between these variables such as definitions, theorems, axioms, etc. The part of mivar network of the considered subject domain is represented in the form of the matrix M (see Table 2).

Table 2: Part of the mivar network. Geometry

Rules \ Parameters	side AB	side BC	side AC	Perimeter	...	AB is .. greater than BC	AC is .. greater than BC
The perimeter of the triangle (using the lengths of three sides)	**X**	**X**	**X**				**Y**
The side BC (using the perimeter and the ratios of the sides)		**Y**		**X**		**X**	**X**
The side AC (using the ratio of the sides)		**X**	**Y**				**X**
The side AB (using the ratio of the sides)	**Y**	**X**				**X**	

3 Reasoning

On the basis of subject domain representation described above it is possible to design an algorithm allowing us to implement information search inside the mivar network, set open and hidden interconnections between data inside mivar network, construct

computational algorithms of the set task in corresponding subject domain on the basis of data from mivar network. Such a methodology for searching logical inference path allows us to avoid brute forcing of all possible rules on each step. This algorithm includes forming aggregates of known parameters and setting one or more required parameters, then the processing for each known parameter (which was not processed before) is carried out to find required parameters. This processing involves the following stages:

– determines rules (which were not launched before) in which, firstly, known parameter serves as the input variable and secondly, all other variables are known;

– launch these found rules and add output variables of launched rules to the aggregate of known parameters; moreover, if all the required parameters are found, the processing is stopped;

– design the sequence of launched rules in the order of launching, thus, the designed sequence is logical inference path.

Let us illustrate the basic variant of the scheme described above using a simple example from the subject domain "Geometry. Triangles" . To do this are given steps of the solution of the following task from this subject domain: *It is needed to find the lengths of the sides AB and AC of the triangle ABC, if the perimeter of triangle is 28 cm, the side BC is 4 cm less than AB and 9 cm less than AC.* It should be noted that in the task described above input parameters (Z) are the perimeter of the triangle and the ratios between the sides of the triangle. It is required to find (W) – the sides AB and AC of the triangle ABC. This information can be presented in the matrix M – to do this let us add an additional service row and a service column to the matrix. The row is designed to track the changes in known data. The column is designed to track the rules used. The result of the first step of working with mivar matrix is represented in Table 3.

Table 3: The mivar matrix after first step.

Param. / Rules	side AB	side BC	side AC	Perimeter	...	AB is .. greater than BC	AC is .. greater than BC	service column
Perimeter of triangle (using lengths of three sides)	**X**	**X**	**X**				**Y**	
Side BC (using perimeter and ratios of sides)		**Y**		**X**		**X**	**X**	
Side AC (using ratio of sides)		**X**	**Y**				**X**	
Side AB (using ratio of sides)	**Y**	**X**				**X**		
Service row	**W**		**W**	**Z**		**Z**	**Z**	

According to the mechanism described above, on the second step the rule "The side BC using the perimeter and the ratios of the sides" can be launched, which is indicated in the corresponding cell of the service column. Output of the rule after launching indicated as known variable Z. The results of this step are represented in Table 4.

Table 4: The mivar matrix after second step.

Param. / Rules	side AB	side BC	side AC	Perimeter	...	AB is .. greater than BC	AC is .. greater than BC	service column
Perimeter of triangle (using lengths of three sides)	**X**	**X**	**X**				**Y**	
Side BC (using perimeter and ratios of sides)		**Y**		**X**		**X**	**X**	✓
Side AC (using ratio of sides)		**X**	**Y**				**X**	
Side AB (using ratio of sides)	**Y**	**X**				**X**		
Service row	**W**	**Z**	**W**	**Z**		**Z**	**Z**	

In another words, by doing corresponding actions, we obtain the following algorithm to solve the set task: "The side BC using the perimeter and the ratios of the sides" $\longrightarrow$ "The side AC using the ratio of the sides" $\longrightarrow$ "The side AB using the ratio of the sides" .

The described mechanisms of information representation and information processing allow us to reduce the number of experts involved to develop AI systems and simplify subject domain descriptions, which allows to describe them more fully.

References

[1] Chibirova M.O. (2015). Analysis of existing approaches: ontology, cognitive maps and mivar nets. *Radio Industry*. Vol. **3**, pp. 55-66. (In Russian)

[2] Lakemeyer G., Nebel B. (1994). *Foundation of knowledge representation and reasoning*. Springer-Verlag, Berlin.

[3] Luger G.F. (2009). *Artificial intelligence: structures and strategies for complex problem solving*. Pearson Education, Boston.

[4] Varlamov O.O. (2003). Multidimensional informational space of data and rules representing overview. *Information technologies*. Vol. **5**, pp. 42-47. (In Russian)

Index

Abdushukurov, 81
Agabekova, 234
Ageeva, 117
Alama-Bućko, 79
Apanasovich, 143

Bachar, 135
Badziahin, 121
Baklanov, 268
Baklanova, 268
Baraille, 90
Baranovskiy, 122
Bernatavičienė, 64
Bokun, 237
Breiteneder, 87, 104
Brodinova, 87, 104
Burkatovskaya, 147
Byk, 296

Cheplyukova, 152
Chernov, 88
Chibisov, 126
Croux, 94

Dürre, 13
Danilkin, 308
Datta, 12
Deutsch, 44
Dreiziene, 127
Ducinskas, 127
Dus, 112
Dutta, 12
Dzemyda, 64

Egorov, 131

Filina, 156
Filzmoser, 44, 87, 94, 104, 111, 278
Fokianos, 20
Fried, 13

Gabko, 278
Gorbat, 296
Greičius, 291
Gurevich, 160

Hoang, 90
Hoffmann, 94
Hron, 111
Hubin, 274

Imbrasien, 64
Ionkina, 284
Iskakova, 165

Jakimauskas, 21

Kavaliou, 300
Kharin A., 96, 278
Kharin Yu., 25, 168
Khatskevich, 201
Kirlitsa, 195
Klimenok, 280
Kolchin, 284
Kolesnikova, 243
Kopats, 287
Krivko-Krasko, 227
Kruglov, 190
Kulak, 247, 256

Lappo, 199
Leri, 172
Lialikova, 201
Liventseva, 296
Lysenko, 227

Maltsew, 168
Malugin, 205, 296
Matalytski, 287
Matilainen, 30
Matkovskaya, 250
Medvedev, 210
Menshenin, 176
Miettinen, 30
Minkevičius, 291
Mishura, 38

Monti, 44
Mrochek, 296
Mukha, 215
Muradov, 81

Naumenko, 287
Navitskaya, 219
Nikitsionak, 135
Nikolov, 100
Nordhausen, 30
Novikov, 253
Novopoltsev, 205, 223, 296

Ogurtsova, 296
Oja, 30
Oliveira, 45
Orlova, 136
Orsingher, 53
Ortner, 87, 104

Pacheco, 45
Paunksnis, 64
Pavlov, 178
Pavlova, 296

Radavičius, 182
Ralchenko, 138

Sakalauskas, 21
Sakhno, 142
Salvador, 45
Savelov, 105
Sergeev, 300
Serov, 55
Sharilova, 247, 256
Shevlyakov, 58, 107
Shirokov, 107
Shvets, 268
Smirnov, 107
Soshnikova, 259
Sprindzuk, 300
Stabingis, 64
Starodubtsev, 304
Stoimenova, 68
Storvik, 274
Svidrytski, 143

Taskinen, 30
Ton, 96
Troush, 122
Tuzikov, 300

Valadas, 45
Varlamov, 308
Vasilevskiy, 58
Vencalek, 74
Vexler, 160
Vilela, 45
Visotski, 263
Vogel, 13
Voloshko, 185
Vorobeychikov, 147

Walach, 111
Walczak, 111

Yakushava, 199
Yatskou, 143

Zaharieva, 87, 104
Zaigrajew, 79
Zhalezka, 219
Zhao, 160
Zhuk, 112
Zhurak, 25
Zmitrovich, 227
Zubkov, 55, 156, 190
Zuev, 231

Научное издание **АНАЛИЗ И**

МОДЕЛИРОВАНИЕ КОМПЬЮТЕРНЫХ ДАННЫХ

ТЕОРЕТИЧЕСКАЯ И ПРИКЛАДНАЯ СТОХАСТИКА

Материалы XI Международной конференции

Минск, 6—10 сентября 2016 г.

На английском языке Ответственный за

выпуск *В. А. Волошко*

Подписано в печать 19.07.2016. Формат 60x84 V,. Бумага офсетная.
Ризография. Усл. печ. л. 36,73. Уч.-изд. л. 39,50.
Тираж 125 экз. Заказ 418.

Республиканское унитарное предприятие
«Издательский центр Белорусского государственного университета».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 1/159 от 27.01.2014.
Ул. Красноармейская, 6, 220030, Минск.

Отпечатано с оригинала-макета заказчика
в республиканском унитарном предприятии
«Издательский центр Белорусского государственного университета».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 2/63 от 19.03.2014.
Ул. Красноармейская, 6, 220030, Минск.