VILNIAUS UNIVERSITETAS

Gintautas Jakimauskas

DUOMENŲ TYRYBOS EMPIRINIŲ BAJESO METODŲ TYRIMAS IR TAIKYMAS

Daktaro disertacijos santrauka

Fiziniai mokslai, informatika (09 P)

Vilnius, 2014

Disertacija rengta 2013–2014 metais Vilniaus universiteto Matematikos ir informatikos institute.
Disertacija ginama eksternu.

Mokslinis konsultantas

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

Disertacija ginama Vilniaus universiteto Informatikos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Antanas Žilinskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Nariai:

prof. habil. dr. Kazys Kazlauskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),

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

prof. habil. dr. Rimantas Rudzkis (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P),

prof. dr. Gerhard-Wilhelm Weber (Vidurio Rytų technikos universiteto Taikomosios matematikos institutas, Ankara, Turkija, fiziniai mokslai, informatika – 09 P).

Disertacija bus ginama viešame Vilniaus universiteto Informatikos mokslo krypties tarybos posėdyje 2014 m. balandžio 15 d., 13 val. Vilniaus universiteto Matematikos ir informatikos instituto 203 auditorijoje.

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

Disertacijos santrauka išsiuntinėta 2014 m. kovo 14 d.
Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

VILNIUS UNIVERSITY

Gintautas Jakimauskas

ANALYSIS AND APPLICATION OF EMPIRICAL BAYES METHODS IN DATA MINING

Summary of Doctoral Dissertation

Physical sciences, Informatics (09 P)

Vilnius, 2014

The doctoral dissertation was prepared at the Vilnius University Institute of Mathematics and Informatics in 2013–2014.
The dissertation will be defended as external dissertation.

Scientific Consultant

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

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

Chairman:

Prof. Dr. Habil. Antanas Žilinskas (Vilnius University, Physical Sciences, Informatics – 09 P).

Members:

Prof. Dr. Habil. Kazys Kazlauskas (Vilnius University, Physical Sciences, Informatics – 09 P),

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

Prof. Dr. Habil. Rimantas Rudzkis (Vilnius University, Physical Sciences, Mathematics – 01 P).

Prof. Dr. Gerhard-Wilhelm Weber (Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey, Physical Sciences, Informatics – 09 P).

The dissertation will be defended at the public meeting of the Council of Informatics Science of Vilnius University in the auditorium 203 at the Vilnius University Institute of Mathematics and Informatics at 1 p.m. on 15 April, 2014.

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

The summary of the doctoral dissertation was distributed on 14 March, 2014.
A copy of the doctoral dissertation is available for review at the Library of Vilnius University.

1. INTRODUCTION

1.1. RESEARCH FIELD

The research field is data mining methods and algorithms of large dimensional data and rare events.

Efficient searching and using of the information, hidden in the data, is one of the most important factors in a dynamic field of the current research and business. Data mining is a contemporary field of information analysis, arising from the interconnection of database technologies, artificial intelligence, and statistical data analysis. Data mining has a wide scope, encompassing many methods, algorithms, software systems, and applications. If the usual methods of data analysis help to disclose the dependency variables, data mining is unique in that the result of the analysis is discovery of new dependencies unknown, or their existence was not even suspected. The modern data mining technology is based on the patterns, that represent the relationship between the data.

The main factors taken into account in solving data mining problems are as follows:

- a large amount of various types of information to be processed,
- results of the analysis could be presented to various users with different interests.

Data mining methods in decision making have two phases. The first one is taking advantage of a sample of the data collected, disclosing data structures and properties. The second one is forecasting and decision-making, based on the disclosed data structures and properties. The wide variety of the methods and algorithms reveals the complexity of data mining and its technologies adapted to the different situations. Often several methods and complex combinations of methods are used to solve data mining problems. The variety of the tasks and methods is complemented by a group of data mining algorithms. None of them are universal or beyond reproach. The choice of an algorithm takes into account the complexity of its operational and logical analysis, the time required, as well as the memory required, and reliability of the analysis.

The knowledge discovered using data mining methods and technologies is formally presented as a hypotheses on data models or estimates of model parameters, e.g., it is

necessary to assess whether the data have a tendency to form clusters or groups of objects, or certain features are correlated, and whether the data can be squashed without losing the essential information, etc. In data mining the following methods are widely used: clusterization, multidimensional scaling, classification, support vector machines, regression analysis, principal component analysis, etc.

The Bayesian decision theory is wide used in data mining, when information about the parameters is given in the form of a probability distribution function. The methods based on Bayesian decision theory have many advantages (see., e.g., DeGroot (1970), Carlin, Louis (1996), Rossi *et al.* (2003), Diaconis (2009), Press *et al.* (2007), Richey (2010)), as compared to the classical (*„frequentist“*) methods. However, these methods were started to be applied only in recent decades mainly because of two reasons. One of the reasons is the fact that they may not be objective, i.e., if the statistical calculations are performed without a thorough examination and evaluation of the characteristics of the object under consideration, it might make a subjective impact on the results. The other important reason is a rapid development of computer hardware from about 1980, because the Bayesian methods, even in quite simple cases, need a considerable amount of calculations.

1.2. RELEVANCE OF THE PROBLEM

Sufficiently accurate and fast processing of large data sets is one of the main goals of data mining. One of the solutions is usage of a more efficient hardware and software; however, this kind of solution is not always available in practice. Another solution is to replace the initial data set with a smaller data set preserving, much as possible, the main properties of the initial data set.

Data mining problems that use large populations and large dimensions often occur in biometrics, medicine, insurance, computer networks, etc. For example, estimation of rare events (e.g., probabilities of some disease, homicides, suicides, etc.) in large populations is of high relevance in statistical epidemiology. An adequate estimation of the probabilities of insured events can have a significant practical effect on the insurance.

Thus far, there is no generally accepted methodology for the multivariate nonparametric hypothesis testing. Traditional approaches to multivariate nonparametric

hypothesis testing are based on the empirical characteristic function (Baringhaus and Henze (1988)), nonparametric distribution density estimators and smoothing (Bowman and Foster (1993), Huang (1997)), as well as on the classical univariate nonparametric statistics calculated for data projected onto the directions found via the projection pursuit (L. Zhu, Fang, and Bhatti (1997), Szekely and Rizzo (2005)). A more advanced technique is based on the Vapnik-Chervonenkis theory, the uniform functional central limit theorem and inequalities for large deviation probabilities (see, e.g., Marcoulides, Hershberger (1997), Hirukawa (2012)). Recently, especially in applications, the Bayes approach and Markov chain Monte-Carlo methods have been widely used (see, e.g., Andrieu *et al.* (2003), Berg (2004), Asmussen, Glynn (2007), Sakalauskas, Vaičiulytė (2012), Vaičiulytė, Sakalauskas (2011)).

Hence the analysis and application of empirical Bayes methods and algorithms in testing nonparametric hypotheses if wee use large dimensional data and in estimating the parameters of statistical models of large populations are a relevant theoretical and practical data mining problem.

1.3. THE RESEARCH OBJECT

The research object is data mining empirical Bayes methods and algorithms applied in the analysis of large populations of large dimensions.

1.4. THE AIM AND OBJECTIVES OF THE RESEARCH

The aim and objectives of the research are to create methods and algorithms for testing nonparametric hypotheses for large populations and for estimating the parameters of data models.

The following problems are solved to reach these objectives:

1. To create an efficient data partitioning algorithm of large dimensional data.
2. To apply the data partitioning algorithm of large dimensional data in testing nonparametric hypotheses.
3. To apply the empirical Bayes method in testing the independence of components of large dimensional data vectors.

4. To develop an algorithm for estimating probabilities of rare events in large populations, using the empirical Bayes method and comparing Poisson-gamma and Poisson-Gaussian mathematical models, by selecting an optimal model and a respective empirical Bayes estimator.

5. To create an algorithm for logistic regression of rare events using the empirical Bayes method.

1.5. SCIENTIFIC NOVELTY

The following new results have been obtained:

1. A new binary data partitioning method is developed, based on the CART algorithm, which enables very fast and efficient partitioning of large dimensional data.

2. A new method for testing the independence of selected components of large dimensional data.

3. A new method for selecting the optimal model in the estimation of probabilities of rare events, using the Poisson-gamma and Poisson-Gaussian mathematical models and empirical Bayes estimators. A new nonsingularity condition in the case of the Poisson-gamma model is presented.

1.6. PRACTICAL SIGNIFICANCE OF THE WORK

The following practical results have been obtained:

1. Th presented data partitioning algorithm enables us to reduce the calculation time of clusterization procedures of multidimensional Gaussian mixtures.

2. The criterion presented for testing the independence of the components of high dimensional random vectors has higher power as compared to the classical criterion for larger dimensions.

3. The algorithms presented for empirical Bayes estimation of the model parameters were applied to the analysis of medical and sociological data, taking into account the nonsingularity condition of the Poisson-gamma model.

1.7. APPROVAL OF THE RESULTS

The research results have been presented in the following national and international conferences:

1. Jakimauskas, Gintautas. Gamma and logit models in empirical Bayesian estimation of probabilities of rare events // STOPROG 2012: Stochastic programming for implementation and advanced applications: international workshop, July 3-6, 2012, Neringa, Lithuania.
2. Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian estimation for Poisson-gamma model // 24th Mini EURO conference on continuous optimization and information-based technologies in the financial sector (MEC EurOPT 2010), Vilnius
3. Gurevičius, Romualdas; Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian estimation of small mortality rates // 5th international Vilnius conference [and] EURO-mini conference "Knowledge-based technologies and OR methodologies for decisions of sustainable development" (KORSD-2009): September 30 – October 3, 2009, Vilnius, Lithuania.
4. Sakalauskas, Leonidas; Jakimauskas, Gintautas; Sušinskas, Jurgis. Analysis of medical data by empirical Bayes method // Computer data analysis and modeling: complex stochastic data and systems: Ninth international conference: Minsk, September 7-11, 2010.
5. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Empirical Bayes testing goodness-of-fit for high-dimensional data // Computer data analysis and modeling: complex stochastic data and systems: Ninth international conference: Minsk, September 7-11, 2010.
6. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Testing of independency for high-dimensional data // Computer data analysis and modeling: complex stochastic data and systems: Eighth international conference: Minsk, September 11-15, 2007.
7. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Clustering and Testing in High-Dimensional Data, 8th Tartu Conference on Multivariate Statistics, Tartu, 26-29 June 2007.

8. Jakimauskas, Gintautas; Sušinskas, Jurgis. Testing independence of high-dimensional random vectors, Nordic Conference on Mathematical Statistics 2008, Vilnius, 16-19 June 2008.

1.8. STRUCTURE OF THE DISSERTATION

The dissertation consists of 5 chapters, references, and appendices.

Chapter 1 is introductory. In this chapter, we present a research field of the dissertation, relevance of the problem, the aim and objectives of the research, scientific novelty, practical significance of the work, approval of the results and publications.

In Chapter 2, we present a large dimensional data partitioning algorithm for data squashing in data mining and other implementations, which is one of the procedures of the software for classification of Gaussian mixtures, created in the Institute of Mathematics and Informatics, Vilnius. The methods using this data partitioning algorithm are used in the next chapter.

In Chapter 3, we present implementations of the data partitioning algorithm, given in Chapter 2.

In section 3.1, we present the procedure for testing the goodness-of-fit hypothesis that, for some k we have the standard Gaussian distribution in the complementary space $\mathbf{R}^{d-k}$, as well as multidimensional Gaussian mixture in the space $\mathbf{R}^k$ (testing of this hypothesis is used in the projection pursuit method). Th main results of this section are presented in the journal *Liet. mat. rink. LMD darbai* – Jakimauskas (2009).

In section 3.2, the procedure for testing the independence of components of large dimensional random vectors is presented. The main results of this section are published in the journal *Austrian Journal of Statistics* – Jakimauskas, Radavičius, Sušinskas (2008).

In section 3.3, we propose a more efficient statistics, for problems given in previous two sections, which is based on the empirical Bayes method. The main results of this section are presented in the journal *Liet. mat. rink. LMD darbai* – Jakimauskas, Sušinskas (2010).

In Chapter 4, we analyze the implementation of the empirical Bayes method in the problem of estimation of small probabilities (e.g., probabilities of some disease, suicides, etc.) in large populations.

In section 4.1, simulation of rare events using the empirical Bayes method, is considered. The main results of this section are presented in the proceedings of the Vilnius International Conference KORSD-2009 – Gurevičius, Jakimauskas, Sakalauskas (2009). We use the data submitted by the Lithuanian Institute of Hygiene.

In section 4.2, the Poisson-Gaussian and Poisson-gamma models are considered. The main results of this section are presented in the proceedings of the Neringa International Conference STOPROG-2012 – Jakimauskas (2012). We use the data from the Database of the Statistics Lithuania.

In section 4.3, we consider a modified regression Poisson-Gaussian model. The main results of this section are presented in the journal *Liet. mat. rink. LMD darbai* – Jakimauskas, Sakalauskas (2012). In this section we also use the data from the Database of the Statistics Lithuania.

In Chapter 5, we present the results and conclusions.

At the end of the dissertation, we present the list of references and appendices.

2. DATA SQUASHING IN DATA MINING

Data mining is a contemporary field of information analysis on the intersection of database technologies, artificial intelligence and statistical data analysis. Data mining is a very wide area, including many methods, algorithms and applied software systems.

Processing of large data sets is one of the main goals of data mining, applying interactive on-line analytical processing systems. In order to define the term ‘large data set’, we need to take into account the productivity of hardware and software, and the considered problem. Realization of some complex data models can raise great problems even for moderate data sizes.

One of the solutions is to use more efficient hardware and software, however this type of solution is not always available in practice. Another solution is to replace the initial data set by a smaller data set, preserving the main properties of the initial data set. A

trivial solution is a random selection of a smaller data set from the initial data set (random sampling). However, in this case, the variance of the parameters of the mathematical model can increase drastically, so it must be taken into account. Sampling methods are very widely used in the cases where collection of data is expensive (e.g., in demographics statistics, economics statistics, sociology research, etc.).

Assume that us have a large data set and we need to replace the initial data set by a smaller data set, preserving the main characteristics of the initial data set and using the information from all the elements of the initial data set. Various methods are used for this purpose, which can be divided into two groups: partitioning and hierarchical (Zhou, Sander, 2003) methods. Partitioning algorithms partition data set into clusters, hierarchical algorithms present a hierarchical cluster structure, however they do not define the clusters in explicit form.

Recently the data squashing method (DuMouchel et al, 1999) is wide used. This method 'squashes' the data so that the statistical analysis, using the squashed data, gives the results that are similar to that obtained using the full data set. In such a case, data analysis could be made using standard methods and the results are much more precise than that obtained using a random sample of the corresponding size. There is an example (Madigan et al., 2002) that, in the case of logistic regression with 750000 observations, the calculations with the squashed data set with 8443 observations have yielded 500 times less mean square error of regression coefficients than that with a randomly selected data set of size 7543. In this article, the data squashing method (likelihood-based data squashing) uses not necessarily a rectangular grid, so the partitions can be irregular.

One of the wide used methods in the classification and regression analysis is the CART (*classification and regression tree*) method (see, e.g. Hastie *et al*. (2001).

In this method the selected region (multidimensional rectangular parallelepiped) is step-by-step divided into partitions, until a certain stopping condition is fulfilled. In terms of graph theory, the steps performed by the CART method, can be described as a tree starting from the root node, then the nodes branching into two or more nodes connected to the parent node by edges. If, at each step, the node branches into two nodes (binary partitions), this method is a binary tree method. The CART method is mostly used (according to its name) for the classification and regression analysis.

Let us consider a modification of the CART algorithm for fast grouping of large data sets of large dimensions. This problem is similar to the application of the CART algorithm to the regression analysis, but, in our case, we use a large number of partitions and the simplest partitioning in order to minimize the calculation time. Our aim is to apply grouped data to a certain classification procedure of the Gaussian mixtures at a certain stage of this procedure in order to minimize the total calculation time.

Suppose, we have a sample X^N of size N. If N is sufficiently large, e.g., N=2000 (software limit to the abovementioned classification procedure N=10000), we can group the sample and obtain a shorter sample $X_{Grp}{}^{NGrp}$, where $NGrp$ is substantially smaller than N (e.g., $NGrp$ can be equal around 150). The first steps of the classification procedure are performed using initial sample X^N, then main calculations are performed with the grouped sample $X_{Grp}{}^{NGrp}$ (many steps: adding a new cluster, deleting a cluster, refinement of the parameters using iterations of the EM algorithm, testing goodness-of-fit, etc.). When the main calculations are finished, we return to the initial sample X^N and perform final refinements using the initial sample. It is very important to implement a fast grouping procedure, only in this case, we can get the total decrease in calculation time of both the grouping procedure and the classification procedure.

In 1991–1993, the Institute of Mathematics and Informatics (MII), Vilnius, the software for classification of multidimensional Gaussian mixtures was developed. This software was created in collaboration with the Central Economic-Mathematical Institute (CEMI), Moscow, where it was thoroughly tested and included int the software Class Master by the company Stat-Dialogue. This software later was distributed as the commercial software. One of the original algorithms in the software created in MII, is binary data partitioning procedure for large dimensional data which is used to reduce calculation time of the classification procedure.

Mathematical description of the algorithm. Let $\mathbf{X} = (X(1);\ldots;X(N))$ be a sample in $\mathbf{R}^d$, i.e.,

$$X(j) = (X_1(j), X_2(j),..., X_d(j))^{\mathrm{T}} \in \mathbf{R}^d,\ j = 1,2,...,N.$$

We will use counts of observations (this situation often occurs in practice) q_j, $j = 1, 2, \ldots, N$, i.e., the observation $X(j)$ is repeated q_j times. If there are no repeated observations, then simply $q_j = 1$, $j = 1, 2, \ldots, N$.

Let us have d-dimensional parallelepiped S, defined by intervals on each axis $[a_i, b_i]$, $i = 1, 2, \ldots, d$, (it can be selected freely, but it is recommended to use as small parallelepiped as possible, e.g., using the minimal and maximal values on each axis), such that all observations fit into this rectangular parallelepiped:

$$a_i \le X_i(j) \le b_i,\ i = 1,2,\ldots,d,\ j = 1,2,\ldots,N.$$

The aim of the procedure is to split a d-dimensional rectangular parallelepiped into sets of d-dimensional rectangular parallelepipeds (each of them contains at least one observation of the sample $\mathbf{X}$) $S(k) = \{S_k(1), S_k(2),\ldots, S_k(k)\}$, $k = 1,2,\ldots,k_{\max}$, $k_{\max} \le N$ (by definition $S(1) = S$), in such a way that the grouping error of the sample is minimized. The number M, $M > 1$, $M \in \{2, 4, 8, 16,\ldots\}$, is preset at the initial stage, which defines the smallest division of each axis. Let $Z_k(j)$, $j = 1, 2, \ldots, k$, be grouping points at each finished step k, and the corresponding total grouping error is

$$E_k^2 = E_k^2(\mathbf{X}, S(k)) = \sum_{j=1}^{k} E_k^2(j, \mathbf{X}, S(k)) = \sum_{j=1}^{k}\sum_{l=1}^{N} 1_{X(l)\in S_k(j)} \cdot q_l \cdot |X(l) - Z_k(j)|^2 .$$

After finishing the k-th step we calculate in advance the maximum possible decreases of the grouping error, dividing all the k rectangular parallelepipeds for each dimension. We select a rectangular parallelepiped and the axis with the maximum decrease of the grouping error to be in line for the next step. At the (k+1)-th step we split this rectangular parallelepiped into two rectangular parallelepipeds. We cut (several times, if needed) these two rectangular parallelepipeds, if one half contains no observations.

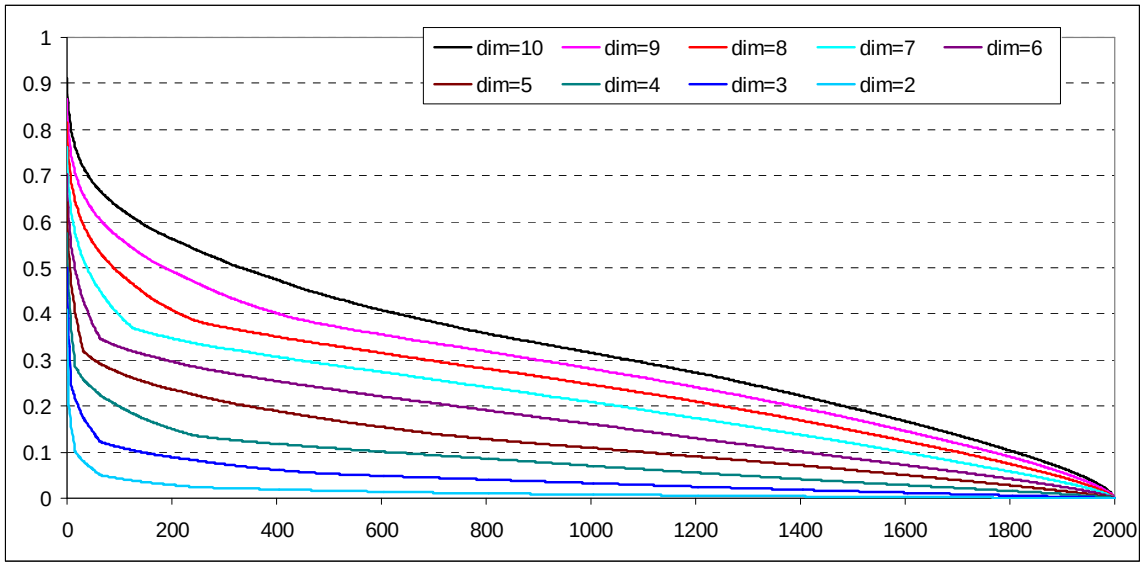

Fig. 2.1. Behaviour of the mean square grouping error depending on the dimension of uniformly distributed observations.

The algorithm stops when all the rectangular parallelepipeds are partitioned to the smallest size, defined by the number M. We can force a stop of the algorithm, if certain k is reached, or a certain mean square grouping error $(E_k^2 / N)^{1/2}$ is reached.

Simulation results. Let us consider uniformly distributed observations in a d-dimensional cube. We compare mean square grouping error $(E_k^2 / N)^{1/2}$ for various dimensions (see Fig. 2.1, here $N = 2000$). On the axis x we have the number of partitions k.

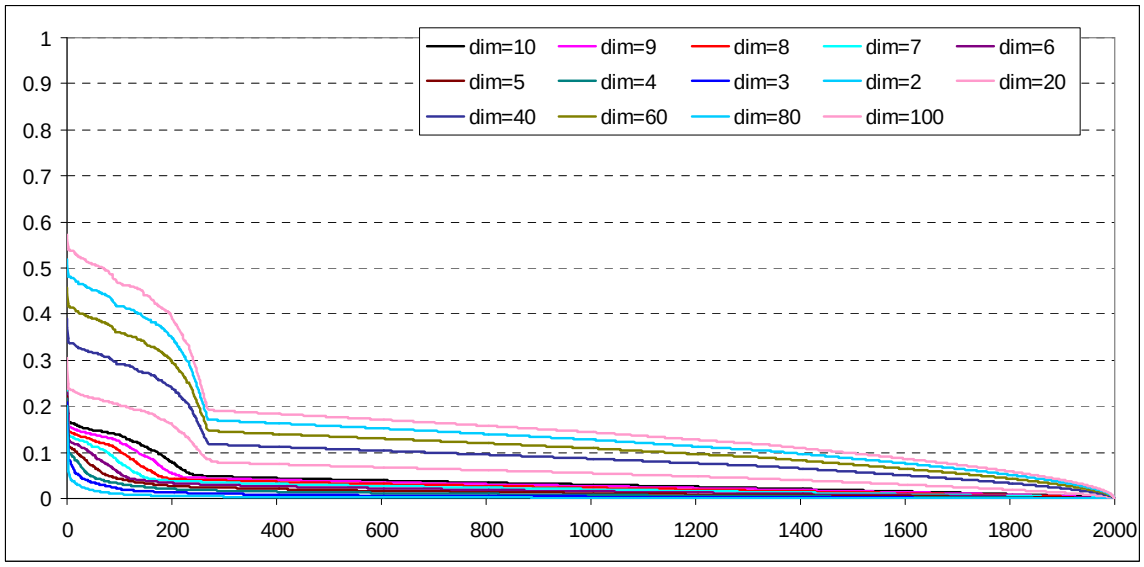

Fig. 2.2. Behaviour of the mean square grouping error depending on the dimension of observations of the 3-component Gaussian mixture.

Alternatively, let us consider observations well suited for the implementation of the abovementioned algorithm. Assume that we have observations in a d-dimensional cube (sample size $N = 2000$), a small part of which (3 per cent) is uniformly distributed and the other observations are the sample of the 3-component (equal probabilities) Gaussian mixture with unit covariance matrices, multiplied by the constant $c = 0.02$, and means (0.5–0.15, 0.5–0.1, 0.5, 0.5, ... , 0.5), (0.5, 0.5+0.2, 0.5, 0.5, ... , 0.5), (0.5+0.15, 0.5–0.1, 0.5, 0.5, ... , 0.5). The behaviour of $(E_k^2 / N)^{1/2}$ is shown in Fig. 2.2.

Conclusions. Partitioning algorithms are widely used for data squashing in order to use information of all the elements of the data. The presented partitioning algorithm is well suited for use if the data have a cluster structure. After a comparatively small number of partitions, we can reduce the mean square grouping error to a level such that we can successfully perform classification procedures of the Gaussian mixture, using a substantially smaller number of observations, and thus reducing the calculation time.

3. APPLICATION OF THE DATA PARTITIONING PROCEDURE IN TESTING GOODNESS-OF-FIT

In this chapter, application of the data partitioning procedure in testing goodness-of-fit is considered. We present implementations of this procedure for testing goodness-of-fit in the projection pursuit procedure, for testing the independence of components of a large dimensional random vector. We present a more powerful criterion using the empirical Bayesian approach.

When considering the testing goodness-of-fit, we can apply a certain direct method. Assume that we have distinct subsets A_1, A_2. ... , A_k, in a d-dimensional space, and that we know probabilities of a certain d-dimensional random variable of these subsets, i.e., $p_1 = \mathbf{P}(A_1)$, $p_2 = \mathbf{P}(A_2)$, ... , $p_k = \mathbf{P}(A_k)$. It is important that the sum of the probabilities is near to 1. Then goodness-of-fit of the d-dimensional data $X^N = (X_1, X_2, \dots, X_N)$ can be tested by comparing the probabilities $p_1, p_2, \dots, p_k$ with the corresponding empirical probabilities $q_1, q_2, \dots, q_k$, where q_j is the number of observations of X^N contained in the subset A_j, divided by N, $j = 1, 2, \dots k$.

Thus far, there is no generally accepted methodology for multivariate nonparametric hypothesis testing. Traditional approaches to multivariate nonparametric hypothesis

testing are based on the empirical characteristic function (Baringhaus, Henze (1988)), nonparametric density estimators and smoothing (Bowman, Foster (1993), Huang (1997)), and classical nonparametric statistics calculated for data projected onto the directions found via the projection pursuit (Zhu *et al.* (1997), Szekely, Rizzo (2005)). A more advanced technique is based on Vapnik-Chervonenkis theory, the uniform functional central limit theorem, and inequalities for large deviation probabilities (Vapnik (1988), Bousquet *et al.* (2004)).

In Jakimauskas *et al.* (2008), a simple data-driven and computationally efficient procedure is proposed for testing the independence of high-dimensional random vectors. The procedure is based on the randomization and bootstrap, sequentional data partitioning procedure, and χ^2-type statistics. In Jakimauskas (2009), it was implemented for testing goodness-of-fit in some stage of the projection pursuit algorithm. In Jakimauskas, Sušinskas (2010), a more powerful statistics as compared to χ^2-type statistics is proposed implementing the empirical Bayesian approach.

3.1. EFFICIENT ALGORITHM FOR TESTING GOODNESS-OF-FIT FOR CLASSIFICATION OF LARGE DIMENSIONAL DATA

Projection pursuit is used in data mining to reduce the dimension of initial data (see, e.g. Aivazyan S. A. (1996)).

Let $X = X^N$ be a sample of size N of d-dimensional Gaussian mixture (let the dimension d be large) with the distribution function (d.f.) F.

Because of the high dimension it is natural to project the sample X to linear subspaces of the dimension k (k = 1, 2, …) using a projection pursuit method (see, e.g., Aivazyan, S. A. (1996), Friedman, J. H. (1987)). Having the estimate of the discriminant subspace, provided by the projection pursuit method, it is easier to classify using the projected sample.

One of the problems in the projection pursuit method is calculation of a certain nonparametric estimate in a high-dimensional space. As an alternative we use Monte-Carlo method and the data partitioning procedure thus avoiding ‘curse of dimensionality’ (see, e.g., Hastie *et al.* (2001)). We use a joined sample of the initial sample and the simulated sample with the known distribution for which the hypothesis holds. The

number of observations of the each sample will be counted corresponding to partition elements, and a certain statistics will be used. The critical value of the criterion is obtained by simulating sufficient number of realizations for which the hypothesis holds. The efficiency of the criterion is based on a weak dependence on the dimension and the considered distribution.

Test statistics. Let $n^{j,k}$, resp., $m^{j,k}$, be the number of initial sample elements, respectively, the number of simulated sample elements, in the j-th element of the k-th partition of the joined sample. Define the χ^2-type statistics

$$T_k = \frac{S_k - (k-1)}{\sqrt{(2k-1)}}, k = 1, 2, \dots, K,$$

where

$$S_k = \frac{1}{2N} \sum_{j=1}^{k} (n^{j,k} - m^{j,k})^2, k = 1, 2, \dots, K,$$

Simulation results. Let us consider a sample of 3-component Gaussian mixture in a 10-dimensional space with means (–4, –1, 0, … , 0), (0, 2, 0, … , 0), (4, –1, 0, … , 0) and unit covariance matrices. It is known that the dimension of the discriminant subspace equals 2.

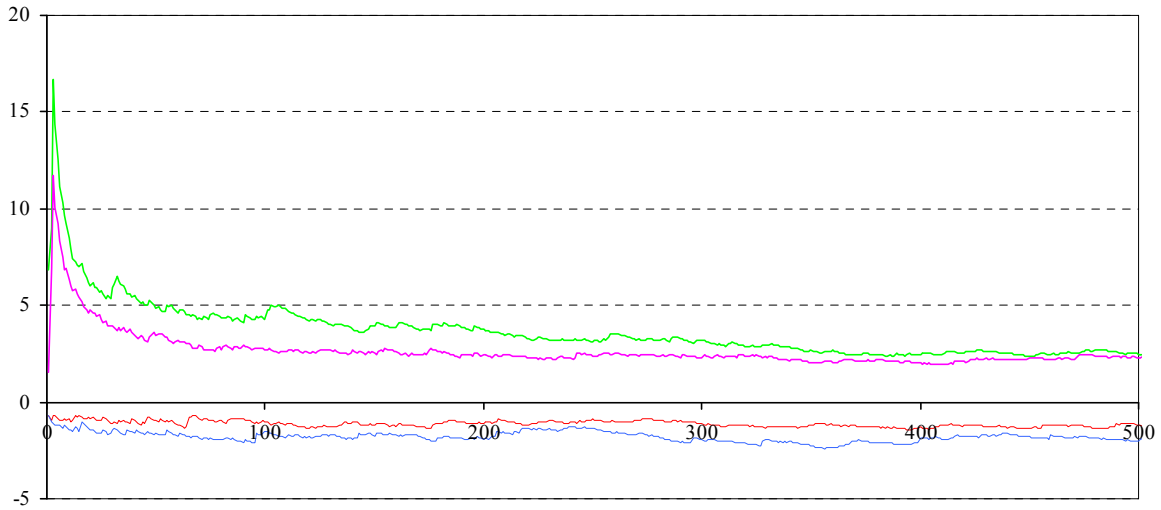

Fig. 3.1.1.: Minima and maxima of T_k (projection to a 1-dimensional subspace).

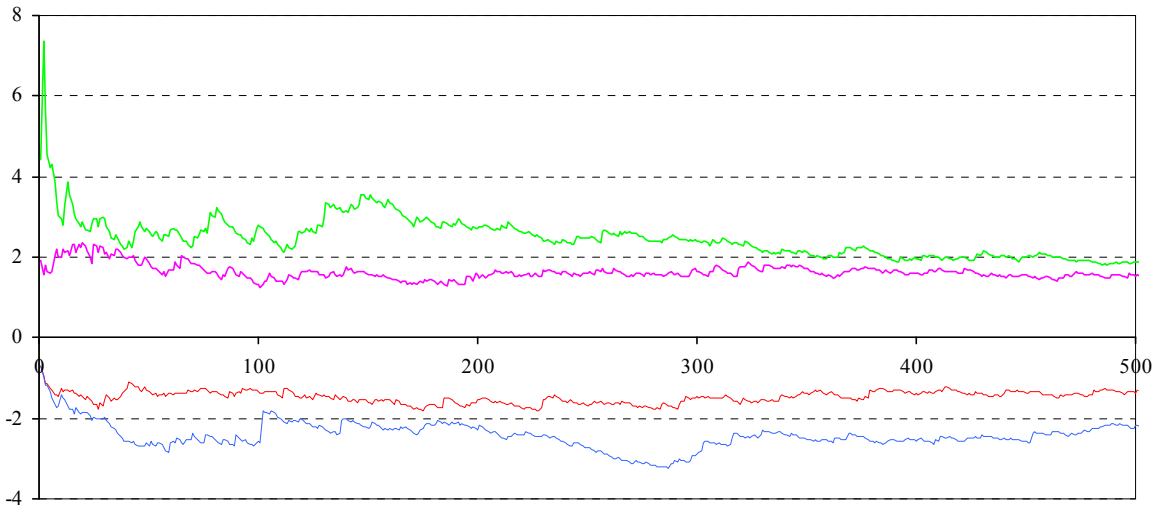

Fig. 3.1.2.: Minima and maxima of T_k (projection to a 2-dimensional subspace).

We compare the behaviour of the test statistics projecting data to a 1-dimensional subspace (i.e., the dimension that is not sufficient) and projecting data to 2-dimensional subspace (see Fig. 3.1.1, resp., Fig. 3.1.2). We present minima and maxima of the test statistics of 100 independent realizations and minima and maxima excluding 5 per cent of the largest and 5 per cent of the smallest values.

Conclusions. The simulation results show that there is a weak dependence on the dimension and distribution. The criterion based on the test statistics, excluding 5 per cent of the largest and 5 per cent of the smallest values is suitable to test the hypothesis on the dimension of the discriminant subspace.

3.2. TESTING THE INDEPENDENCE OF COMPONENTS OF LARGE DIMENSIONAL DATA

Our goal is to propose a relatively simple, data-driven and computationally efficient procedure for testing the independence of components of the d-dimensional random vector X, in case the dimension d is large. We will compare the power of the proposed criterion with that of classical criterion proposed by Blum, Kiefer, Rosenblatt (1961).

Let $\mathbf{X} = (X(1),\ldots,X(N))$ be a sample of the size N of i.i.d. observations of a random vector X with the distribution function F on $\mathbf{R}^d$. We are interested in testing some properties of F. Let $\mathcal{F}_H$ and $\mathcal{F}_A$ be disjoint classes of d-dimensional distributions.

Let us consider a non-parametric hypothesis testing problem $H: F \in \mathcal{F}_H$ vs. $A: F \in \mathcal{F}_A$. Testing the independence of two components $X_1 \in R^{d_1}$ and $X_2 \in R^{d_2}$, $d_1 + d_2 = d$, of $X = (X_1', X_2')$ corresponds to

$$\mathcal{F}_H = \left\{G : G(x) = G_1(x_1) \cdot G_2(x_2), x = (x_1, x_2), x_1 \in \mathbf{R}^{d_1}, x_2 \in \mathbf{R}^{d_2}\right\},$$

where G_1 and G_2 are marginal distributions of G that correspond to the components X_1 and X_2, respectively.

The proposed procedure is based on randomization and bootstrap, elements of the sequential testing, and data partitioning procedure. Monte-Carlo simulations are used to assess the performance of the procedure.

Test statistics. Let $\mathbf{X}^{(H)} := (X^{(H)}(1), \dots, X^{(H)}(M))$ be a sample of i.i.d. random vectors, for which the independence hypothesis holds. The joint sample is denoted by $\mathbf{Y}$.

We will use the χ^2-type statistics defined in the previous section. The critical value of c_α is obtained by the Monte-Carlo method using the selected significance level α.

Simulation results. To generate a sample from $F_H = F_1 \cdot F_2$, we simulate two independent samples with distribution F, and combine sample elements by taking first d_1 coordinates from the first sample and the remaining d_2 components from the second sample.

For the tests we use standard multivariate Student distribution with m degrees of freedom. Although the components of X are uncorrelated they are dependent. Since X converges in distribution to a standard Gaussian random vector as $m \to \infty$, the dependence of the components vanishes for large m.

The results of Monte Carlo simulations show that there is a weak dependence on the wide range of dimensions, sample sizes and null distributions.

The computer simulations were performed for $d \le 20$, $200 \le N, M \le 1000$, and $m = 1; 2, \dots; 7; 25; 100$. The dimensions d_1 and d_2 of the independent components X_1 and X_2, respectively, were chosen in two ways. In the first case $d_1 = d_2 = d/2$, and in the second case $d_1 = 1$, $d_2 = d - 1$. The typical number of simulations $R = 1000$. Below we present the results for $d = 20$ and $N = 1000$ (see Fig. 3.2.1).

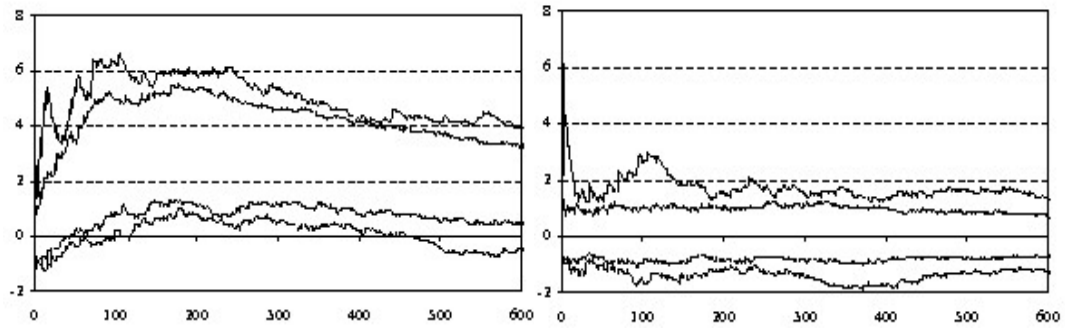

Fig. 3.2.1 Maxima, minima and two-side 0.9 confidence levels of statistic T_k for a sample from the Cauchy distribution ($m = 1$) and for the corresponding control data; $d = 20$, $d_1 = d_2 = 10$, $N = 1000$.

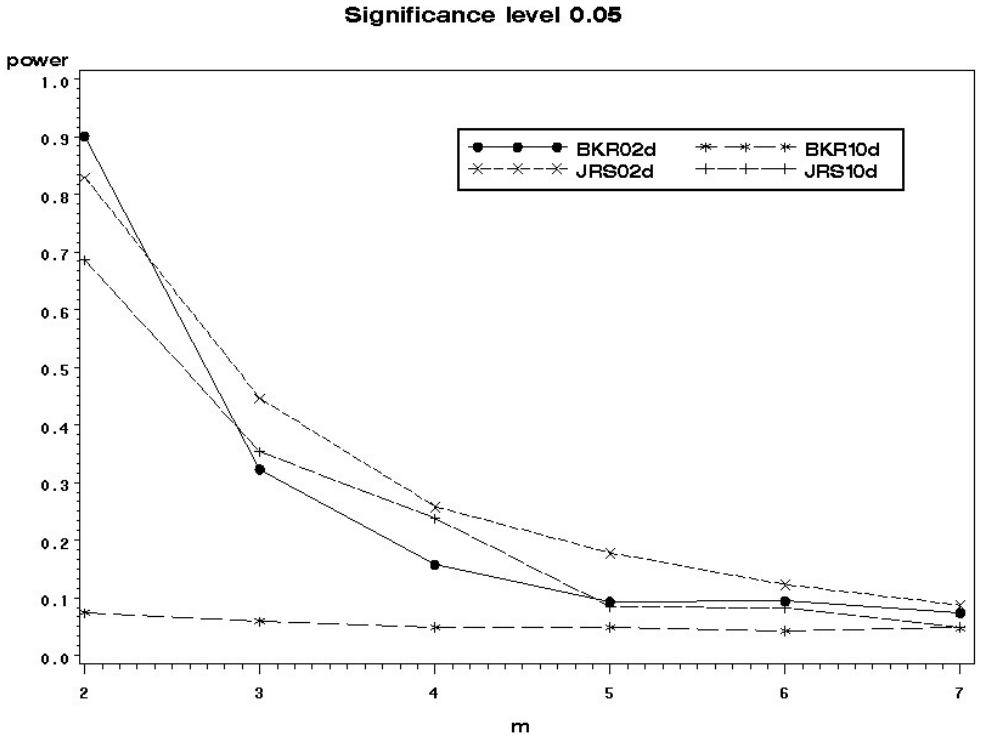

Fig. 3.2.2. Power functions of BKR and JRS tests, significance level $\alpha = 0.05$.

The proposed test procedure is referred to as JRS test (see Jakimauskas, Radavičius, Sušinskas (2008)) for brevity. The performance of the procedure is compared with the classical criterion of Blum, Kiefer, Rosenblatt (1961) (for brevity, BKR test) based on the Cramer-Von Mises-type test statistics for testing the independence:

$$\omega_{BKR}^2 = N \int\limits_{\mathbf{R}^{d_1}} \int\limits_{\mathbf{R}^{d_2}} (\hat{F}(u,v) - \hat{F}_1(u)\hat{F}_2(v))^2 \, d\hat{F}(u,v).$$

Here $\hat{F}_i$ is the empirical distribution function of the component X_i based on the sample $\mathbf{X}$ ($i = 1; 2$).

The power of the JRS test is compared with that of BKR test. To evaluate the power functions of the independence tests, Monte-Carlo simulations with $R = 1000$ realizations have been performed. The results are presented in Fig. 3.2.2 for the significance level $\alpha = 0.05$ and dimensions $d = 2$ and $d = 10$ with $d_1 = d_2 = d/2$. The power of the JRS test slightly decreases for growing dimension d, and for $d = 10$ it is close to the power of the BKR test for $d = 2$. The power of the BKR test for $d = 10$ is very low.

Conclusions. The results of the Monte Carlo simulations show that the proposed procedure is promising. It outperforms the classical Blum-Kiefer-Rosenblatt test even for low dimensional data. The dependence of the critical value c_α on the dimensionality d and the partition procedure is weak and can be reduced by imposing appropriate additional requirements on it.

3.3. APPLICATION OF THE EMPIRICAL BAYES APPROACH TO NONPARAMETRIC TESTING FOR HIGH-DIMENSIONAL DATA

The abovementioned χ^2-type statistics does not take into account the distribution of the number of elements in individual partition sets. In some cases (e.g., if distributions are similar in a large part of the space $\mathbf{R}^d$ and differ significantly only in a small part), the large number of insignificant deviations can mask a smaller number of significant deviations, thus reducing efficiency of hypothesis testing. We can apply a simple method: reject some part (e.g., one fourth) of the smallest deviations in absolute value and use for hypothesis testing only the deviations that are largest in absolute values. However, this simple method, besides advantages, has a disadvantage – the distribution of the statistics for a null hypothesis strongly depends on the distribution under the null hypothesis. Below we will present (based on Jiang, Zhang (2009)) the method based on the empirical Bayesian approach.

Let $\mathbf{X} = (X(1), \ldots, X(N))$ be a sample of the size N of i.i.d. observations of a random vector X with a distribution P on $\mathbf{R}^d$. We consider testing of nonparametric properties of P in case the dimension d of observations is large.

In Jakimauskas *et al.* (2008), a simple, data-driven and computationally efficient procedure of nonparametric testing for high-dimensional data has been introduced. The procedure is based on randomization and resampling (bootstrap), a special sequential data partition procedure, and χ^2-type statistics. The goal is to find a more efficient test statistics based on the nonparametric maximum likelihood (NML) and the empirical Bayes (NEB) estimators in an auxiliary nonparametric mixture model.

Auxiliary testing problem and empirical Bayes approach. Let us consider an auxiliary testing problem:

$$H_0^{\eta} : \mathbf{E}\,\eta_0 = 0_n \quad \text{vs.} \quad H_1^{\eta} : \mathbf{E}\,\eta_0 \neq 0_n .$$

where $\eta \sim \mathcal{N}(\theta, I_n)$ and $\theta \in \mathbf{R}^n$ is an unknown mean vector. In the (empirical) Bayes approach, the unknown parameter θ is treated as random. Thus, we consider a nonparametric Gaussian mixture model with a mixture distribution G:

$\eta = \theta + z$, θ and z are independent,

$z \sim \mathcal{N}(0_n, I_n)$,

$\theta_i \sim G$, $\{\theta_i, i = 1, 2, \dots, n\}$ are i.i.d. r.v's.

Computer simulation results. The following three alternatives of θ_i are analyzed:

(a1) $\theta_i = a u_i$, $u_i \sim \mathcal{N}(0, 1)$,

(a2) $\theta_i = a(2z_i - 1)$, $z_i \sim \mathcal{B}(1/2, 1)$,

(a3) $\theta_i = a(-1)^i \cdot \mathbf{1}\{i \le m\}$, $1 < m < n$.

For various combinations of the parameters a, n, and m, simulations with 1000 replications have been performed. The parameter $a > 0$ represents the difficulty of the testing problem. The simulations show some improvements in power of the NEB test in comparison with the power of the χ^2-type test. Fig. 3.3.1–3.3.3 illustrate the typical results. Here we give power plots for the empirical Bayes statistics and for χ^2 statistics depending on the parameter a. We use $n = 50$, and $m = 8$.

Conclusions. When applying the empirical Bayes method, we use an auxiliary Gaussian mixture model. Empirical Bayes estimates are obtained using the EM algorithm with some restrictions. The simulation results have showed a certain improvement of the power of the criterion in some natural cases (e.g., when distributions coincide on a large area and significantly differ only in a small area).

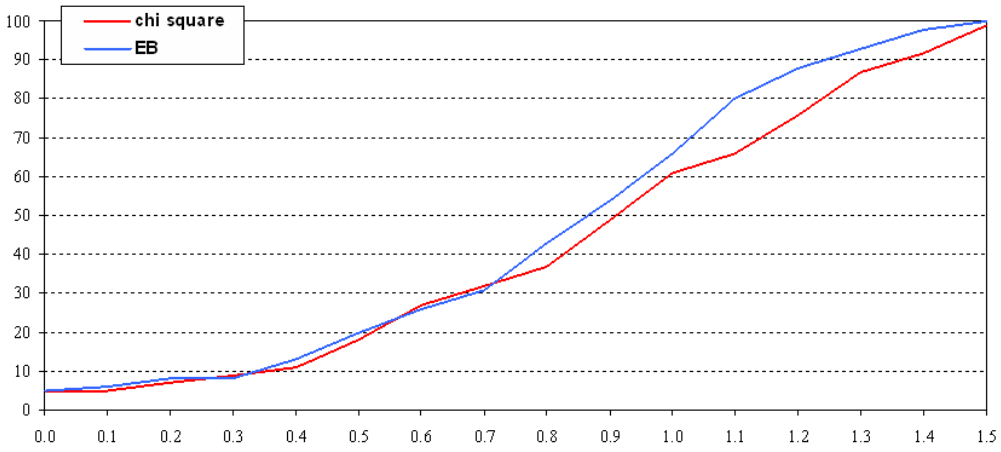

Fig. 3.3.1. Test power for the alternative (a1).

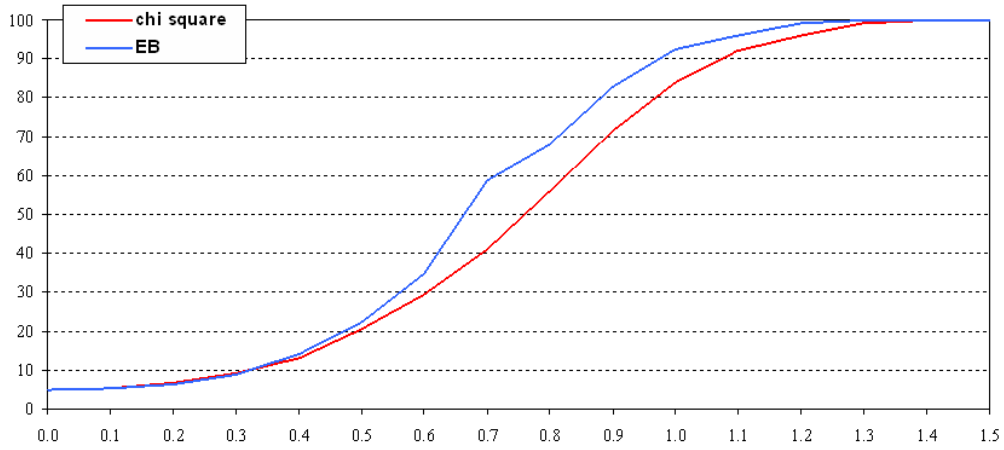

Fig. 3.3.2. Test power for the alternative (a2).

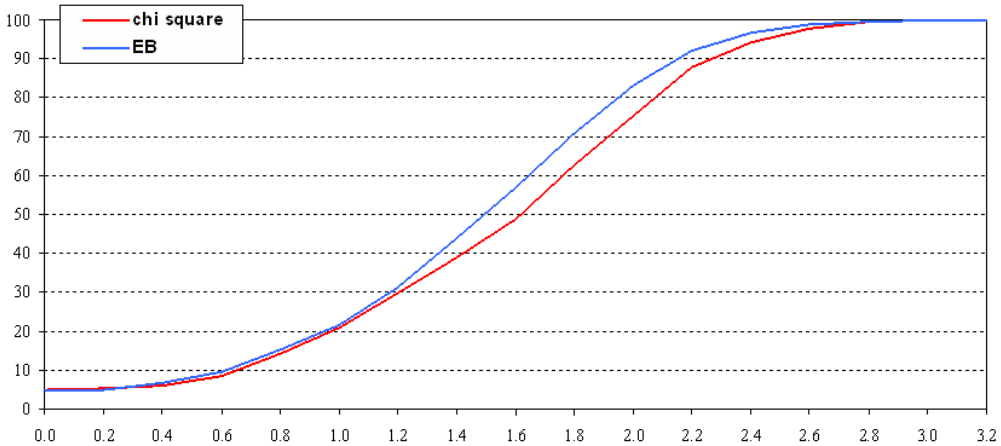

Fig. 3.3.3. Test power for the alternative (a3).

4. ANALYSIS OF PROBABILITIES OF RARE EVENTS USING THE EMPIRICAL BAYES METHOD

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. The classical estimator of probabilities is often not suitable because of too large estimation errors. Assume that the number of events in populations has the Poisson distribution with certain parameters. Note that this approximation is sufficiently accurate for large populations and for small (not too small) probabilities.

In the empirical Bayesian approach, an assumption is made that probabilities of events are random and have a certain distribution. It is well known that empirical Bayes estimators have a significantly less mean square error as compared to that obtained using simple mean relative risk estimates (see, e.g., Clayton, Caldor, (1987), Meza (2003), Sakalauskas (2010)).

4.1. SIMULATION OF RARE EVENTS USING THE EMPIRICAL BAYES METHOD

Let us consider the Lithuanian mortality data set in the period of 2003–2004 submitted by the Lithuanian Institute of Hygiene. The main purpose of mapping is to describe the geographical variation of mortality or decease in an attempt to demonstrate that a particular event may be caused by factors of a spatial structure. We investigate important numerical features of empirical Bayesian estimation techniques for the Poisson-Gaussian model, when a prior distribution of logits is normal with the parameters estimated by the maximum likelihood (ML) method (Tsutakava et al. (1985), Sakalauskas (2009)). We utilize a Lithuanian mortality data set of 2003–2004 years to estimate the underlying true risks and show the applicability of the approach considered.

Implementation in data analysis. The method developed has been applied in the data analysis on homicide and suicide mortality in 2003–2004 in Lithuania (all the events in population, for men and women). Integration and minimization of the ML function was performed using mathematical software MATHCAD and the Pascal programming language.

Let us consider K populations, where each population consists of N_j individuals. Assume that some event (e.g., suicide or homicide) can occur in these populations. Our goal is to estimate unknown probabilities of the events P_j, when the number of events Y_j, $j = \overline{1,K}$, in populations is observed. The mean relative risk estimate P of the probabilities is obtained dividing the total number of events by the total size of populations.

	$\sum_{j=1}^{K}(Y_j - N_j \cdot P)^2 / \sum_{j=1}^{K} Y_j$	P^*10^5	μ	σ
AllSuic	14255.1/1434=9.941	41.656	-7.652	0.281
AllHom	453.3/325=1.395	9.440	-9.260	0.136
MenSuic	9484.2/1199=8.297	74.582	-7.707	0.288
MenHom	356.0/232=1.534	14.431	-8.840	0.159
WomenSuic	692.9/256=2.705	13.952	-8.826	0.371
WomenHom	76.8/100=0.768	5.450	-9.817	0.000

Table 4.1.1. Empirical Bayesian estimation of suicide/homicide mortality in 2003.

	$\sum_{j=1}^{K}(Y_j - N_j \cdot P)^2 / \sum_{j=1}^{K} Y_j$	P^*10^5	μ	σ
AllSuic	15421/1381	40.334	-7.652	0.278
AllHom	287/294	8.587	-9.260	0.000
MenSuic	11889/1124	70.337	-7.707	0.305
MenHom	313/200	12.515	-8.840	0.234
WomenSuic	1573.2/257	14.075	-8.826	0.257
WomenHom	84.1/93	5.093	-9.817	0.000

Table 4.1.2. Empirical Bayesian estimation of suicide/homicide mortality in 2004.

We will use the Poisson-Gaussian model, where the probabilities are considered random, and their logits are independent Gaussian random variables with mean μ and variance σ^2. Some spatial analysis of data in administrative territories was made. The results of analysis of the nonsingularity condition and estimation of probabilities using the empirical Bayes method illustrated in Tables 4.2.1, 4.2.2.

Conclusions. Implementation of the empirical Bayesian approach allows detecting certain spatial effects of distribution of suicide/homicide probabilities in administrative territories of Lithuania.

4.2. GAMMA AND LOGIT MODELS IN THE EMPIRICAL BAYESIAN ESTIMATION OF PROBABILITIES OF RARE EVENTS

Let us consider the problem of estimation of small probabilities in large populations (e.g., estimation of probability of some disease, death, suicides, etc.). We consider two models of distribution of unknown probabilities: the probabilities have the gamma distribution (model (A)), or logits of the probabilities have Gaussian distribution (model (B)). We have selected real data from Database of Indicators of Statistics Lithuania (see *http://www.stat.gov.lt/*): Working-age persons recognized as disabled for the first time by the administrative territory (Table M3140706), in 2010 (number of populations $K = 60$). We have used average annual population data by the administrative territory (Table M3010211). We have obtained initial parameters (using simple iterative procedures described below) for models (A) and (B). At the second stage, we performed various tests using the Monte-Carlo simulation (using model (A) or model (B)) of sample data varying one selected parameter and obtaining maximum likelihood estimates by means of (independently) model (A) and model (B). The main goal was to select an appropriate model for a specified parameter set and to propose some recommendations for using gamma and logit models for the empirical Bayesian estimation.

Let us have K populations $A_1, A_2, \ldots, A_K$, consisting of N_j individuals, respectively, and some event (e.g., death or some disease), can occur in these populations. We observe the number of events $\{Y_j\} = Y_j$, $j = 1, 2, \ldots, K$.

We assume that the number of events is caused by unknown probabilities $\{\lambda_j\} = \lambda_j$, $j = 1, 2, \dots, K$, which are equal for each individual from the same population. Then the number of events $\{Y_j\}$ is a sample of independent random variables (r.v.) $\{\mathbf{Y}_j\} = \mathbf{Y}_j$, $j = 1, 2,\dots, K$, with a binomial distribution (resp., with the parameters (λ_j, N_j), $j = 1, 2, \dots, K$. Clearly,

$$\mathbf{E}\mathbf{Y}_j = \lambda_j N_j, j = 1, 2, \dots, K. \tag{1}$$

An assumption is often made (see, e.g., (Tsutakawa *et al.*, 1985), (Clayton, Caldor, 1987)), that r.v.'s $\{\mathbf{Y}_j\}$ have a Poisson distribution with the parameters $\lambda_j N_j$, $j = 1, 2,\dots, K$,

$$\mathbf{P}\{\mathbf{Y}_j = m\} = h(m, \lambda_j N_j), m = 0, 1, \dots; \; j = 1, 2, \dots, K, \tag{2}$$

where

$$h(m, z) = e^{-z}\frac{z^m}{m!}, \;\; m = 0, 1, \dots, \; z > 0, \tag{3}$$

Under such an assumption, (1) hlds as well.

We consider the mathematical model assuming that unknown probabilities $\{\lambda_j\}$ are independent identically distributed (i.i.d.) r.v.'s with the distribution function F from a certain class $\mathcal{F}$. Our problem is to find estimates of unknown probabilities $\{\hat{\lambda}_j\}$ from the observed number of events $\{Y_j\}$, assuming that $F \in \mathcal{F}$.

Assume that $\{\lambda_j\}$ are i.i.d. gamma r.v's with a shape parameter $\nu > 0$ and a scale parameter $\alpha > 0$, i.e., d.f. F has a distribution density

$$f(x) = f(x; \nu, \alpha) = \frac{\alpha \cdot (\alpha \cdot x)^{\nu-1}}{\Gamma(\nu)} e^{-\alpha x}, \;\; 0 \le x < \infty. \tag{4}$$

Then, $\mathbf{E}\lambda_j = \nu / \alpha$, and $\mathbf{D}\lambda_j = \nu / \alpha^2$. Moreover,

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = \frac{Y_i + \nu}{N_j + \alpha}, \;\; j = 1, 2, \ldots, K. \tag{5}$$

Denote this model by (A).

Regardless of distribution of $\{\lambda_j\}$ we can use the mean relative risk (MRR) estimate

$$\overline{\lambda}^{MRR} = \frac{\sum_{k=1}^{K} Y_k}{\sum_{k=1}^{K} N_k}, \tag{6}$$

so we assume that $\{\overline{\lambda}_j^{MRR}\} \equiv \overline{\lambda}^{MRR}$, $j = 1, 2, \ldots, K$. Also we can use the relative risk (RR) estimate $\{\overline{\lambda}_j^{RR}\} = \overline{\lambda}_j^{RR}$, $j = 1, 2, \ldots, K$, where

$$\overline{\lambda}_j^{RR} = \frac{Y_j}{N_j}, \;\; j = 1, 2, \ldots, K. \tag{7}$$

Let us consider an empirical Bayes estimate $\{\hat{\lambda}_j\}$, that is a certain compromise between the mean relative risk estimate $\{\overline{\lambda}_j^{MRR}\}$ and the relative risk estimate $\{\overline{\lambda}_j^{RR}\}$. This estimate is obtained by (5) using parameter estimates $(\hat{\nu}, \hat{\alpha})$.

Alternatively, we consider Bayes estimate $\{\widetilde{\lambda}_j\}$, obtained under the assumption that unknown probabilities are i.i.d. r.v.'s, such that their logits

$$\alpha_j = \ln \frac{\lambda_j}{1 - \lambda_j}, \; j = 1, 2, \ldots, K, \tag{8}$$

are i.i.d. Gaussian r.v.'s with mean μ and variance σ^2. Denote this model by (B). In this case, the conditional expectation of $\{\lambda_j\}$ is of the following form (see (Sakalauskas (1995), Gurevičius *et al.*, (2009)):

$$\mathbf{E}(\lambda_j \mid \mathbf{Y}_j = Y_j) = \frac{\int_{-\infty}^{\infty} \frac{1}{1 + e^{-x}} h(Y_j, \frac{N_j}{1 + e^{-x}}) \varphi(x; \mu, \sigma^2) dx}{D_j(\mu, \sigma^2)}, \tag{9}$$

$$D_j(\mu,\sigma^2)=\int_{-\infty}^{\infty}h(Y_j,\frac{N_j}{1+e^{-x}})\varphi(x;\mu,\sigma^2)dx. \tag{10}$$

When considering model (A), the respective maximum likelihood function is of the following form:

$$L_A(\nu,\alpha)=\sum_{j=1}^{K}\left(\ln\frac{\Gamma(Y_j+\nu)}{\Gamma(\nu)}+\nu\ln(\alpha)-(Y_j+\nu)\ln(N_j+\alpha)+Y_j\ln N_j\right) \tag{11}$$

In this case, if for the number of observed events $\{Y_j\}=Y_j$, $j=1,\ 2,\ \ldots\ ,K$, in populations A_1, A_2, … , A_K, consisting of N_j, $j=1,\ 2,\ \ldots\ ,K$, individuals, the non-singularity condition

$$\sum_{j=1}^{K}Y_j-\sum_{j=1}^{K}(Y_j^2-(N_j\cdot P)^2)<0\,, \tag{22}$$

holds, then there exists a maximum of the maximum likelihood function (11) for some finite α and ν, because, for sufficiently small values of α and ν, the derivative by ν of the maximum likelihood function is larger than zero.

Considering the model (B), the corresponding maximum likelihood function is of the following form:

$$L_B(\mu,\sigma^2)=\sum_{j=1}^{K}\left(\ln D_j(\mu,\sigma^2)\right). \tag{12}$$

The maximum likelihood estimates are obtained by maximizing (11), resp., (12) and replacing parameter values in (5) or (10) by (ν^*,α^*) or $(\mu^*,(\sigma^*)^2)$. In practice, approximate estimates $\{\hat{\lambda}_j\}$ and $\{\widetilde{\lambda}_j\}$ are obtained using numerical methods (usually iterative procedures) for finding the approximate parameter values $(\hat{\nu},\hat{\alpha})$, resp. $(\widetilde{\mu},\widetilde{\sigma}^2)$.

Simulation results. As initial data for the simulation of $\{Y_j\}$ we have selected real data from Database of Indicators of Statistics Lithuania: Working-age persons

recognized as disabled for the first time by the administrative territory, in 2010 (number of populations ($K = 60$), 9 data sets in total (Table M3140706).

We have used average annual population data by the administrative territory (Table M3010211); the total population in all administrative territories equals 3286820.

At the initial stage, real sample data were evaluated using model (A) and model (B), i.e. we obtained starting estimates $\{\hat{\lambda}_j\}_0$ and $\{\tilde{\lambda}_j\}_0$. The simulation and estimation were performed by both models for 1000 independent realizations, and the respective values of maximum likelihood function were compared (see Fig.'s 4.2.1, 4.2.2). Also simulation of number of events using various values of α and ν, was performed as well.

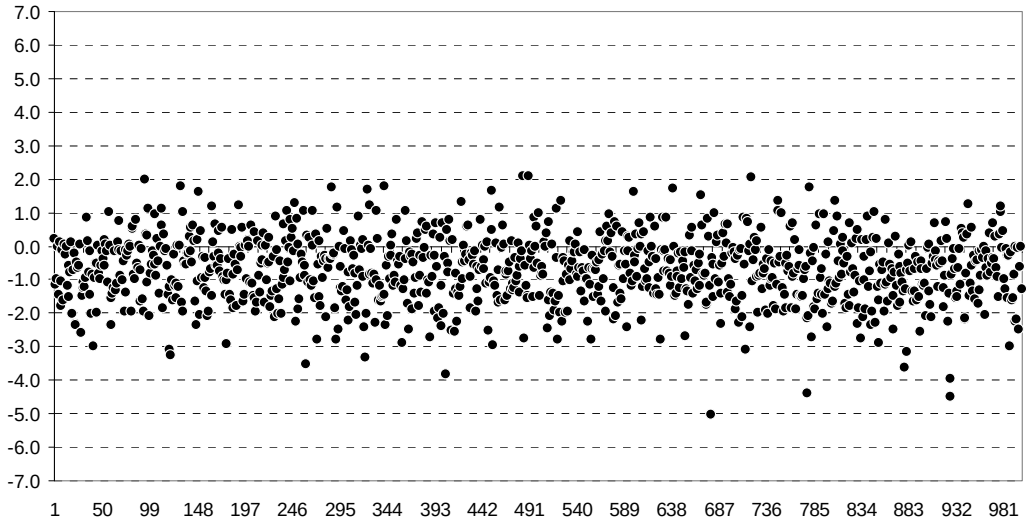

Fig. 4.2.1. Differences of $L_A - L_B$ (simulation by model (A), data set no. 1).

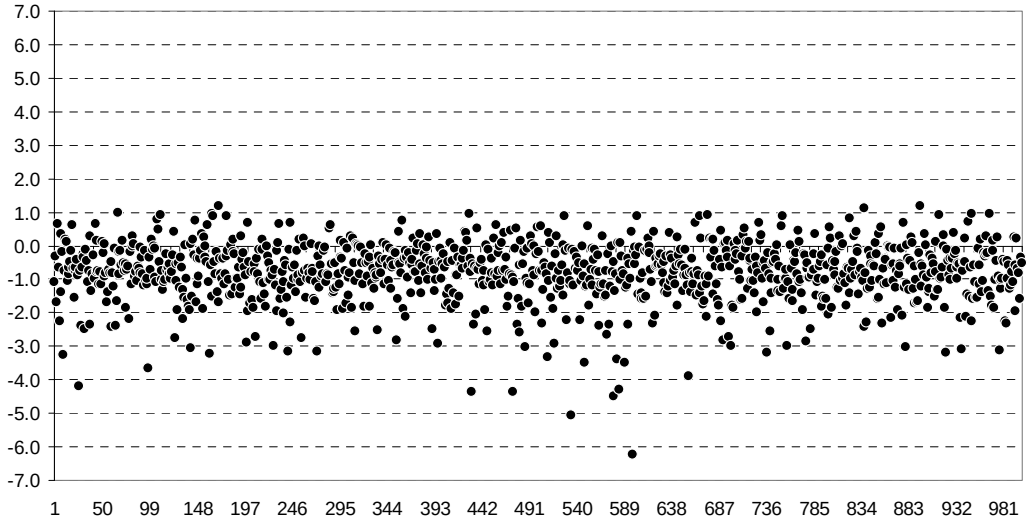

Fig. 4.2.2. Differences of $L_A - L_B$ (simulation by model (B), data set no. 1).

Conclusions. The results show that for most realizations of data sets 1–9 (simulation by model (A), and simulation by model (B)) are typically $L_B(\tilde{\mu},\tilde{\sigma}^2) > L_A(\hat{\nu},\hat{\alpha})$. For simulated data sets, a preferable model depends mostly on the value ν/α, specifically, for lower values the Poisson-gamma model is preferable, and for bigger values the Poisson-Gaussian model becomes preferable. The results show that the Monte-Carlo simulation method enables us to determine which estimation model is preferable.

4.3. EMPIRICAL BAYESIAN REGRESSION MODEL FOR ESTIMATION OF PROBABILITIES OF RARE EVENTS

The problem of adding an additional regression variable to the Poisson-Gaussian model when estimating probabilities of rare events in large populations is investigated here. We deal with two models of distribution for unknown probabilities: the probabilities have the gamma distribution (Poisson-gamma model, model (A)), or, alternatively, logits of the unknown probabilities have the Gaussian distribution (Poisson-Gaussian model, model (B)). In a modified regression model (B), the additional regression variable will be used for the mean of Gaussian distribution (model BR).

As a basis for the regression variable we use real data from Database of Indicators of Statistics Lithuania: Number of hospital discharges by administrative territory, in 2010 (Table M3140312).

Mathematical models. We use the same models as in the previous section. In addition, we introduce model (BR), adding an auxiliary regression variable Z_j, assuming that $\mu = \mu(j) = \mu_0 + \mu_1 Z_j$, $j = 1, 2, \ldots, K$. This variable is considered non-random.

Computer simulation results. The results show that the Monte-Carlo simulation method enables us to determine which estimation model is preferable and, in some cases, model (BR) is preferable, whilst model (A) is preferable to model (B) (see Fig.'s 4.3.1, 4.3.2).

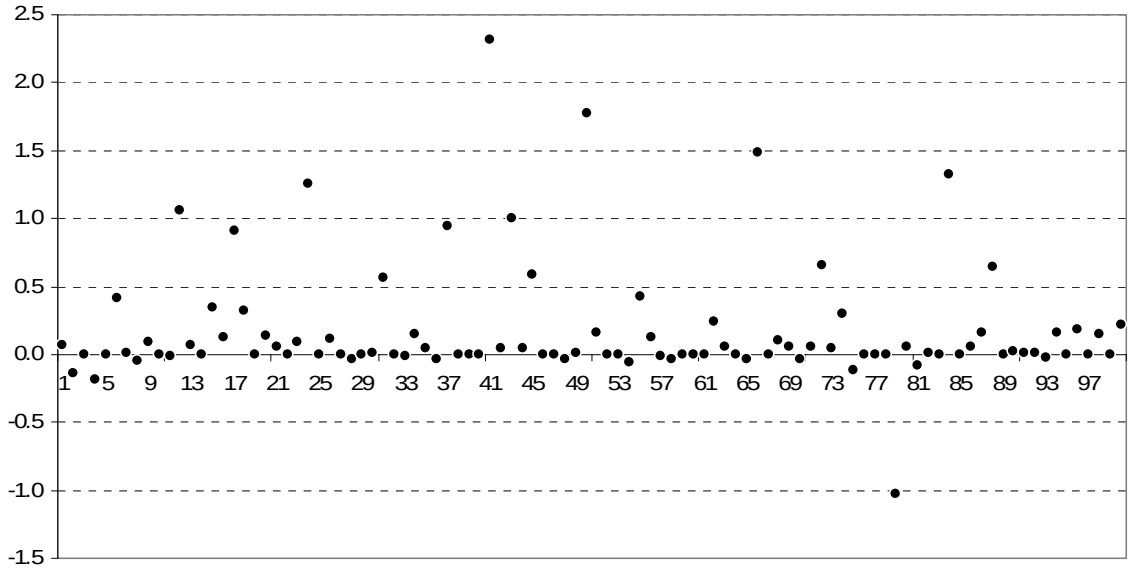

Fig. 4.3.1. Difference of L_A–L_B (data set 6, simulation by model (B).

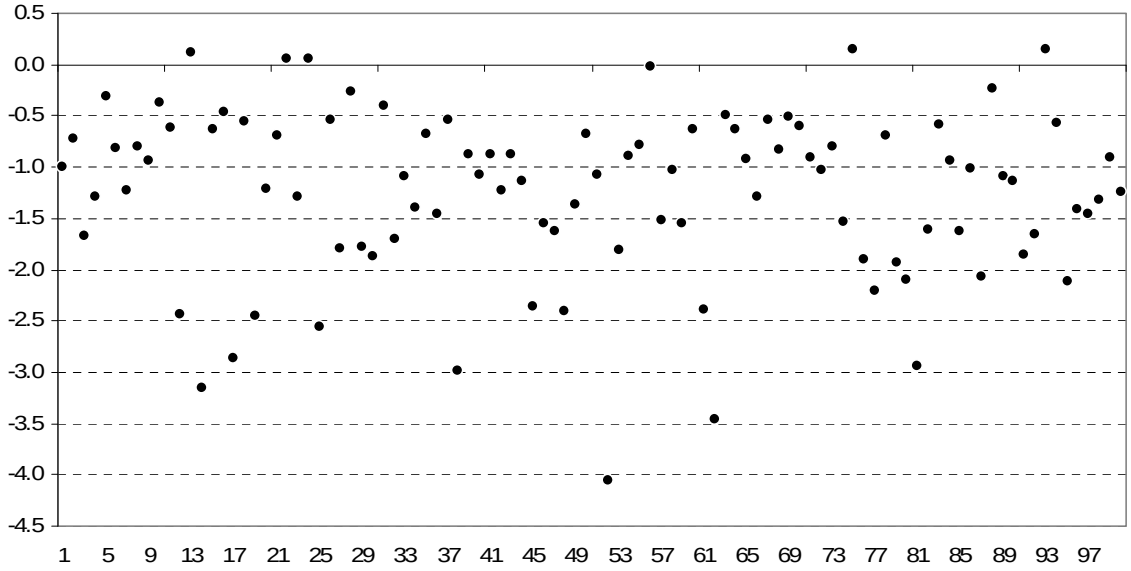

Fig. 4.3.2. Difference of L_A–L_{BR} (data set 6, simulation by model (B).

5. RESULTS AND CONCLUSIONS

The following new results have been obtained:

1. The presented binary data partitioning algorithm of large dimensional data enables us to perform efficient data partitioning with implementations in the classification of Gaussian mixtures and testing goodness-of-fit.

2. The new method for testing independence of components of random vectors enables testing the independence of large dimensional random vectors.

3. The new method of selecting the method for estimating probabilities of rare events in populations allows us to select the preferable model for empirical Bayes estimation.

The following practical results have been obtained:

1. The presented data partitioning algorithm enables us to reduce the calculation time of clusterization procedures of multidimensional Gaussian mixtures.

2. Th criterion proposed for testing the independence of components of high-dimensional random vectors has a higher power as compared to the classical criterion for larger dimensions.

3. The algorithms presented for empirical Bayesian estimation of model parameters were applied in the analysis of medical and sociological data, taking into account the nonsingularity condition of the Poisson-gamma model.

The following conclusions can be drawn:

1. The binary data partitioning procedure is best suited for data with an explicit cluster structure. After a comparatively small number of steps of the procedure we can achieve quite a small value of mean grouping error, so we can use the grouped data for the classification by a selected distribution model, reducing the calculation time.

2. The method of testing goodness-of-fit has a weak dependence on the dimension and on the distribution, thus allowing us to use minima and maxima of the test statistics as a criterion to test goodness-of-fit.

3. The criterion for testing the independence of components of random vectors has a higher power as compared to the classical Blum-Kiefer-Rosenblatt criterion for larger dimensions.

4. Implementation of the empirical Bayesian approach enables us to construct more powerful criterion for testing nonparametric hypotheses as compared to the χ^2-type criterion in some natural cases (e.g., when distributions coincide on a large area and significantly differ only in a small area). Note that this criterion requires considerably larger amount of calculations.

5. In the estimation of probabilities of rare events in some cases of estimation of the parameters we have a singularity issue. The nonsingularity condition for the Poisson-gamma model enables us to avoid using such data for estimation.

6. The numerical algorithm for empirical Bayes estimation enables us to obtain the estimates of the parameters of the Poisson-gamma model, including the nonsingularity condition.

7. The Monte-Carlo simulation is helpful in selecting the preferable model for empirical Bayes estimation with regard to Poisson-gamma, Poisson-Gaussian and modified regression Poisson-Gaussian models.

List of publications of the author

1. G. Jakimauskas. Efficiency analysis of one estimation and clusterization procedure of one-dimensional Gaussian mixture // Informatica, ISSN 0868-4952, 8(3), 1997, p. 331-343.
2. G. Jakimauskas, R. Krikštolaitis. Influence of projection pursuit on classification errors: computer simulation results // Informatica, ISSN 0868-4952, 11(2), 2000, p. 115-124.
3. G. Jakimauskas, R. Krikštolaitis. Bootstrap methods in selection of the discriminant subspace // Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISSN 0132-2818, T. 40, 2000, p. 281-286.
4. G. Jakimauskas. Procedure of the removal of the outliers from the sample satisfying the multidimensional Gaussian mixture model // Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISBN 9986-680-16-6, T. 42, 2002, p. 523-528.
5. Jakimauskas, Gintautas; Radavičius, Marijus; Sušinskas, Jurgis. A simple method for testing independence of high-dimensional random vectors // Austrian journal of statistics, ISSN 1026-597X, Vol. 37, no. 1, 2008, p. 101-108.
6. Jakimauskas, Gintautas. Efficient algorithm for testing goodness-of-fit for classification of high dimensional data // Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISSN 0132-2818, T. 50, 2009, p. 293-297.
7. Jakimauskas, Gintautas; Sušinskas, Jurgis. Application of the empirical Bayes approach to nonparametric testing for high-dimensional data // Lietuvos matematikos

rinkinys. Lietuvos matematikų draugijos darbai, ISSN 0132-2818, T. 51, 2010, p. 402-407.

8. Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian regression model for estimation of small rates // Lietuvos matematikos rinkinys. Lietuvos matematikų draugijos darbai, ISSN 0132-2818, T. 53, ser. A, 2012, p. 42-47.
9. Gurevičius, Romualdas; Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian estimation of small mortality rates // 5th international Vilnius conference [and] EURO-mini conference "Knowledge-based technologies and OR methodologies for decisions of sustainable development" (KORSD-2009): September 30 – October 3, 2009, Vilnius, Lithuania. Vilnius: Technika, ISBN 9789955284826, 2009, p. 290-295.
10. Jakimauskas, Gintautas; Sakalauskas, Leonidas. Empirical Bayesian estimation for Poisson-gamma model // 24th Mini EURO conference on continuous optimization and information-based technologies in the financial sector (MEC EurOPT 2010). Vilnius: Technika, ISBN 9789955285984, 2010, p. 254-257.
11. Jakimauskas, Gintautas. Gamma and logit models in empirical Bayesian estimation of probabilities of rare events // STOPROG 2012: Stochastic programming for implementation and advanced applications: proceedings of international workshop, July 3-6, 2012, Lithuania. Vilnius: Technika, ISBN 9786099524146, 2012, p. 43-48.
12. Radavičius, Marijus; Jakimauskas, Gintautas. Robust projection pursuit // Computer data analysis and modeling: complex stochastic data and systems: proceedings of the Seventh international conference: Minsk, September 6-10, 2004. Vol. 1. Minsk: Publishing centre BSU, ISBN 985-445-492-4, 2004, p. 114-117.
13. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Testing of independency for high-dimensional data // Computer data analysis and modeling: complex stochastic data and systems: proceedings of the Eighth international conference: Minsk, September 11-15, 2007. Vol. 1. Minsk: Publishing centre BSU, ISBN 9789854765082, 2007, p. 174-177.
14. Radavičius, Marijus; Jakimauskas, Gintautas; Sušinskas, Jurgis. Empirical Bayes testing goodness-of-fit for high-dimensional data // Computer data analysis and modeling: complex stochastic data and systems: proceedings of the Ninth international

conference: Minsk, September 7-11, 2010. Vol. 1. Minsk: Publishing center BSU, ISBN 9789854768472, 2010, p. 199-202.

15. Sakalauskas, Leonidas; Jakimauskas, Gintautas; Sušinskas, Jurgis. Analysis of medical data by empirical Bayes method // Computer data analysis and modeling: complex stochastic data and systems: proceedings of the Ninth international conference: Minsk, September 7-11, 2010. Vol. 1. Minsk: Publishing center BSU, ISBN 9789854768472, 2010, p. 203-206.

Short information about the author

Birth date and place

December 3, 1956, Utena, Lithuania.

Education

1974 Utena High School.

1979 Vilnius University, Faculty of Mathematics, master's degree in mathematics.

Work experience

1979–1980 – engineer mathematician-programmer, Institute of Mathematics and Cybernetics, Probability Theory Sector (from 14 April, 1980 – Department).

1980–1983 – doctoral student, Institute of Mathematics and Cybernetics, Probability Theory Department, Time Series Analysis Sector.

1983–1989 – junior researcher, Institute of Mathematics and Cybernetics, Probability Theory Department, Time Series Analysis Sector (from 1988 – Applied Statistics Department).

1989–2003 – researcher, Institute of Mathematics and Cybernetics (from 9 July, 2990 – Institute of Mathematics and Informatics), Applied Statistics Department.

1995–2012 – chief specialist (half staff) Statistics Lithuania, Econometric Studies Consultants Division (later – Econometric Studies Division, Methodology and Quality Division).

2003–2009 – junior researcher, Institute of Mathematics and Informatics, Applied Statistics Department.

2009 to date – specialist, Institute of Mathematics and Informatics (from October, 2010 –

Vilnius University Institute of Mathematics and Informatics), System Analysis Department, Operational Research Sector.

Reziumė

Darbo tyrimų objektas yra duomenų tyrybos empiriniai Bajeso metodai ir algoritmai, taikomi didelio matavimų skaičiaus didelių populiacijų duomenų analizei.

Darbo tyrimų tikslas yra sudaryti metodus ir algoritmus didelių populiacijų neparametrinių hipotezių tikrinimui ir duomenų modelių parametrų vertinimui.

Šiam tikslui pasiekti yra sprendžiami tokie uždaviniai:

1. Sudaryti didelio matavimo duomenų skaidymo algoritmą.
2. Pritaikyti didelio matavimo duomenų skaidymo algoritmą neparametrinėms hipotezėms tikrinti.
3. Pritaikyti empirinį Bajeso metodą daugiamačių duomenų komponenčių nepriklausomumo hipotezei tikrinti su skirtingais matematiniais modeliais, nustatant optimalų modelį ir atitinkamą empirinį Bajeso įvertinį.
4. Sudaryti didelių populiacijų retų įvykių dažnių vertinimo algoritmą panaudojant empirinį Bajeso metodą palyginant Puasono-gama ir Puasono-Gauso matematinius modelius.
5. Sudaryti retų įvykių logistinės regresijos algoritmą panaudojant empirinį Bajeso metodą.

Darbo metu gauti šie nauji rezultatai:

1. Didelio matavimo duomenų binarinio skaidymo metodas, paremtas erdvės skaidymu, naudojamu duomenų klasifikavime, kuris įgalina atlikti didelio matavimo duomenų skaidymą, pritaikomą duomenų grupavimui bei neparametrinių hipotezių tikrinimui.
2. Naujas metodas didelio matavimo nekoreliuotų duomenų pasirinktų komponenčių nepriklausomumo tikrinimui.
3. Naudojant skirtingus matematinius modelius parengtas naujas metodas parenkant didelių populiacijų retų įvykių optimalų modelį ir atitinkamą empirinį Bajeso įvertinį. Pateikta nesinguliarumo sąlyga Puasono-gama modelio atveju.

Disertaciją sudaro 5 skyriai, literatūros sąrašas ir priedai.

1-asis skyrius yra įvadinis. Jame pateikiama disertacijos tyrimų sritis, problemos aktualumas, tyrimų objektas, tyrimų tikslas ir uždaviniai, mokslinis naujumas, praktinė darbo reikšmė bei darbo rezultatų aprobavimas ir publikavimas.

2-ajame skyriuje pateikiamas didelės apimties ir didelio matavimo duomenų skaidymo algoritmas ir juo paremtos procedūros pritaikymas klasifikavimo procedūros skaičiavimų laiko sumažinimui. Kiti šios procedūros taikymai yra pateikiami sekančiame skyriuje, o jos aprašymas patogumo dėlei išskirtas į atskirą skyrių, taip pat priede pateikiamas algoritminis procedūros aprašymas. Šiuo algoritmu paremta viena iš procedūrų iš 1992–1995 metais MII sukurtos programinės įrangos daugiamačių Gauso mišinių klasifikavimui.

3-ajame skyriuje pateikiami 2-ajame skyriuje aprašyto didelės apimties ir didelio matavimo duomenų skaidymo algoritmo taikymai. 1-ajame skyrelyje pateikiama procedūra duomenų modelio verifikavimui. 2-ajame skyrelyje pateikiama procedūra neparametriniam didelio matavimo atsitiktinių vektorių nepriklausomumo testavimui. 3-ajame skyrelyje pateikiama procedūra naudojanti empirinį Bajeso metodą, kuri leidžia gauti efektyvesnius hipotezių tikrinimo kriterijus uždaviniams pateiktiems pirmuose dviejuose skyreliuose.

4-ajame skyriuje pateikiami empirinio Bajeso metodo taikymai retų dažnių populiacijose analizei. 1-ajame skyrelyje nagrinėjamas retų įvykių modeliavimas naudojant empirinį Bajeso metodą. 2-ajame skyrelyje nagrinėjami Puasono-Gauso ir Puasono-gama modeliai empiriniame Bajeso mažų tikimybių vertinime. 3-ajame skyrelyje pateikiamas modifikuotas regresinis empirinis Bajeso įvertis mažų tikimybių vertinimui.

5-ajame skyriuje pateikiami rezultatai ir išvados.

Pabaigoje pateikiamas literatūros sąrašas ir priedai.

Trumpa informacija apie autorių

Gimimo data ir vieta

1956 m. gruodžio 3 d., Utena.

Išsilavinimas

1974 Utenos IV vidurinė mokykla.

1979 Vilniaus Universitetas, Matematikos fakultetas, matematikos magistras.

Darbo patirtis

1979–1980 – inžinierius-matematikas programuotojas, Matematikos ir kibernetikos institutas, Tikimybių teorijos sektorius (nuo 1980 m. balandžio 14 d. – skyrius).

1980–1983 – aspirantas, Matematikos ir kibernetikos institutas, Tikimybių teorijos skyrius, Dinaminių sekų tyrimo sektorius.

1983–1989 – jaun. moksl. bendradarbis, Matematikos ir kibernetikos institutas, Tikimybių teorijos skyrius, Dinaminių sekų tyrimo sektorius (nuo 1988 – Taikomosios statistikos skyrius).

1989–2003 – mokslo darbuotojas, Matematikos ir kibernetikos institutas (nuo 1990 m. liepos 9 d. – Matematikos ir informatikos institutas), Taikomosios statistikos skyrius.

1995–2012 – vyr. specialistas (0,5 etato) Lietuvos Statistikos departamentas, Ekonometrinių tyrimų konsultantų skyrius (vėliau – Ekonometrinių tyrimų skyrius, Metodologijos ir kokybės skyrius).

2003–2009 – jaun. mokslo darbuotojas, Matematikos ir informatikos institutas), Taikomosios statistikos skyrius.

nuo 2009 – specialistas, Matematikos ir informatikos institutas (nuo 2010 m. spalio – Vilniaus universiteto Matematikos ir informatikos institutas), Sistemų analizės skyrius, Operacijų tyrimų sektorius.