VILNIUS UNIVERSITY

INGRIDA VAIČIULYTĖ

STUDY AND APPLICATION OF MARKOV CHAIN MONTE CARLO METHOD

Summary of Doctoral Dissertation
Physical Sciences, Informatics (09 P)

Vilnius, 2014

The doctoral dissertation was prepared at the Institute of Mathematics and Informatics of Vilnius University in 2009–2014.

Scientific Supervisor

Prof. Dr. Habil. Leonidas Sakalauskas (Vilnius University, Physical Sciences, Informatics – 09 P).

The dissertation will be defended at the Council of the Scientific Field of Informatics of Vilnius University:

Chairman

Prof. Dr. Habil. Gintautas Dzemyda (Vilnius University, Physical Sciences, Informatics – 09 P).

Members:

Prof. Dr. Romualdas Kliukas (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07 T);

Prof. Dr. Audrius Lopata (Vilnius University, Physical Sciences, Informatics – 09 P);

Prof. Dr. Gediminas Stepanauskas (Vilnius University, Physical Sciences, Mathematics – 01 P);

Prof. Dr. Habil. Rimantas Šeinauskas (Kaunas University of Technology, Physical Sciences, Informatics – 09 P).

Opponents:

Prof. Dr. Kęstutis Dučinskas (Klaipėda University, Physical Sciences, Mathematics – 01 P);

Assoc. Prof. Dr. Olga Kurasova (Vilnius University, Physical Sciences, Informatics – 09 P).

The dissertation will be defended at the public meeting of the Council of the Scientific Field of Informatics in the auditorium number 203 at the Institute of Mathematics and Informatics of Vilnius University, at 1 p. m. on 26th of November 2014.

Address: Akademijos st. 4, LT-08663 Vilnius, Lithuania.

The summary of the doctoral dissertation was distributed on the 24th of October 2014.

A copy of the doctoral dissertation is available for review at the Library of Vilnius University.

VILNIAUS UNIVERSITETAS

INGRIDA VAIČIULYTĖ

MARKOVO GRANDINĖS MONTE-KARLO METODO TYRIMAS IR TAIKYMAS

Daktaro disertacijos santrauka

Fiziniai mokslai, informatika (09 P)

Vilnius, 2014

Disertacija rengta 2009–2014 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinis vadovas

prof. habil. dr. Leonidas Sakalauskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Disertacija ginama Vilniaus universiteto Informatikos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Gintautas Dzemyda (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Nariai:

prof. dr. Romualdas Kliukas (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07 T);

prof. dr. Audrius Lopata (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P);

prof. dr. Gediminas Stepanauskas (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P);

prof. habil. dr. Rimantas Šeinauskas (Kauno technologijos universitetas, fiziniai mokslai, informatika – 09 P).

Oponentai:

prof. dr. Kęstutis Dučinskas (Klaipėdos universitetas, fiziniai mokslai, matematika – 01 P);

doc. dr. Olga Kurasova (Vilniaus universitetas, fiziniai mokslai, informatika – 09P).

Disertacija bus ginama Vilniaus universiteto viešame Informatikos mokslo krypties tarybos posėdyje 2014 m. lapkričio mėn. 26 d. 13 val. Vilniaus universiteto Matematikos ir informatikos instituto 203 auditorijoje.

Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2014 m. spalio mėn. 24 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

INTRODUCTION

Research Area

Stochastic processes can be modeled and predicted by probabilistic statistical methods, using the data that describes the course of the process. The following various stochastic techniques are often used to describe and research random processes: Markov chain Monte Carlo, Gibbs sampler, Metropolis-Hastings algorithm, stochastic approximation, etc. (Rubinstein and Kroese, 2007; Spall, 2003). Markov chain Monte Carlo (MCMC) is a computer simulation method, which is widely used in statistics, technology, physics, bioinformatics, etc. MCMC method is often applied to calculate probabilities or rare events by importance sampling, in data analysis by EM (*expectation maximization*) algorithm, for practical application of Bayesian method by modeling the posterior distribution and using numerical methods in determining their parameters, etc. As the area of probabilistic models where MCMC methods can be applied is very wide, they are limited to the multidimensional distributions, which could be constructed in hierarchical way from elliptical distributions. Distributions that are obtained in this way can be applied when solving most of practical and theoretical exercises of data analysis.

Relevance of the Problem

Known MCMC algorithms usually generate some or several chains, determining convergence by empirical method and recording large enough Monte Carlo sample size in all chains (Bradley and Thomas, 2000). It is evident that these procedures are not very effective from the computational viewpoint as generation of chains uses too much computer's time, and in case of empirical termination of chain generation, statistically significant convergence might be not achieved yet. What is more, while applying MCMC, the problem often occurs in deciding what Monte Carlo sample size should be generated for separate chains.

The other relevant MCMC computational problems include selection of the parameters for the initial chain, computations when there is singularity of stochastic models, calculations with very large or very small intermediate values. Another relevant

problem of using MCMC is construction of skew heavy-tailed distributions and estimation of their parameters.

Research Object

The research object of dissertation is adaptive Markov chain Monte Carlo method study, its numerical realization and application in data analysis, regulation techniques of assessment of accuracy of estimators, selection of number of chains, algorithm termination, and Monte Carlo sample size for separate chains.

Aim and Objectives of the Research

The aim of the research is to examine Markov chain Monte Carlo adaptive methods by creating computationally effective algorithms for decision-making of data analysis with the given accuracy, and to research the effectiveness of these algorithms.

Algorithms and software for their realization were created, in order to reach this aim. This software was designed for regulation of Monte Carlo sample size in separate chains, assessment of accuracy of the estimators, and for the termination of Markov chain process. The algorithms constructed are employed for statistical estimation of data by MCMC method, using the data that was collected in practice or is known from literature. Effectiveness of this methods and algorithms is analyzed by statistical modeling method, constructed in this research. The numerical problems of MCMC method, which are examined in this study, were researched by solving several exercises of data analysis (estimation of parameters for skew *t* distribution, Poisson-Gaussian model, and stable distribution). The aforementioned exercises are characterized by properties that are common to other similar tasks, thus, the obtained results can be successfully applied for other statistical exercises.

Scientific Novelty

The following results have been obtained in the research:

1) termination rule of Markov chain generating;
2) the termination rule for Markov chain generating;
3) the method of testing the efficiency of Markov chain Monte Carlo algorithms;

4) application of the adaptive Markov chain Monte Carlo method for solving the exercises, the probabilistic models of which are built in a hierarchical way from elliptical distributions (for estimation of parameters of skew *t* distribution, Poisson-Gaussian model and multivariate α-stable distribution).

Practical Significance of the Results

In this study, MCMC algorithms are constructed for estimation of parameters of multivariate skew *t* distribution, Poisson-Gaussian model and multivariate α-stable distribution with the given accuracy. Furthermore, the efficiency of algorithms is verified by computer simulation. These algorithms can be applied to solve the practical tasks (financial sequences prediction, research on biological populations and insured events, etc.). Importantly, the results obtained can be applied to solve various tasks of statistical estimation by MCMC method: importance sampling method, EM algorithm, maximal likelihood method, and other.

The following practical results were obtained in this dissertation:

1) the algorithm has been created to estimate the parameters of skew *t* distribution;
2) the algorithm has been created to estimate the parameters of Poisson-Gaussian model;
3) the algorithm has been created to estimate the parameters of stable symmetric distribution;
4) the statistical modeling approach has been created to research the effectiveness of MCMC algorithms.

Defended Statements

1. The algorithms created and equipment for their realization is dedicated for:
    a) Monte Carlo sample size regulation at Markov chains;
    b) assessment of accuracy of the estimators;
    c) terminating the Markov chain process.
2. The methods and algorithms created can be applied for statistical estimation of data by adaptive MCMC method for solving practical and test exercises.

3. The algorithm created allows to solve exercises by MCMC method with the given accuracy. It reduces the volume of calculations (approximately twice) in comparison with the known.

Approbation and Publications of the Research

The results of the dissertation were presented at 10 international and 8 national scientific conferences. The main results of the dissertation were published in 15 periodical scientific publications: 1 of them is included in *ISI Web of Science* data base with own citation index, 1 of them is included in *Web of Science* data base, several of them are published at *CEEOL*, *Index Copernicus*, and other data bases.

Structure of the Dissertation

The dissertation consists of introduction, 4 chapters, conclusions, the list of references, and appendixes.

The introduction provides the aim of the dissertation, objectives, methods and the list of approbation and publications of the dissertation results.

In the first chapter, the relevance of selected topic is discussed, the problems are formulated.

In the second chapter, the Markov chain Monte Carlo method is constructed for estimation of the parameters for multivariate skew *t* distribution, and the application of algorithm for estimation of maximum likelihood by Monte Carlo method is presented.

In the third chapter, the multivariate empirical Bayesian Poisson-Gaussian model is constructed; other aspects of the Bayesian calculation are discussed.

In the fourth chapter, Markov chain Monte Carlo algorithm for estimation of the parameters for multivariate α-stable distribution is created.

Chapter 1. ANALYTICAL RESEARCH OF MARKOV CHAIN MONTE CARLO ALGORITHMS

This chapter presents the analytical overview of Markov chain Monte Carlo method and assumptions for solving the exercises by creating adaptive MCMC method.

1.1. Markov chain Monte Carlo method

In statistics, MCMC method forms a class of algorithms for imitation of probability distributions by constructing Markov chain, which allows us to obtain the necessary distribution as the distribution of chain balance state. Markov chain usually consists of several sequentially generated Monte Carlo samples, the so-called chains (or iterations), and estimators, calculated by using these samples.

If each chain generated depends only on estimators that were calculated in the previous chain, and does not depend on samples and estimators of earlier chains, then the chain, constructed from these iterations, is characterized by Markov feature.

A significant number of software, implementing MCMC method, have been developed, for example, *BUGS*, *Laplace's Demon*, *JAGS*, etc., and included in the list of *R* package. Algorithms have also been written to develop these programs, e.g., *Hamiltonian Monte Carlo*, *Metropolis within Gibbs*, *Griddy-Gibbs*, *Slice Sampler*, *t-walk*, *Robust Adaptive Metropolis*, *Elliptical Slice Sampler*, etc. (http://www.bayesian-inference.com/mcmc#algorithms). When using MCMC method, the following problems are encountered: selection of number of chains and regulation of Monte Carlo sample size for separate chains. Some or several chains are usually generated in the known MCMC algorithms, when fixed and large enough Monte Carlo sample size is detected in all chains and by estimating convergence empirically (Bradley and Thomas, 2000). It is evident that these procedures are not very effective from the computational viewpoint as generation of chains uses too much computer's time, and in case of empirical termination of chain generation, statistically significant convergence might be not achieved yet.

One way of solving the problem of selecting the Markov chain size is to terminate the generation of chain if samples, calculated in adjacent chains, do not differ statistically after applying statistical methods for verification of hypothesis on differences and matches of aforementioned samples (see section 1.4; Brooks and Gelman, 1998; Sakalauskas, 2000). Some authors attempted to introduce tests for comparison of two adjacent chains, however, these tests are one dimensional or allow to compare two vectors at best (Brooks and Gelman, 1998), while in practical exercises

probability distributions are often described by several vectors and several matrixes. Methods and algorithms for statistical estimation of Markov chains differences are proposed and analyzed in this dissertation, using standard Hoteling and Anderson's criteria (Anderson, 1958; Krishnaiah, 1984).

Another problem, related to reduction of calculation volume, is regulation of Monte Carlo sample size in separate chains. In fact, there is no need to generate large Monte Carlo samples when constructing first chains of Markov chain because smaller sample sizes are enough for iterative modification of model parameters. Large Monte Carlo samples should be generated only at the end of Markov chain, when statistical criterion is compatible with hypothesis on concurrence of the last chains of probabilistic models. Methods of Monte Carlo sample size regulation are proposed and simulated by computer in this dissertation by using statistical criterion about uniformity of Monte Carlo sample distributions in two Markov chains. Sample sizes can be taken as inversely proportional to the ratio of termination statistic and quantile of termination criterion.

From computational viewpoint, MCMC approach allows us to solve the equations, which include complex multivariate integrals, by constructing Markov chain of Monte Carlo samples. These equations can often be derived as necessary condition of optimality for some stochastic criteria (Dennis and Schnabel, 1996). In this dissertation, likelihood functions that describe these criteria are accepted as continuous and smoothly differentiated, therefore, MCMC method can be interpreted as gradient descent method for this likelihood function. Usually, it is possible to prove that EM algorithm, widely used in statistics, is a separate case of stochastic gradient search (see section 2.3). Application of MCMC method for multi-extreme tasks is not researched in this dissertation.

The algorithms have been created to estimate parameters of skew *t* distribution, Poisson-Gaussian model and of multivariate stable symmetric distribution by adaptive MCMC method. Statistical tasks, solved by adaptive MCMC method, described in dissertation, reveal these characteristics, allowing to solve other statistical exercises by adaptive MCMC method. Adaptive MCMC method is characterized by rule of sample size regulation and estimation of modeling errors by statistical method in separate chains by regulating the Markov chains Monte Carlo sample size and generated number of chains accordingly.

1.2. Estimation of parameters of statistical models

MCMC method is widely used to obtain estimators of parameters of statistical probability models. MCMC method is also often employed when constructing EM algorithms that are used to calculate estimators by maximum likelihood or Bayesian methods. These algorithms are applied for Monte Carlo Markov chain construction in next chapters.

Maximum likelihood (ML) approach allows to obtain the values of parameter sets of model, which maximize the likelihood function for fixed independent uniformly distributed model data sample. The higher the size is the higher is probability to obtain estimators, which will almost not differ from the actual parameter values. The realization of ML approach by MCMC method is researched in the present dissertation.

In the theory of estimators and decisions, Bayesian method allows to research the hypotheses, to calculate the probabilities of events. As there is very small amount of statistical data collected about rare occurrence events, probability estimation methods for estimating rare event probabilities are developed in this dissertation (see chapter 3).

1.3. Approximate integral calculation by Monte Carlo method

Probability theory and statistics often requires to calculate the means of various random values and vectors that are expressed in complex multivariate integrals. Although averaged random values are distributed widely, their means or mean-related statistical distributions can be quite well approximated by one-dimensional or multi-dimensional normal distribution or distributions that are related to it. It is possible to calculate multidimensional integrals by Monte Carlo method by using law of large number and central limit theorem. It is quite often required to calculate the estimators, which have non-linear dependence on Monte Carlo estimators, for example:

$$\hat{L} = -\sum_{i=1}^{K} \ln(P_i), \tag{1}$$

where $P_i = \frac{1}{N}\sum_{j=1}^{N} u_{ij}$, u_{ij} – are independent uniformly distributed random variables, $Eu_{ij} = \mu_i$, N – Monte Carlo sample size, K – any number. Monte Carlo estimator (1) is consistent estimator of expression $L = -\sum_{i=1}^{K} \ln(\mu_i)$ because $E\hat{L} = -\sum_{i=1}^{K} \ln(\mu_i) + O\left(\frac{1}{N}\right)$. In order to estimate function (1) confidence interval, its sample dispersion must be known. When sample consists of independent random variables and $EP_i = \mu_i$, estimator dispersion can be calculated as $D\hat{L} = \sum_{i=1}^{K} \frac{E(P_i - \mu_i)^2}{\mu_i^2} + O\left(\frac{1}{N}\right)$. From there and from law of large numbers it follows that the likelihood function of Monte Carlo estimator (1) dispersion is approximated as $D\hat{L} \approx \sum_{i=1}^{K}\left(\frac{P1_i}{P_i^2} - 1\right)$, where $P1_i = \frac{1}{N}\sum_{j=1}^{N} u_{ij}^2$.

After making some simple rearrangements and by using asymptotical Monte Carlo condition of normality, the 95% confidence interval of the estimator (1) can be approximated as follows:

$$\left[\hat{L} - \frac{2}{\sqrt{N}} \cdot \sqrt{\sum_{i=1}^{K}\left(\frac{P1_i}{P_i^2} - 1\right)}, \hat{L} + \frac{2}{\sqrt{N}} \cdot \sqrt{\sum_{i=1}^{K}\left(\frac{P1_i}{P_i^2} - 1\right)}\right]. \quad (2)$$

Estimation of confidence intervals of statistical estimators is no less important task than calculation of the estimators themselves. Frequently, statistical modeling estimators must be obtained with certain accuracy, for example, confidence intervals must be of required length. The length of confidence interval, which is the most important measure of estimator accuracy, can be adjusted by choosing Monte Carlo sample size N accordingly. When random sample is generated, it is possible to recurrently calculate that the sums, which are included into sample mean and dispersion, and to terminate generation as necessary accuracy of Monte Carlo estimators is reached.

1.4. Statistical hypothesis testing

Hypothesis about match of mean vectors and covariance matrix is discussed and tested in this section, which is further applied to identify the differences between two Markov chains. Let $X^1, X^2, \ldots, X^K$ be a random sample from a *d*-variate normal distribution with mean vector μ and covariance matrix Ω. The null hypothesis to be tested is $\mu = \mu_0$ or $\Omega = \Omega_0$, where μ_0 and Ω_0 are known. When the null hypothesis is rejected, then either $\mu \neq \mu_0$ or $\Omega \neq \Omega$, or $\mu \neq \mu_0$ and $\Omega \neq \Omega$.

The criteria for testing the null hypothesis are known (Anderson, 1958),

$$-2\log\lambda = N\log|\Omega_0| + N\cdot\mathrm{tr}\left(\Omega_0^{-1}\right)\cdot\left[S + (\bar{x}-\mu_0)(\bar{x}-\mu_0)^T\right] - N\log|S| - N\cdot d, \qquad (3)$$

where $N\cdot S = \sum_{j=1}^{N}(x_j-\bar{x})(x_j-\bar{x})^T$, $N\cdot\bar{x} = \sum_{j=1}^{N}x_j$, when the statistic $\chi^2 = \frac{-2\log\lambda}{(1-C_1)^2}$, $C_1 = \frac{2d^2+9d+11}{6N(d+3)}$ is approximated as a $\chi^2_{p_1}$ distribution with $p_1 = \frac{1}{2}d(d+3)$ degrees of freedom $F > F_{p_1,p_2}(\gamma)$ (Krishnaiah, 1984). *F* approximation is more accurate, when the statistic $F = \frac{-2\log\lambda}{b}$, where $b = \frac{p_1}{1-C_1-\frac{p_1}{p_2}}$, is approximated as an *F* distribution. In these approximations the null hypothesis is rejected if the computed value is too high, i.e., $\chi^2 > \chi^2_{p_1}(\gamma)$, $F > F_{p_1,p_2}(\gamma)$ (Krishnaiah, 1984).

Chapter 2. ESTIMATION OF MULTIVARIATE SKEW *t* DISTRIBUTION PARAMETERS

In this chapter, application of adaptive MCMC method for estimation of skew *t* distribution parameters by ML approach is analyzed.

2.1. Definition of multivariate skew *t* distribution

In general, the skew *t* distribution is defined by hierarchical statistical model through multivariate normal distribution, the vector of mean of which is also distributed as a normal distribution, when both covariate matrices depend on the parameter, distributed according to the gamma distribution (Azzalini and Genton, 2008).

Let $X=(X_1, X_2,..., X_d)$ be a random vector that is distributed as a multivariate normal vector $X \sim N(z,\Omega)$ with density $f(x|z,t,\Omega)=(t/\pi)^{\frac{d}{2}} \cdot |\Omega|^{-\frac{1}{2}} \cdot e^{-t\cdot(x-z)^T\cdot\Omega^{-1}\cdot(x-z)}$, where the vector of mean $z=(z_1, z_2,..., z_d)$ in its turn is distributed as a multivariate normal distribution $z \sim N\left(\mu, \frac{\Theta}{2t}\right)$ in the cone $W=\{\eta\cdot(z-\mu)\geq 0, \eta\subset\Re^d\}$, Ω, Θ – full rank matrices and random variable t follows from the Gamma distribution with the parameter α (Azzalini and Capitanio, 2003) that has the density $f_1(t|\alpha)=\frac{t^{\frac{\alpha}{2}-1}}{\Gamma(\alpha/2)}\cdot e^{-t}$.

Using this definition, the skew *t* distribution density is expressed as a multi-dimensional integral:

$$
\begin{aligned}
&p(x|\mu,\Omega,\Theta,\alpha,\eta)=2\cdot\int_0^\infty \int_{\eta\cdot(z-\mu)\geq 0} f(x|z,t,\Omega)\cdot f(z|\mu,t,\Theta)\cdot f_1(t|\alpha)dzdt= \\
&=\int_0^\infty \int_{\eta\cdot(z-\mu)\geq 0} \frac{2}{\pi^d\cdot|\Omega|^{\frac{1}{2}}\cdot|\Theta|^{\frac{1}{2}}\cdot\Gamma\left(\frac{\alpha}{2}\right)}\cdot t^{\frac{\alpha}{2}+d-1}\times \\
&\times e^{-t\cdot\left((x-z)^T\cdot\Omega^{-1}\cdot(x-z)+(z-\mu)^T\cdot\Theta^{-1}\cdot(z-\mu)+1\right)}dzdt
\end{aligned}
\tag{4}
$$

This model can be interpreted statistically in the following way: for example, the area *W* describes biological infection source, pollution source or investors' preference area, etc., which are related to further randomized dispersion, distributed by normal or other elliptic distribution. For simplicity, this dissertation considers that the aforementioned area is described only by one linear restriction, but in general case, several linear restrictions can be involved.

2.2. Estimators of maximum likelihood approach

Let $X = \left(X^1, X^2, ..., X^K\right)$ is the matrix of observation where X^i, $i = \overline{1,K}$, are independent random vectors, distributed as multivariate skew *t* distribution $X \sim ST(\mu, \Omega, \Theta, \alpha, \eta)$, the parameters of which are estimated by the ML approach (see section 1.2). By using integral (4), log-likelihood function is described in the following way:

$$L(\mu,\Omega,\Theta,\alpha,\eta) = -\sum_{i=1}^{K} \ln\left(p\left(X^i \middle| \mu,\Omega,\Theta,\alpha,\eta\right)\right) = -\sum_{i=1}^{K} \ln\left(E\left(f\left(X^i \middle| Z,G,\Omega\right)\right)\right), \tag{5}$$

where means are calculated by random variables *Z* and *G* (see (17)) with common density:

$$f\left(z,t \middle| \mu,\Omega,\alpha\right) = \begin{cases} 2 \cdot f\left(z \middle| \mu,t,\Theta\right) \cdot f_1\left(t \middle| \alpha\right), & \text{if } \eta \cdot (z-\mu) \geq 0, \\ 0, & \text{if } \eta \cdot (z-\mu) < 0, \end{cases} \tag{6}$$

where $\eta \subset \Re^d$.

Estimators of distribution parameters must minimize the log-likelihood function $L(\mu,\Omega,\Theta,\alpha,\eta) \to \min\limits_{\mu,\Omega,\Theta,\alpha,\eta}$. ML estimators $\hat{\mu}, \hat{\Omega}, \hat{\Theta}, \hat{\alpha}, \hat{\eta}$ of multivariate skew *t* distribution (4) are calculated by equating the respective derivatives of likelihood function to zero and solving the system of received equations (see section 1.2).

Using the method of fixed-point iteration, parameter estimators are obtained from following equations:

$$\frac{1}{K}\sum_{i=1}^{K} E\left(t \cdot \left(X^i - z\right) \middle| X^i, \hat{\mu}, \hat{\Omega}, \hat{\Theta}, \hat{\alpha}, \hat{\eta}\right) = 0, \tag{7}$$

$$\hat{\Omega} = \frac{2}{K}\sum_{i=1}^{K} E\left(t \cdot \left(X^i - z\right) \cdot \left(X^i - z\right)^T \middle| X^i, \hat{\mu}, \hat{\Omega}, \hat{\Theta}, \hat{\alpha}, \hat{\eta}\right), \tag{8}$$

$$\hat{\Theta} = \frac{2}{K}\sum_{i=1}^{K} E\left(t \cdot \left(z - \hat{\mu}\right) \cdot \left(z - \hat{\mu}\right)^T \middle| X^i, \hat{\mu}, \hat{\Omega}, \hat{\Theta}, \hat{\alpha}, \hat{\eta}\right), \tag{9}$$

$$\hat{\alpha} = \frac{K \cdot \sum\limits_{i=0}^{d-1} \dfrac{1}{\dfrac{2 \cdot i}{\hat{\alpha}} + 1}}{\sum\limits_{i=1}^{K} E\left(\dfrac{1}{2}\ln\left(\hat{A}\right) \middle| X^i, \hat{\mu}, \hat{\Omega}, \hat{\Theta}, \hat{\alpha}, \hat{\eta}\right)}, \tag{10}$$

$$\hat{\eta} = \frac{1}{K}\sum_{i=1}^{K} E\left(t \cdot \hat{\Lambda} \middle| X^i, \hat{\mu}, \hat{\Omega}, \hat{\Theta}, \hat{\alpha}, \hat{\eta}\right), \tag{11}$$

where

$$\hat{A} = \left(X^i - z\right)^T \cdot \hat{\Omega}^{-1} \cdot \left(X^i - z\right) + \left(z - \hat{\mu}\right)^T \cdot \hat{\Theta}^{-1} \cdot \left(z - \hat{\mu}\right) + 1,$$

$$\hat{\Lambda}_r = \left(z - \hat{\mu}\right)_r \cdot \left(\hat{\Omega}^{-1} \cdot \left(X^i - z\right) - \hat{\Theta}^{-1} \cdot \left(z - \hat{\mu}\right)\right)_d, \quad r = 1, 2, \ldots, d-1, \quad \hat{\Lambda}_d = 0.$$

Using the derivatives of likelihood function, gradient descent method is obtained for likelihood function optimization:

$$\mu^{k+1} = \mu^k - \zeta_\mu \cdot \frac{\partial L\left(\mu^k, \Omega^k, \Theta^k, \alpha^k, \eta^k\right)}{\partial \mu}, \tag{12}$$

$$\Omega^{k+1} = \Omega^k - 2 \cdot \Omega^k \cdot \frac{\partial L\left(\mu^k, \Omega^k, \Theta^k, \alpha^k, \eta^k\right)}{\partial \Omega} \cdot \Omega^k, \tag{13}$$

$$\Theta^{k+1} = \Theta^k - 2 \cdot \Theta^k \cdot \frac{\partial L\left(\mu^k, \Omega^k, \Theta^k, \alpha^k, \eta^k\right)}{\partial \Theta} \cdot \Theta^k, \tag{14}$$

$$\alpha^{k+1} = \alpha^k - \zeta_\alpha \cdot \frac{\partial L\left(\mu^k, \Omega^k, \Theta^k, \alpha^k, \eta^k\right)}{\partial \alpha}, \tag{15}$$

$$\eta^{k+1} = \eta^k - \zeta_\eta \cdot \frac{\partial L\left(\mu^k, \Omega^k, \Theta^k, \alpha^k, \eta^k\right)}{\partial \eta}, \tag{16}$$

where

$$\zeta_\mu = \frac{1}{\sum_{i=1}^{K} E\left(t \middle| X^i, \mu^k, \Omega^k, \Theta^k, \alpha^k, \eta^k\right)}, \quad \zeta_\alpha = \frac{\alpha^k}{\sum_{i=1}^{K} E\left(\frac{\ln\left(A^k\right)}{2} \middle| X^i, \mu^k, \Omega^k, \Theta^k, \alpha^k, \eta^k\right)},$$

$$\zeta_\eta = 1, \; A^k = \left(X^i - z\right)^T \cdot \left(\Omega^k\right)^{-1} \cdot \left(X^i - z\right) + \left(z - \mu^k\right)^T \cdot \left(\Theta^k\right)^{-1} \cdot \left(z - \mu^k\right) + 1.$$

EM algorithm follows from gradient descent method. Its realization by MCMC method is described in section 2.3.

2.3. Markov chain Monte Carlo algorithm

Since the likelihood function in this case is also expressed as multi-dimensional integrals, MCMC algorithm was created in dissertation, in order to obtain ML estimators by selecting Monte Carlo sample sizes in separate chains in the way as to decrease a total number of trials and by terminating the generation of chains when the samples in two adjacent procedure steps differ insignificantly.

Estimators of parameters can be calculated by iteration method, using EM algorithm (see section 1.2), when some initial values are chosen, and by calculating the integrals, included into equations (7)–(11), by Monte Carlo method. Using a random sample:

$$G_j \sim \text{Gamma}\left(\frac{\alpha^k}{2}\right),\ \varphi_j \sim \text{N}\left(0, \Theta^k\right),\ Z_j = \begin{cases} \mu^k + \varphi_j, \text{ if } \eta \cdot \varphi_j \geq 0, \\ \mu^k - \varphi_j, \text{ if } \eta \cdot \varphi_j < 0, \end{cases} \tag{17}$$

where $j = 1, 2, \ldots, N^k$, $k = 0, 1, 2, \ldots$, Monte Carlo estimates are calculated on k^{th} Markov chain:

$$P_i^k = \frac{1}{N^k} \sum_{j=1}^{N^k} f\left(X^i \middle| Z_j, G_j, \Omega^k\right), \tag{18}$$

$$M_i^k = \frac{1}{N^k} \sum_{j=1}^{N^k} \left(X^i - Z_j\right) \cdot G_j \cdot f\left(X^i \middle| Z_j, G_j, \Omega^k\right), \tag{19}$$

$$S_i^k = \frac{1}{N^k} \sum_{j=1}^{N^k} \left(X^i - Z_j\right) \cdot \left(X^i - Z_j\right)^T \cdot G_j \cdot f\left(X^i \middle| Z_j, G_j, \Omega^k\right), \tag{20}$$

$$T_i^k = \frac{1}{N^k} \sum_{j=1}^{N^k} \left(Z_j - \mu^k\right) \cdot \left(Z_j - \mu^k\right)^T \cdot G_j \cdot f\left(X^i \middle| Z_j, G_j, \Omega^k\right), \tag{21}$$

$$Q_i^k = \frac{1}{N^k} \sum_{j=1}^{N^k} d \cdot \Lambda^k \cdot G_j \cdot f\left(X^i \middle| Z_j, G_j, \Omega^k\right), \tag{22}$$

where

$$\Lambda^k = \left(\Lambda_1^k, \Lambda_2^k, \ldots, \Lambda_d^k\right)$$

$$\Lambda_r^k = (Z_r - \mu^k)_r \cdot \left(\left(\Omega^k\right)^{-1} \cdot \left(X^i - Z_r\right) - \left(\Theta^k\right)^{-1} \cdot \left(Z_r - \mu^k\right)\right)_d, \quad r = 1, 2, \ldots, d-1$$

$$\Lambda_d^k = 0,\ 1 \leq i, k \leq K.$$

From (12)–(16) follow these formulas:

$$\mu^{k+1} = \mu^k + \frac{1}{K \cdot \zeta^k} \sum_{i=1}^{K} \frac{M_i^k}{P_i^k} \tag{23}$$

$$\Omega^{k+1} = \frac{2}{K} \sum_{i=1}^{K} \frac{S_i^k}{P_i^k} \tag{24}$$

$$\Theta^{k+1} = \frac{2}{K} \sum_{i=1}^{K} \frac{T_i^k}{P_i^k} \tag{25}$$

$$\alpha^{k+1} = \frac{1}{h^{k+1}} \sum_{i=0}^{d-1} \frac{1}{\frac{2 \cdot i}{\alpha^k} + 1} \tag{26}$$

$$\eta^{k+1} = \eta^k - \frac{1}{K} \sum_{i=1}^{K} \frac{Q_i^k}{P_i^k} \tag{27}$$

By using equations (5) and (18), a consistent estimator of log-likelihood function is obtained (see (1)). The confidence intervals of received estimators are determined by using asymptotical normal approximation of Monte Carlo estimators that was described in the first chapter (see section 1.3). Thus, Monte Carlo chains are generated, according to the formulas (23)–(27) until the length of confidence interval (2) becomes lower than the chosen value ε, $\varepsilon > 0$, and statistical hypotheses about matching of mean vectors and covariance matrices in two adjacent iterations $H_0 : \mu^{k+1} = \mu^k$, $\Omega^{k+1} = \Omega^k$, $\Theta^{k+1} = \Theta^k$, $\alpha^{k+1} = \alpha^k$, $\eta^{k+1} = \eta^k$ is not rejected (see section 1.4). Thus, statistical hypothesis is rejected, if termination criterion (see (3)) $H^k > \Psi_{\delta,p}$, where $\Psi_{\delta,p}$ is a quantile of χ_p^2 distribution, $p = d(d+3)$ degrees of freedom, δ is a significance level. If this hypothesis is not rejected, generation of Markov chain can be terminated and the estimates, obtained in the last iteration, can be accepted.

Taking into account the fact that distribution in expressions of Monte Carlo estimators (23)–(27) can be asymptotically approximated by multivariate normal distribution (see section 1.3); the following law for regulation of Monte Carlo sample size is introduced:

$$N^{k+1} \geq \Psi_{\nu,p} \cdot \frac{N^k}{H^k}, \tag{28}$$

where ν is significance level. Trying to avoid too small or too large Monte Carlo sample size, which is calculated according to (28), it is restricted by bottom value $N_{\min}$ and by top value $N_{\max}$. Seeking to reduce the time, required for generation of Monte Carlo samples, it is possible to terminate generation, when the length of confidence interval (2) becomes lower than the initial value, while the value (28) is not reached yet.

2.4. Computer modeling

The algorithm created was tested with simulated data. It was observed, that estimates obtained by a stochastic algorithm are close to the estimates obtained by analytical approach. To check the assumption on asymptotical distribution of statistics, obtained by adaptive MCMC method in the optimal point by the one-dimensional or multidimensional normal law, another computational experiment was performed. $M = 100$ samples were generated, each consisting of $K = 100$ skew *t* distribution values, and by using the program that calculates integrals, the skew *t* distribution parameters for each sample were analytically obtained.

Then, $N = 2000$ volume independent Monte Carlo samples (17) were generated, and termination test and others MCMC method tests were calculated for each sample. Statistical χ_p^2 and Kolmogorov-Smirnov criteria did not object to assumptions about the asymptotic termination statistic as distributed a χ_p^2 law with $p = 10$ degrees of freedom made with 0,05 significance level. Efficiency of MCMC algorithm created, when the sample size is regulated, and of standard algorithm, where the sample size in all iterations is fixed, is compared in this dissertation. Efficiency of algorithms is compared by the sample size, which is necessary to meet the conditions of exercise convergence. The tested adaptive MCMC algorithm allowed to reduce the calculations by 2,5 times.

The algorithm, created in the dissertation, was applied for analysis of data of Australian Sports Institute and for analysis of data of USA companies for financial years of 2007–2011.

A simple illustration is provided by analyzing a subset of the Australian Institute of Sport (AIS) data, examined by Cook and Weisberg (1994), which contains various biomedical measurements of a group of Australian athletes (Smyth, GK, 2011). For each data pair – the body mass index (*BMI*), the percentage of body fat (*Bfat*), sum of the skin folds (*ssf*) and the lean body mass (*LBM)* – two-dimensional skew *t* distribution model is estimated. Fig. 1 displays the surface plot and the contour lines of the fitted skew *t* distribution density together with data dissemination for 100 women. The diagrams presented visually illustrate the asymmetry of skew *t* distribution density and distribution properties.

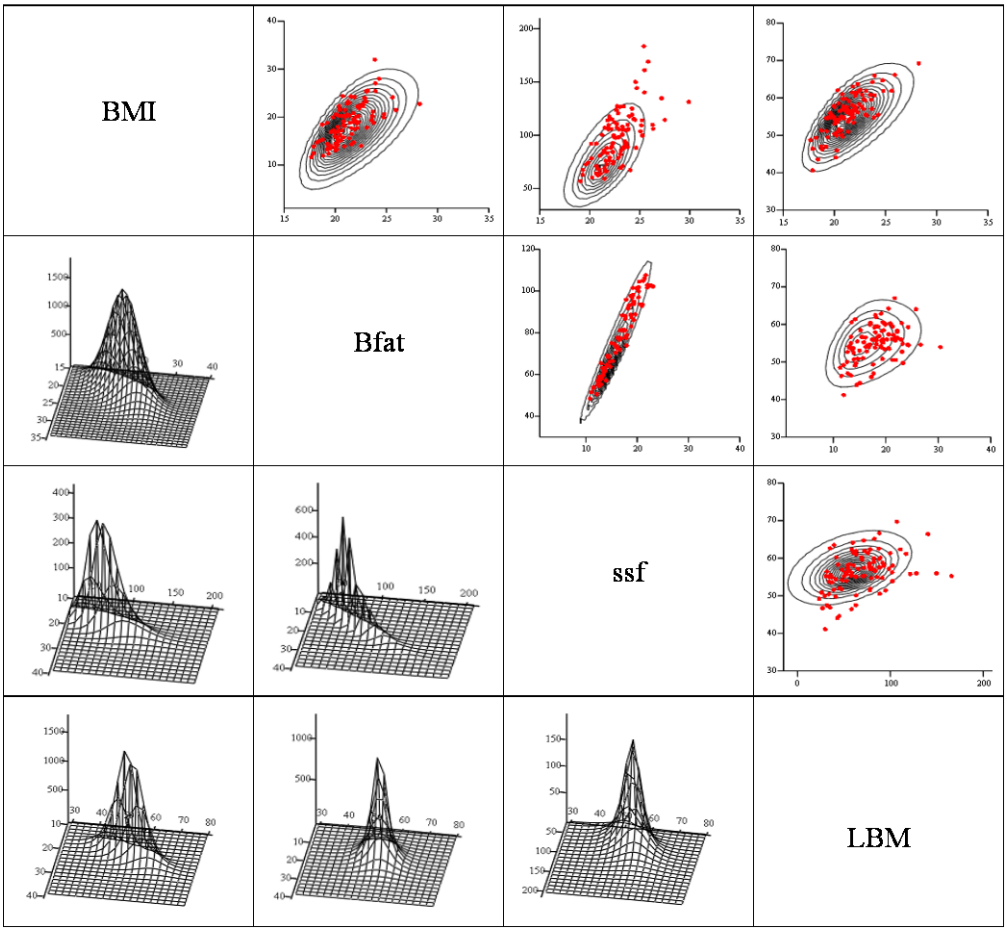

Fig. 1. Surface plots of the skew *t* distribution and contour levels together with variables dissemination diagram for AIS data

Further, the example is given, where skew *t* distribution is analyzed in the context of economical situation, representing changes in investors' expectation preference area during crisis and post crisis periods.

Estimation of two-dimensional skew *t* distribution parameters is applied by using rating data of enterprises. Day stock selling price (*Close*) and selling volume of 130 USA enterprises are examined during the period from 2007 till 2011.

The skew *t* distribution density contour lines, drawn in fig. 2, show the dissemination of data pair *Close* and *Volume* for each year. By using statistical interpretation of skew *t* distribution (see section 2.1), it is possible to represent the investors' preference area that comes from skew *t* distribution density integration half-plane, drawn in red arrows. In 2008, investors' preference area decreased significantly –

this represents the real economical situation in that year: investors' trust decreased. After the crisis, in 2009, it increased and approached the limit of 2007 again.

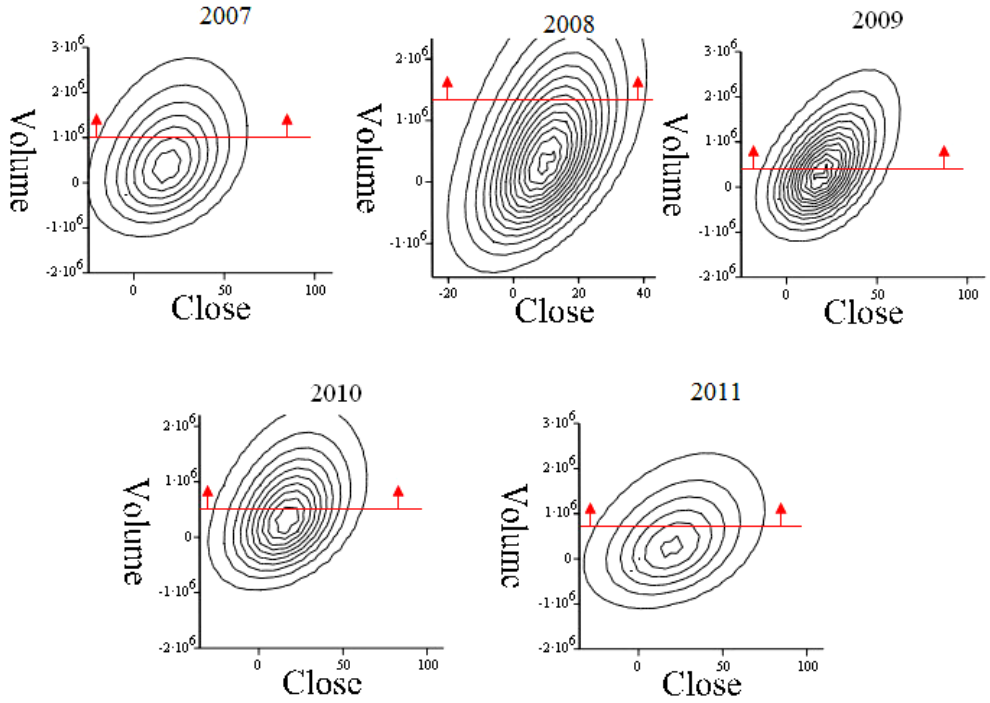

Fig. 2. Contour levels of the skew *t* distribution to *Close* and *Volume* data for the period of 2007–2011

As the results in fig. 2 show, economical crisis had influence on investors' preference area. In the rating data distribution, shown in fig. 2, investors' preference area reflects their expected results of the enterprises.

Chapter 3. POISSON-GAUSSIAN MODEL FOR ESTIMATION OF MULTIVARIATE RARE EVENT PROBABILITIES

Adaptive MCMC algorithm to estimate some rare event probabilities by empirical Bayesian approach was constructed in this dissertation. The algorithm of event probability estimation by empirical Bayesian approach was constructed by Tsutakawa *et al*. (1985), Sakalauskas (2010) postulating a prior assumption that *logit* transformations are approximated as a normal distribution. Generalization of this algorithm in multivariate case and the method for initial data selection for multivariate Poisson-Gaussian model is further analyzed in this dissertation.

3.1. Poisson-Gaussian model

Let's consider a set $C = \left(c^1, c^2, \ldots, c^K\right)$ of *K* populations, where each population c^i consist of I_i individuals, $i = 1, 2, \ldots, K$. Assume that some events of *M* type can occur in the populations (i.e., illness, death, insured events, and so on). The aim is to estimate the unknown probabilities of events P_i^m, where number of events Y_i^m is m^{th} number of appearance at the population i^{th}, $i = 1, 2, \ldots, K$, $m = 1, 2, \ldots, M$.

Empirical Bayesian approach assumes that $Y_i^m \sim \text{Pois}\left(\lambda_i^m\right)$, where $\lambda_i^m = I_i \cdot P_i^m$. Poisson-Gaussian model assumes that *logit* transformation of event probabilities $\rho_i^m = \ln\dfrac{P_i^m}{1-P_i^m}$ in the populations is distributed, according to multivariate normal distribution with parameters μ, Ω (Bradley and Thomas, 2000; Tsutakawa *et al.*, 1985), when the density of ρ_i^m is:

$$g\left(\rho_i^m, \mu, \Omega\right) = \frac{\exp\left(-\left(\rho_i^m - \mu\right)^T \Omega^{-1}\left(\rho_i^m - \mu\right)\right)}{\sqrt{|\Omega|} \cdot (2\pi)^{\frac{M}{2}}}, \tag{29}$$

$i = 1, 2, \ldots, K$, $m = 1, 2, \ldots, M$.

In the case observed, the unknown parameters μ, Ω are estimated by ML approach after minimizing the log-likelihood function:

$$L(\mu, \Omega) = -\sum_{i=1}^{K} \ln\left(p\left(P_i^m \middle| Y\right)\right) = -\sum_{i=1}^{K} \ln\left(\int_{-\infty}^{+\infty} \prod_{m=1}^{M} f\left(Y_i^m, \frac{I_i}{1+e^{-\rho_i^m}}\right) g\left(\rho_i^m, \mu, \Omega\right) d\rho_i^m\right), \tag{30}$$

where $f_{Y_i^m}\left(\varpi, \lambda_i^m\right) = P\left(Y_i^m = \varpi; \lambda_i^m\right) = e^{-\lambda_i^m} \dfrac{\left(\lambda_i^m\right)^{\varpi}}{\varpi!}$, $\varpi = 0, 1, 2, \ldots$.

When respective first derivatives of likelihood function are equated to zero and the system is solved, the estimators are obtained:

$$\mu^{k+1} = \frac{1}{K}\sum_{i=1}^{K} \frac{\int_{-\infty}^{+\infty} \rho_i^m \prod_{m=1}^{M} f\left(Y_i^m, \frac{I_i}{1+e^{-\rho_i^m}}\right) \cdot g\left(\rho_i^m, \mu^k, \Omega^k\right) d\rho_i^m}{B_i\left(\mu^k, \Omega^k\right)}, \tag{31}$$

$$\Omega^{k+1} = \frac{1}{K}\sum_{i=1}^{K} \frac{\int_{-\infty}^{+\infty} \left(\rho_i^m - \mu^k\right)\left(\rho_i^m - \mu^k\right)^T \prod_{m=1}^{M} f\left(Y_i^m, \frac{I_i}{1+e^{-\rho_i^m}}\right) \cdot g\left(\rho_i^m, \mu^k, \Omega^k\right) d\rho_i^m}{B_i\left(\mu^k, \Omega^k\right)}, \tag{32}$$

where $B_i\left(\hat{\mu}, \hat{\Omega}\right) = \int_{-\infty}^{+\infty} \prod_{m=1}^{M} f\left(Y_i^m, \frac{I_i}{1+e^{-\rho_i^m}}\right) \cdot g\left(\rho_i^m, \hat{\mu}, \hat{\Omega}\right) d\rho_i^m$.

EM algorithm is useful to solve equations (31) and (32), in order to get ML estimates (see section 1.2). As the probabilities of rare events $\left(P_i^m \approx 0\right)$ and $logitP_i^m \sim N\left(\mu, \Omega\right)$ are analyzed, it might be assumed that probabilities are approximately distributed by a log-normal law. Using this assumption, the following heuristic initial point $\left(\mu^0, \Omega^0\right)$ in the equations (31) and (32) can be taken:

$$\mu^0 = \ln\left(P\right) - \frac{1}{2}\Omega^0, \tag{33}$$

$$\Omega^0 = \ln\left(\frac{\sum_{i=1}^{K} \Psi^{-1} \cdot \left(Y_i - I_i P\right) \cdot \left(Y_i - I_i P\right)^T \cdot \Psi^{-1}}{\left(\sum_{i=1}^{K} I_i\right)^2} + 1\right), \tag{34}$$

where $P = \left(\frac{\sum_{i=1}^{K} Y_i^1}{\sum_{i=1}^{K} I_i}, \frac{\sum_{i=1}^{K} Y_i^2}{\sum_{i=1}^{K} I_i}, \ldots, \frac{\sum_{i=1}^{K} Y_i^M}{\sum_{i=1}^{K} I_i}\right)$, $\Psi = \mathrm{Diag}\left(P\right)$, $Y_i = \left(Y_i^1, Y_i^2, \ldots, Y_i^M\right)$, $i = 1, 2, \ldots, K$.

3.2. Markov chain Monte Carlo algorithm

Let's say that *k* number of Markov chains is generated and estimates μ^k, Ω^k in each chain with initial values (33) and (34) are calculated. To avoid the computational problems that can occur due to very small values of intermediate results an auxiliary function is introduced. Statistical criterion H^k (see (3)) to test hypothesis $H_0 : \mu^{k+1} = \mu^k$, $\Omega^{k+1} = \Omega^k$ is introduced for algorithm termination. This criterion,

calculated in the optimal point of likelihood function, is approximated by an *F*, if the sample size is large enough (see section 1.4; Krishnaiah, 1984).

The algorithm is terminated, if statistical criterion does not oppose to hypothesis $H^k \le F_{\delta,p}$ and the length of confidence interval (2) of likelihood function is lower than the chosen initial value ε, $\varepsilon > 0$: $2 \cdot \tau_\gamma \cdot \sqrt{\frac{D^k X}{N^k}} \le \varepsilon$, and covariance matrix estimate is not singular, where $F_{\delta,p}$ and τ_γ are a quantiles of *F* and normal distributions, respectively, δ, γ – significance level.

Moreover, seeking to avoid the singularity of models, Monte Carlo chains are terminated, when the ratio between maximum and minimum eigenvalues estimates of the covariance matrix exceeds the selected critical value, which usually is 10–30 (Lin and Zhu, 2004). This value, chosen in this dissertation, is equal to 10.

New Monte Carlo sample is generated, if at least one of the termination conditions is not met. Sample size is regulated, following the rule that is analogical to (28):

$$N^{k+1} \ge \frac{N^k \cdot \nu}{H^k} \cdot F_{\nu,p}, \tag{35}$$

where $F_{\nu,p}$ – quantile of an *F* distribution, ν – significance level. Application of this rule allows to choose the Monte Carlo sample size in Markov chain rationally, also ensures the convergence of this exercise into optimal value of likelihood function (Sakalauskas, 2000).

3.3. Computer modeling

In order to test the behaviour of created algorithm, the experiments were made with selected simulated data and real data – the data of suicides (*suic*) and homicides (*hom*) in Lithuania in 2003. $M = 2$ events in populations set, made of $K = 60$ municipalities, are analyzed. Initial parameters are calculated by formulas (33) and (34). 100 Markov Monte Carlo chains are generated and the probabilities P_i^m of events *suic* and *hom* are estimated

by using the described MCMC algorithm applied to Poisson-Gaussian model, $i = 1, 2, \ldots, 60$ and $m = 1, 2$.

Table 1 presents the probability estimates of men *suic* $\left(P_{\text{suic}} \cdot 10^5\right)$ and *hom* $\left(P_{\text{hom}} \cdot 10^5\right)$ in the largest cities of Lithuania, which were obtained in this dissertation. What is more, the estimates $P_{\text{suic\_1}} \cdot 10^5$ and $P_{\text{hom\_1}} \cdot 10^5$, which were obtained by Sakalauskas (2010) without taking into consideration the correlation between these events, are included in the table for comparison. Upon introducing correlation, lower value of likelihood function is obtained. As different estimates are obtained, it might be concluded that introduction of the correlation is obligatory.

Table 1. ***Suicide and homicide stats in the largest cities of Lithuania in 2003***

City	Quantity of men	Quantity of *suic*	Quantity of *hom*	$P_{\text{suic}} \cdot 10^5$	$P_{\text{hom}} \cdot 10^5$	$P_{\text{suic\_1}} \cdot 10^5$	$P_{\text{hom\_1}} \cdot 10^5$
Vilnius	246412	110	22	72,63	15,68	47,50	12,03
Kaunas	167308	85	27	72,40	15,02	54,08	15,19
Klaipėda	88308	45	16	70,48	15,40	56,58	15,49
Šiauliai	60221	43	14	66,89	14,39	74,21	16,26
Panevėžys	53913	26	9	69,39	14,92	57,25	14,98

The dependences on number of generated chains, obtained from analysis of men data, when the length of confidence interval does not exceed $\varepsilon = 0{,}1$, are depicted in fig. 3–7. As it might be observed in fig. 3, the minimum value of log-likelihood function has already been achieved in the first iterations. The termination criterion $\frac{H^k}{F_{\delta,p}}$ (see *test* in fig. 4) is decreasing until the critical value 1 of termination is achieved, where $F_{\delta,p}$ is an F distribution with $p = 5$ degrees of freedom at 0,95-quantile. Fig. 5 shows the dependence on sample size (marked as *N predictable*) that is calculated by rule (35). The real sample size (marked as *N real*), depicted in this figure, is obtained by terminating the sample generation when the length of confidence interval does not exceed the chosen critical value ε. Fig. 6 shows the dependence of length of confidence interval on the number of chains. The presented dependences show that the confidence interval is large at the beginning of the chain, then it decreases down to the required value of 0,1.

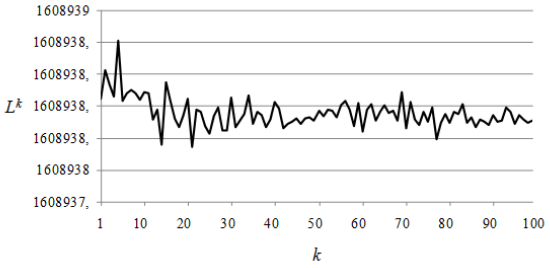

Fig. 3. Likelihood function L^k

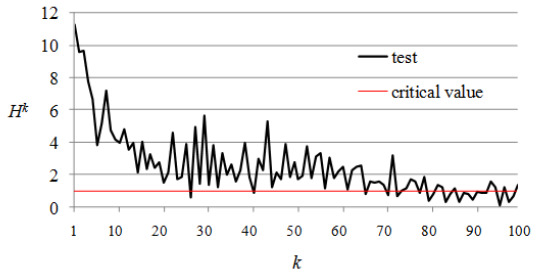

Fig. 4. Termination test H^k

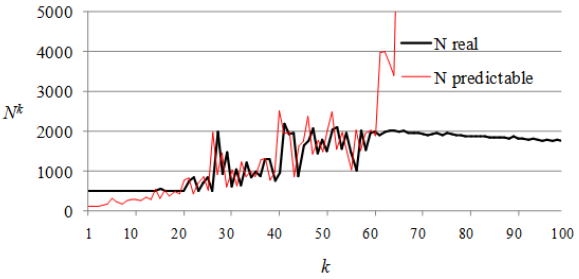

Fig. 5. Sample size N^k

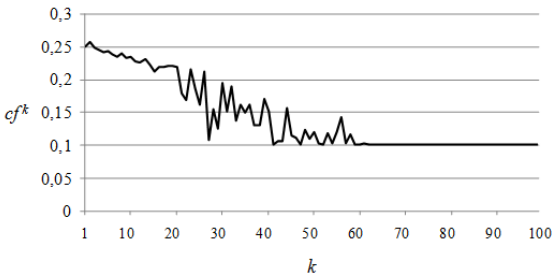

Fig. 6. Confidence interval length cf^k

The research to compare the effectiveness of MCMC algorithm, created when the sample size is regulated, with effectiveness of standard algorithm, where sample size is fixed in all iterations, is also included in this dissertation. Effectiveness of algorithms is compared by sample size, necessary to meet the conditions of convergence. In the case researched, adaptive MCMC method, allows us to calculate Bayesian estimates of rare event probabilities with the required accuracy and by reducing the calculations almost twice.

Chapter 4. ESTIMATION OF STABLE SYMMETRIC VECTOR DISTRIBUTION PARAMETERS

The statistical adaptive stable symmetric vector MCMC algorithm, which allows to estimate the parameters of these vector distributions, by using ML approach, is constructed in this dissertation.

Zoloterev's expression of stable distribution density is used in this dissertation (Золотарев, 1983). In one-dimension case, it is known that $s = s_1 \cdot s_2$, where:

s_1 – random stable variable with skewness parameter $\beta = 1$ and shape parameter $\alpha_1 < 1$;

s_2 – another random stable variable with skewness parameter $\beta = 0$ and shape parameter α_2;

s – random stable variable with skewness parameter $\beta = 0$ and shape parameter $\alpha = \alpha_1 \cdot \alpha_2$ (Rachev and Mittnik, 1993; Ravishanker and Qiou, 1999).

While applying this method, it is usually selected that s_2 would be a random variable, which is normally distributed, i.e., $\alpha_1 = \frac{\alpha}{2}$ and $\alpha_2 = 2$. In this way, the multivariate stable symmetric vector can be expressed through normally distributed random vector, and α-stable variables (Rachev and Mittnik, 1993; Ravishanker and Qiou, 1999) as $X = \mu + \sqrt{s_1} \cdot s_2$, where s_1 – subordinator with parameter α, random vector $s_2 \sim N(0, \Omega)$ and μ is a random vector of mean.

4.1. Estimators of maximum likelihood approach

Let's consider that the sample $X = \left(X^1, X^2, \ldots, X^K\right)$ consists of independent d-variate stable vectors. The log-likelihood function of this sample is

$$L(X, \mu, \Omega, \alpha) = \frac{K}{2} \ln|\Omega| - \sum_{i=1}^{K} \ln\left(\int_0^{\infty} \int_0^{1} B\left(X^i, y_i, z_i, \mu, \Omega, \alpha\right) \cdot \exp\{-z_i\} dy_i dz_i \right), \qquad (36)$$

where

$$t_\alpha(y) = \frac{\sin\left(\frac{\pi \cdot \alpha \cdot y}{2}\right) \cdot \sin\left(\frac{\pi \cdot (2-\alpha) \cdot y}{2}\right)^{\frac{2-\alpha}{\alpha}}}{(\sin \pi \cdot y)^{\frac{2}{\alpha}} \cdot \left(\cos\frac{\pi \cdot \alpha}{4}\right)^{\frac{2}{\alpha}}}, \quad z_i = \left| \frac{t_\alpha(y_i)}{s_i} \right|^{\frac{\alpha}{2-\alpha}}$$

$$B\left(X^i, y_i, z_i, \mu, \Omega, \alpha\right) = \exp\left\{ - \frac{z_i^{\frac{2-\alpha}{\alpha}} \left(X^i - \mu\right)^T \Omega^{-1} \left(X^i - \mu\right)}{2 \cdot t_\alpha(y_i)} \right\} \cdot \frac{z_i^{\frac{2-\alpha}{\alpha} \cdot \frac{d}{2}}}{\sqrt{|\Omega|} \cdot t_\alpha(y_i)^{\frac{d}{2}}}.$$

Estimators of parameters are calculated from these equations by fixed-point method:

$$\hat{\mu} = \frac{\sum_{i=1}^{K} \frac{X^i \cdot g_i}{f_i}}{\sum_{i=1}^{K} \frac{g_i}{f_i}}, \tag{37}$$

$$\hat{\Omega} = \frac{1}{K} \cdot \sum_{i=1}^{K} \frac{\left(X^i - \hat{\mu}\right)\left(X^i - \hat{\mu}\right)^T g_i}{f_i}, \tag{38}$$

where

$$g\left(X, \hat{\mu}, \hat{\Omega}, \alpha\right) = \int_0^{\infty}\int_0^1 \frac{z^{\frac{2-\alpha}{\alpha}}}{t_\alpha(y)} \cdot B\left(X, y, z, \hat{\mu}, \hat{\Omega}, \alpha\right) \cdot \exp\{-z\} dy dz, \tag{39}$$

$$f\left(X, \hat{\mu}, \hat{\Omega}, \alpha\right) = \int_0^{\infty}\int_0^1 B\left(X, y, z, \hat{\mu}, \hat{\Omega}, \alpha\right) \cdot \exp\{-z\} dy dz, \tag{40}$$

EM algorithm can be used to solve the equations (37), (38) after integrals (39) and (40) are calculated by Monte Carlo method (see section 1.2). When μ and Ω are fixed, the shape parameter estimate can be obtained by solving the exercise of one variable minimization.

4.2. Markov chain Monte Carlo algorithm

Let's say the initial values $\mu^0, \Omega^0, \alpha^0$ are selected, then k number of Markov chains is generated, and estimates $\mu^k, \Omega^k, \alpha^k$ in each chain are calculated. Let's say $Y_j \sim U(0,1)$ and $Z_j \sim -\ln(Y_j)$, where $j = 1, 2, \ldots, N^k$, N^k – is Monte Carlo sample size of the k^{th} chain. Then the sums are calculated:

$$P_i^k = \frac{1}{N^k} \sum_{j=1}^{N^k} B\left(X^i, Y_j, Z_j, \mu^k, \Omega^k, \alpha^k\right), \tag{41}$$

$$PP_i^k = \frac{1}{N^k} \sum_{j=1}^{N^k} \left(B\left(X^i, Y_j, Z_j, \mu^k, \Omega^k, \alpha^k\right)\right)^2, \tag{42}$$

$$V_i^k = \frac{1}{N^k}\sum_{j=1}^{N^k}\frac{Z_j^{\frac{2-\alpha^k}{\alpha^k}}}{t_{\alpha^k}\left(Y_j\right)}B\left(X^i, Y_j, Z_j, \mu^k, \Omega^k, \alpha^k\right), \tag{43}$$

$$VV_i^k = \frac{1}{N^k}\sum_{j=1}^{N^k}\left(\frac{Z_j^{\frac{2-\alpha^k}{\alpha^k}}}{t_{\alpha^k}\left(Y_j\right)}B\left(X^i, Y_j, Z_j, \mu^k, \Omega^k, \alpha^k\right)\right)^2, \tag{44}$$

that are necessary to receive estimators in the next iteration, according to (37) and (38) and EM algorithm:

$$\mu^{k+1} = \frac{\sum_{i=1}^{K} X^i \frac{V_i^k}{P_i^k}}{\sum_{i=1}^{K}\frac{V_i^k}{P_i^k}}, \tag{45}$$

$$\Omega^{k+1} = \frac{1}{K}\sum_{i=1}^{K}\left(X^i - \mu^k\right)\left(X^i - \mu^k\right)^T\frac{V_i^k}{P_i^k}. \tag{46}$$

Then the consistent Monte Carlo estimator of log-likelihood function is obtained (see (1)). The 95% confidence interval for likelihood function is (see (2)):

$$\left[L^k - \frac{2}{\sqrt{N^k}}\sqrt{N^k\cdot\sum_{i=1}^{K}\frac{PP_i^k}{\left(P_i^k\right)^2} - K}, L^k + \frac{2}{\sqrt{N^k}}\sqrt{N^k\cdot\sum_{i=1}^{K}\frac{PP_i^k}{\left(P_i^k\right)^2} - K}\right], \tag{47}$$

Monte Carlo chains are generated, according to formulas (45)–(46) the length of confidence interval (47) becomes lower than chosen value ε, $\varepsilon > 0$, and statistical hypothesis about matching of mean vectors and covariance matrices in two adjacent iterations $H_0 : \mu^{k+1} = \mu^k$, $\Omega^{k+1} = \Omega^k$ is not rejected (see section 1.4). To test this hypothesis, the Anderson criterion is used (see (3)): $H^k > \Psi_{\delta,p}$, where $\Psi_{\delta,p}$ – is χ_p^2 distribution quantile with *p* degrees of freedom, δ – significance level (see section 1.4).

To regulate Monte Carlo sample size, the rule, analogical to rule, which is applied in stochastic programming, is introduced. Application of this rule allows to choose the Monte Carlo sample size in Markov chain rationally, also ensures the convergence of

sets (45) and (46) into optimal value of likelihood function with probability 1 (Sakalauskas, 2000).

4.3. Computer modeling

The algorithm created was tested with chosen simulated data and share data of 3 telecommunication enterprises: AT&T, BellSouth and CenturyLink, from 20-01-2012 to 01-04-2012.

By using MCMC algorithm, described in the dissertation, $k = 50$ Markov chains were generated. The sample size limit $N^k \geq 500$ was applied to avoid too small or too large values. In this case, termination conditions of the algorithm were satisfied after $k = 28$ iterations.

Fig. 8–12 depict dependences when the length of confidence interval does not exceed $\varepsilon = 0{,}2$. As it might be observed in fig. 8, the log-likelihood function is decreasing until the zone of possible solution is achieved. The presented dependences in fig. 9 show that the confidence interval decreases down to the required value of 0,2. In fig. 10, *N real* is obtained by terminating sample generation when the length of confidence interval does not exceed the critical value $\varepsilon = 0{,}2$. *N predictable* is Monte Carlo sample size, calculated according to rule (28). Termination test is depicted in fig. 11, where *critical value* is the value of 0,999-quantile of χ_p^2 distribution with $p = 9$ degrees of freedom (equal to 27,88).

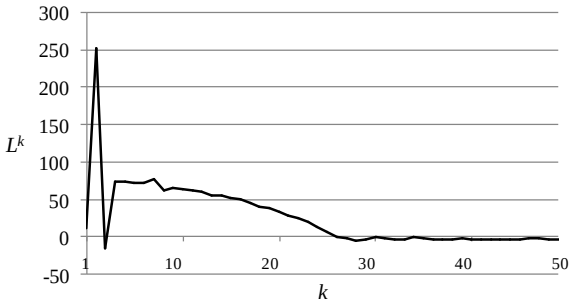

Fig. 8. Likelihood function L^k

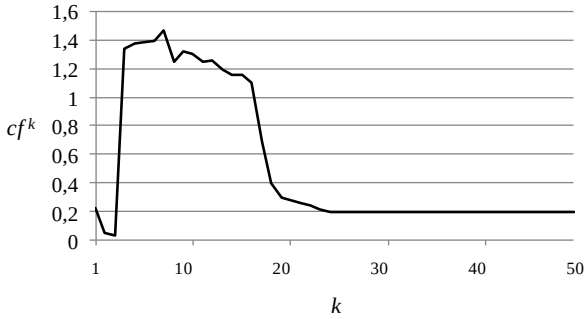

Fig. 9. Confidence interval length cf^k

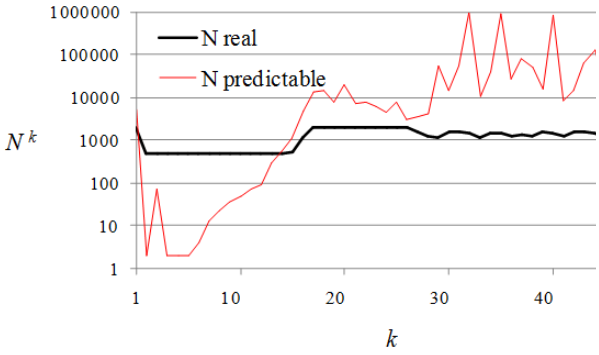

Fig. 10. Sample size N^k

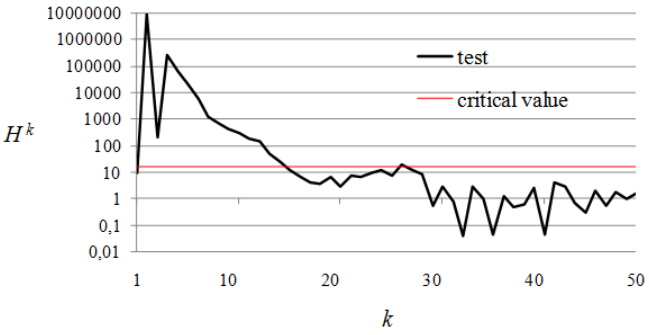

Fig. 11. Termination test H^k

Analogical research of share prices was carried out with the following 5 enterprises: AT&T, Bellsouth, CenturyLink, CBS, and Sprint. In table 2, the fixed and regulated by rule (28) Monte Carlo sample size is presented, required to meet the conditions for algorithm termination.

Table 2. Comparison of standard and adaptive MCMC algorithms of stable symmetric vector distribution

Dimension d	ε	Sample size	k	N^k in the last iteration	Total N
3	0,1	regulated	28	6 478	88 004
		fixed	30	7 000	210 000
	0,2	regulated	28	1 646	32 346
		fixed	29	2 000	58 000
5	0,1	regulated	23	11 724	156 567
		fixed	19	12 000	228 000
	0,2	regulated	12	2 519	16 904
		fixed	19	3 000	57 000

Comparison of created algorithm with a standard MCMC algorithm with fixed sample size has revealed that it allows to obtain the estimators symmetric stable vector law with the necessary accuracy in lower number of chains, and, thus, reducing the volume of calculations by almost two times.

RESULTS AND CONCLUSIONS

1. Markov chain Monte Carlo adaptive methods were created and tested.
2. Rules for Monte Carlo sample size regulation in Markov chains, for assessment of the accuracy of the estimators, and for termination of Monte Carlo chain process were proposed.
3. Algorithm for estimation of skew *t* distribution parameters by adaptive MCMC method was created. It was shown that this method realizes log-likelihood function stochastic gradient search, implementing it with EM algorithm. Tests made with Australian sportsmen data and financial data of enterprises, belonging to health-care industry, confirmed that numerical properties of the method correspond to the theoretical model. Algorithm can be used to test the systems of stochastic type and to solve other statistical tasks by MCMC method.
4. Adaptive MCMC algorithm was constructed to estimate some rare event probabilities by empirical Bayesian approach. Initial data selection method for multivariate Poisson-Gaussian model was proposed. To avoid too small or too large values, the modified likelihood function was introduced. Model for social data analysis was created.
5. The statistical adaptive MCMC algorithm for researching the parameters of stable symmetric vector distributions was constructed. This algorithm was applied for creation of model of share data of telecommunication. This model can be used for data analysis of stock market.
6. Computational problems, typical for most exercises that use MCMC method, were analyzed, including MCMC method for skew *t* distribution, Poisson-Gaussian model and estimations of parameters of stable symmetric vector law. The results, obtained in this way, can be successfully applied for other statistical exercises.
7. Efficiency of MCMC algorithms was tested by statistical modeling. Tests of algorithms behaviour have shown that adaptive MCMC algorithm allows to obtain estimators of examined distribution parameters in lower number of chains, and reducing the volume of calculations approximately two times.

List of Literature Referenced in this Summary

[1] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

[2] Azzalini, A., Capitanio, A. (2003). Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew *t* Distribution. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), 65, 367–389.

[3] Azzalini, A., Genton, M. G. (2008). Robust Likelihood Methods Based on the Skew-t and Related Distributions. *International Statistical Review*, 76 (1), 106–129.

[4] Bayesian-Inference. Prieiga per internetą [žiūrėta 2014-06-18]: <http://www.bayesian-inference.com/mcmc#algorithms>

[5] Bradley, P. C., Thomas, A. L. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman and Hall.

[6] Cook, R. D., Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

[7] Dennis, J. E., Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: Classics in Applied Mathematics. Prieiga per internetą [žiūrėta 2014-06-12]: <http://books.google.lt/books?id=RtxcWd0eBD0C&pg=PA15&hl=lt&source=gbs_toc_r&cad=3#v=onepage&q&f=false>

[8] Brooks, S. P., Gelman, A. (1998). General Methods for Monotoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.

[9] Kabašinskas, A. (2007). *Finansinių rinkų statistinė analizė ir statistinio modeliavimo metodai*. *Daktaro disertacija*. Vilnius: Vytauto Didžiojo universitetas. Prieiga per internetą [žiūrėta 2013-03-15]: < http://www.mii.lt/files/disert_08_akabasinskas.pdf>

[10] Kaklauskas, L. (2012). *Fraktalinių procesų kompiuterių tinkluose stebėsenos ir valdymo metodų tyrimas*. *Daktaro disertacija*. Vilnius: Vilniaus universitetas. Prieiga per internetą [žiūrėta 2013-03-15]:

< http://www.mii.lt/files/mii_dis_2012_kaklauskas.pdf>

[11] Krishnaiah, P. R. (1984). *Handbook of Statistics 1: Analysis of Variance*. New York: Elsevier Science & Technology Books.

[12] Lin, X., Zhu, Y. (2004). Degenerate Expectation-Maximization Algorithm for Local Dimension Reduction. *Studies in Classification, Data Analysis, and Knowledge Organization*, 46, 259–268.

[13] Rachev, S. T., Mittnik, S. (1993). Modeling Asset Returns with Alternative Stable Distributions. *Econometric Reviews*, 12(3), 261– 330.

[14] Ravishanker, N., Qiou, Z. (1999). Monte Carlo EM Estimation for Multivariate Stable Distributions. *Statistics & Probability Letters*, 45(4), 335–340.

[15] Rubinstein, R. Y., Kroese, D. P. (2007). *Simulation and the Monte Carlo Method* (2nd ed.). New York: Wiley.

[16] Sakalauskas, L. (2000). Nonlinear Stochastic Optimization by Monte-Carlo Estimators. *Informatica*, 11(4), 455–468.

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

[18] Smyth, GK (2011). *Australasian Data and Story Library* (OzDASL). Prieiga per internetą [žiūrėta 2014-06-12]: <http://www.statsci.org/data/oz/ais.txt>

[19] Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. New York: Wiley.

[20] Tsutakawa, R. K., Shoop, G. L., Marienfeld, C. J. (1985). Empirical Bayes Estimation of Cancer Mortality Rates. *Statistics in Medicine*, 4(2), 201–212.

[21] Золотарев, В. М. (1983). *Одномерные устойчивые распределения*. Москва: Наука.

List of Publications on Topic of Dissertation

1. Sakalauskas, L., Kalsyte, Z., Vaiciulyte, I., Kupciunas, I. (2014). The relationship between the transparency in provision of financial data and the change in investors'

expectations. *Ekonominė inžinerija – Engineering Economics*, ISSN 1392-2785 (po recenzavimo priimtas spaudai, išduota pažyma Nr. 221).

2. Vaičiulytė, I., Sakalauskas, L. (2014). Markovo grandinės Monte Karlo metodo taikymas tiriant sociologinius duomenis. *Jaunųjų mokslininkų darbai*, ISSN 1648-8776, 2(42) (po recenzavimo priimtas spaudai, išduota pažyma Nr. 036-14).
3. Vaičiulytė, I., Sakalauskas, L. (2014). Sub-gausinio vektoriaus skirstinio parametrų vertinimas Monte-Karlo Markovo grandinės metodu. *Jaunųjų mokslininkų darbai*, ISSN 1648-8776, 1(41), 104–107.
4. Vaičiulytė, I. (2014). Antisimetrinio *t* skirstinio taikymas tiriant finansinius duomenis. *Jaunųjų mokslininkų darbai*, 1(41), 147–151, ISSN 1648-8776.
5. Sakalauskas, L., Vaičiulytė I. (2013). Multidimensional rare event probability estimation algorithm. *Computational Science and Techniques*, ISSN 2029-9966, 1(2), 222–228.
6. Sakalauskas, L., Kalsyte, Z., Vaiciulyte, I., Kupciunas, I. (2013). The application of stable and skew *t*-distributions in predicting the change in accounting and governance risk ratings. *Proceedings of the 8th International Conference „Electrical and Control Technologies“*, ISSN 1822-5934, Kaunas, 53–58.
7. Vaiciulyte, I. (2012). Adaptive Monte-Carlo Markov Chain for Multivariate Statistical Estimation. *Proceedings of International Workshop „Stochastic programming for implementation and advanced applications“*, ISBN 9786099524146, Nida, 119–124.
8. Sakalauskas, L., Vaičiulytė, I. (2012). Daugiamatis mažų dažnių vertinimo algoritmas. *Lietuvos matematikos rinkinys, Lietuvos matematikų draugijos darbai*, ISSN 0132-2818, 53B, 260–263.
9. Sakalauskas, L., Vaiciulyte, I. (2012). Adaptive Monte-Carlo Markov chain. *Proceedings of 2nd International Conference „Stochastic Modeling Techniques and Data Analysis“*, ISBN 978-981-270-968-4, Crete, Greece, 653–660.
10. Sakalauskas, L., Vaiciulyte, I. (2012). Maximum likelihood estimation of multivariate skew *t*-distribution. *Proceedings of 1st International Conference on Operations Research and Enterprise Systems*, ISBN 978-989-8425-97-3, Algarve-Vilamoura, Portugal, 200–203, Vilamoura: SciTePress.

11. Vaičiulytė, I., Sakalauskas, L. (2011). Daugiamačio antisimetrinio *t*-skirstinio parametrinis įvertinimas. *Jaunųjų mokslininkų darbai*, ISSN 1648-8776, 4(33), 157–163.
12. Sakalauskas, L., Vaiciulyte, I. (2011). Estimation of shape parameter of the multivariate *t*-skew distribution. *Proceedings of 14th International Conference „Applied Stochastic Models and Data Analysis"*, ISBN 97888467-3045-9, Roma, Italy, 1211–1218.
13. Vaičiulytė, I., Sakalauskas, L. (2011). Daugiamačio antisimetrinio *t*-skirstinio parametrų vertinimas Monte-Karlo Markovo grandinių metodu. *Jaunųjų mokslininkų darbai*, ISSN 1648-8776, 1(30), 137–141.
14. Sakalauskas, L., Vaiciulyte, I. (2010). Estimation of skew *t*-distribution by Monte-Carlo Markov chain approach. *Proceedings of 9th International Conference „Computer Data Analysis and Modeling"*, ISBN 978-985-476-847-2, Minsk, Belarus, 207–210.
15. Sakalauskas, L., Vaiciulyte, I. (2010). Estimation of skew *t*-distribution by the Monte-Carlo Markov chain approach. *Proceedings of 1th International Conference „Stochastic Modeling Techniques and Data Analysis"*, ISBN 978-981-270-968-4, Crete, Greece, 747–753.

About the Author

Ingrida Vaičiulytė was born on January 4th, 1984 in Radviliškis. In 2002, she graduated Radviliškis Vaižgantas Gymnasium. In 2006, she received a Bachelor's degree in mathematics at Šiauliai University. In 2009, received Master of mathematics and professional qualification of a teacher at Šiauliai University. From 2006 till 2013 worked at Šiauliai Bank. From 2012, is working as the lecturer of mathematics at Šiauliai State College. From 2009 till 2014 was a doctoral student at the Institute of Mathematics and Informatics of Vilnius University.

E-mail: ingrid.vaiciulyte@gmail.com

MARKOVO GRANDINĖS MONTE-KARLO METODO TYRIMAS IR TAIKYMAS

Tyrimų sritis

Atsitiktiniai procesai gali būti modeliuojami bei prognozuojami tikimybiniais statistiniais metodais, pasinaudojus duomenimis apie proceso eigą. Atsitiktiniams procesams aprašyti ir tirti dažnai taikomi įvairūs stochastiniai: Markovo grandinės Monte-Karlo metodas, Gibso imties išrinkimo ir Metropolio-Hastingso algoritmai, stochastinė aproksimacija bei kt. (Rubinstein ir Kroese, 2007; Spall, 2003). Markovo grandinės Monte-Karlo (angl. *Markov Chain Monte Carlo* – MCMC) metodas yra kompiuterinio imitavimo būdas, plačiai taikomas statistikoje, technikoje, fizikoje, bioinformatikoje ir t. t. MCMC metodas dažnai taikomas retų įvykių tikimybėms apskaičiuoti imties išrinkimo metodu (angl. *importance sampling*), duomenų analizėje EM (angl. e*xpectation maximization*) algoritmu, Bajeso metodo praktiniam pritaikymui, modeliuojant aposteriorinius skirstinius bei skaitiniais metodais nustatant jų parametrus, ir pan. Kadangi tikimybinių modelių sritis, kuriems gali būti pritaikyti MCMC metodai, yra labai plati, disertacijoje apsiribojama daugiamačiais skirstiniais, kurie gali būti sukonstruoti hierarchiniu būdu iš elipsinių skirstinių. Tokiu būdu gauti skirstiniai gali būti pritaikyti sprendžiant daugelį praktinių ir teorinių duomenų analizės uždavinių.

Problemos aktualumas

Žinomuose MCMC algoritmuose paprastai sugeneruojamos kelios arba keliolika grandžių, empiriškai nustatant konvergavimą ir užfiksavus visose grandyse pakankamai didelį Monte-Karlo imčių tūrį (Bradley ir Thomas, 2000). Aišku, kad skaičiuojamuoju požiūriu tokios procedūros yra nelabai efektyvios, nes tenka sunaudoti daug kompiuterio laiko grandžių generavimui, o nutraukus grandinės generavimą empiriškai, statistiškai reikšmingas konvergavimas dar gali būti nepasiektas. Taip pat taikant MCMC dažnai kyla problema, kokio dydžio imtys turėtų būti generuojamos atskirose grandyse.

Tokiu būdu, aktualios MCMC skaitmeninimo problemos yra grandinės grandžių skaičiaus nustatymas ir Monte-Karlo imčių tūrio atskirose grandyse reguliavimas. Prie kitų aktualių MCMC skaitmeninimo problemų galima priskirti pradinės grandies

parametrų parinkimą, skaičiavimus esant tikimybinių modelių singuliarumui, skaičiavimus su labai didelėmis arba labai mažomis tarpinėmis reikšmėmis. Gana aktuali MCMC panaudojimo problema yra asimetrinių skirstinių su didelėmis atsitiktinėmis reikšmėmis konstravimas ir parametrų vertinimas.

Tyrimų objektas

Disertacijos tyrimų objektas yra adaptuotos Markovo grandinės Monte-Karlo metodo tyrimas, skaitinis realizavimas ir taikymas duomenų analizėje, tikslumo vertinimo, grandžių skaičiaus parinkimo, algoritmo stabdymo ir Monte-Karlo imčių tūrio reguliavimo būdai.

Tyrimų tikslas ir uždaviniai

Darbo tikslas – ištirti Markovo grandinės Monte-Karlo adaptavimo metodus, sudarant efektyvius skaitinius algoritmus, leidžiančius priimti duomenų analizės sprendimus su iš anksto nustatytu patikimumu, bei ištirti šių algoritmų efektyvumą.

Siekiant šio tikslo disertacijoje sudaryti algoritmai bei juos realizuojanti programinė įranga, skirta Monte-Karlo imčių tūrio adaptavimui atskirose grandyse, įvertinių tikslumo vertinimui bei Markovo grandinės proceso stabdymui. Sukurti metodai ir algoritmai yra pritaikyti duomenų statistiniam vertinimui MCMC metodu, pasinaudojant praktiškai surinktais arba žinomais literatūroje duomenimis. Šių metodų bei algoritmų efektyvumas tiriamas pasinaudojant disertacijoje sudarytu statistinio modeliavimo metodu. MCMC skaitmeninimo problemos, nagrinėjamos disertacijoje, yra ištirtos sprendžiant kelis duomenų analizės uždavinius (asimetrinio *t* skirstinio, Puasono-Gauso modelio ir stabiliojo dėsnio parametrų vertinimo), pasižyminčius ypatybėmis, kurios yra būdingos daugeliui kitų uždavinių, ir tokiu būdu gauti rezultatai gali būti sėkmingai pritaikomi ir kitiems panašiems uždaviniams spręsti.

Mokslinis naujumas

Disertacijoje gauti šie rezultatai:

1) Markovo grandinės Monte-Karlo imčių tūrio adaptavimo taisyklė;
2) Markovo grandinės generavimo stabdymo taisyklė;

3) Markovo grandinės Monte-Karlo algoritmų efektyvumo tyrimo metodas;
4) adaptuotos Markovo grandinės Monte-Karlo metodo pritaikymas uždaviniams, kurių tikimybiniai modeliai konstruojami hierarchiniu būdu iš elipsinių skirstinių, spręsti (asimetrinio t skirstinio, Puasono-Gauso modelio ir daugiamačio α-stabiliojo dėsnio parametrams vertinti).

Praktinė darbo reikšmė

Disertacijoje sudaryti MCMC algoritmai daugiamačio asimetrinio t skirstinio, Puasono-Gauso modelio ir daugiamačio α-stabiliojo dėsnio parametrams vertinti duotu tikslumu bei kompiuterinio modeliavimo būdu ištirtas šių algoritmų efektyvumas. Sudaryti algoritmai gali būti pritaikyti praktikoje iškylantiems uždaviniams spręsti (finansinių sekų prognozavimui, biologinių populiacijų ir draudiminių įvykių populiacijose tyrimams ir pan.). Gauti rezultatai gali būti pritaikyti sprendžiant įvairius statistinio vertinimo uždavinius MCMC metodu: imties išrinkimo metodas, EM algoritmas, didžiausio tikėtinumo metodas ir pan.

Disertacijoje gauti šie praktiniai rezultatai:

1) sudarytas algoritmas asimetrinio t skirstinio parametrams vertinti;
2) sudarytas algoritmas Puasono-Gauso modelio parametrams vertinti;
3) sudarytas algoritmas stabiliojo simetrinio skirstinio parametrams vertinti;
4) sudarytas statistinio modeliavimo metodas MCMC algoritmų efektyvumui tirti.

Ginamieji teiginiai

1. Sudaryti algoritmai bei juos realizuojanti programinė įranga, skirta:
    a) Monte-Karlo imčių tūrio reguliavimui Markovo grandyse;
    b) įvertinių tikslumo vertinimui;
    c) Markovo proceso grandinės stabdymui.
2. Sudaryti algoritmai gali būti pritaikyti duomenų statistiniam vertinimui adaptuotu MCMC metodu sprendžiant praktinius ir testinius uždavinius.
3. Sudaryti algoritmai leidžia spręsti statistinio vertinimo uždavinius MCMC metodu duotu tikslumu sumažinant (maždaug dvigubai) skaičiavimo apimtį palyginus su žinomais algoritmais.

Darbo rezultatų aprobavimas

Tyrimų rezultatai buvo pristatyti 8 respublikinėse ir 10 tarptautinių konferencijų. Paskelbta 15 straipsnių disertacijos tema recenzuojamuose mokslo žurnaluose: 1 iš jų yra *ISI Web of Science* duomenų bazėje ir turi citavimo indeksą, 1 yra *Web of Science* duomenų bazėje, kiti straipsniai – *CEEOL* ir *Index Copernicus* ir kitose duomenų bazėse.

Disertacijos struktūra

Darbą sudaro įvadas, keturi skyriai, išvados, literatūros apžvalga ir priedai.

Įvade pateikiamas disertacijos tikslas, uždaviniai, metodai, darbo rezultatų aprobavimo ir publikavimo sąrašas.

Pirmame skyriuje aptariamas pasirinktos temos aktualumas ir bendra jos problematika.

Antrame skyriuje sudaromas Markovo grandinės Monte-Karlo algoritmas daugiamačio asimetrinio t skirstinio parametrams vertinti, pateikiamas algoritmo pritaikymas Monte-Karlo didžiausio tikėtinumo įvertinimui.

Trečiame skyriuje sudaromas daugiamatis empirinio Bajeso Puasono-Gauso modelis, aptariami kiti Bajeso skaičiavimo aspektai.

Ketvirtame skyriuje aprašomas Markovo grandinės Monte-Karlo algoritmas daugiamačio α -stabiliojo skirstinio parametrams vertinti.

Bendrosios išvados

1. Sudaryti ir ištirti Markovo grandinės Monte-Karlo adaptavimo metodai.
2. Pasiūlytos taisyklės Monte-Karlo imčių tūrio reguliavimui Markovo grandyse, įvertinių tikslumo vertinimui, Markovo proceso grandinės stabdymui.
3. Sukonstruoti trijų hierarchinių modelių adaptuoti MCMC algoritmai.
4. Išnagrinėtos antisimetrinio t skirstinio, Puasono-Gauso modelio, stabiliojo dėsnio parametrų vertinimo MCMC metodu skaitmeninimo problemos.
5. Sudaryti algoritmai pritaikyti realių duomenų tyrimų modeliams sudaryti:
    a) Australijos sporto instituto duomenų analizei;
    b) finansinių 2007–2011 metų JAV įmonių duomenų analizei;

c) 2003 m. Lietuvos savižudybių ir nužudymų skaičiaus duomenų analizei;
d) telekomunikacinių bendrovių akcijų duomenų analizei nuo 2012-01-20 iki 2012-04-01.

6. Statistinio modeliavimo būdu ištirtas MCMC algoritmų efektyvumas.

Trumpai apie autorę

Ingrida Vaičiulytė gimė 1984 m. sausio 4 d. Radviliškyje. 2002 m. baigė Radviliškio Vaižganto gimnaziją. 2006 m. Šiaulių universitete įgijo matematikos bakalauro laipsnį. 2009 m. Šiaulių universitete įgijo matematikos magistro laipsnį su mokytojo profesine kvalifikacija. 2006–2013 dirbo Šiaulių banke. Nuo 2012 m. dirba Šiaulių valstybinėje kolegijoje matematikos dėstytoja. 2009–2014 m. Vilniaus universiteto Matematikos ir informatikos instituto doktorantė.

El. paštas: ingrid.vaiciulyte@gmail.com