Mathematical Problems of Computer Science 42, 63–72, 2014.

# Detection of Heterogeneity in Three-Dimensional Data Sequences: Algorithm and Applications

Evgueni A. Haroutunian, Irina A. Safaryan,
Aram R. Nazaryan and Narine S. Harutyunyan

Institute for Informatics and Automation Problems of NAS RA
e-mail: evhar@ipia.sci.am, irinasafaryan@yandex.ru, aram.nazaryan@gmail.com,
narineharutyunyan57@gmail.com

**Abstract**

We present a nonparametric algorithm which allows reducing the investigations of changes of the joint distribution of chronologically ordered multidimensional random sequence to the investigations of some one-dimensional conditional distributions. The algorithm is implemented with the statistical software package R. The action of the program is demonstrated on applications. The first one concerns the retrospective analysis of the changes in the concentration of chemical components of ground water preceding major seismic events. The second refers to the definition of cut-points in two-dimensional life time data sets of the imatinib-treated chronic myeloid leukemia patients.

**Keywords:** Change-point problem, Rank score test,Threshold copula, Cut-point selection method.

## 1. Introduction

Monitoring of some complex system leads to a vector of observations consisting of input variables (predictors), output variables (responses), and also concomitant (categorizing) variables, possibly influencing significantly on the joint distribution of the input and output variables. Each of these variables can be discrete, continuous or of non-numerical type.

Classification of observations to statistically homogeneous and significantly distinct groups is necessary for forecasting and taking adequate control actions. Such task arises in problems of medical and technical diagnostics in analyzing and forecasting catastrophic events in nature, and also in actuarial and financial mathematics.

The relatively small or moderate dependence is of interest in seismological applications. Nelsen [1] and Mari et al [2] have discussed the families of bivariate distributions with a small or moderate dependence. It is assumed that the data, obtained from different places of observation, are weakly dependent. This dependence is determined only by the fact that the variables belong to the same system, i.e., to the same geographic region and therefore, they are exposed to the same environmental conditions. Change of the structure of dependence is connected with the fact that the whole system is preparing to move from one state to another, and such a change is in some way a precursor to this transition.

In medical applications, such dependence is observable in the research of two-dimensional functions of survival.

In this paper we solve the problem of classification for the case of vectors $(X_n, Y_n, Z_n)$, $n = \overline{1, N}$, where $X$ and $Y$ are continuous, and the concomitant variable $Z$ can be arbitrary.

The case is of particular interest for practice where the shared variable $Z$ is a sequence of ordinal numbers of observations ranked chronologically, or according to some other concomitant variable. If the factor that influences the changes of dependence is the time, then the definition of the moment of changes is a multidimensional version of the famous problem about "disorder" (change-point detection problem). Staging, bibliography and state of the art are presented in the book of Borovkov [3]. Nonparametric methods of detection are presented in the book of Brodsky and Darkhovsky [4]. Theoretical substantiation of nonparametric algorithms based on rank statistics for detecting change-point in one-dimensional case was obtained by Safaryan [5].

For the two-dimensional case the following was stated. Let $(X_n, Y_n)$, $n = \overline{1, N}$ be a chronologically ordered two-dimensional random sequence, statistical properties of which change in some unknown moment (change-point). As in the one-dimensional case, we assume that there exists an index $\lambda \in [\Delta, 1 - \Delta], 0 < \Delta < 1/2$, which determines the index of observation $n_\lambda = [\lambda N]$ such, that the observation $(X_n, Y_n)$ has a two-dimensional distribution function $F^{(n)}(x, y)$ which can be written as:

$$F^{(n)}(x, y) = F_1(x, y)I\{n \le n_\lambda\} + F_2(x, y)I\{n > n_\lambda\},\ n = \overline{1, N}. \tag{1}$$

where $F_1(x, y) \neq F_2(x, y)$ and $I(A)$ is the indicator of the event $A$.

Since we are interested in the change of dependence, the same relation can be written with the copulas

$$C^{(n)}(u, v) = C_1(u, v)I\{n \le n_\lambda\} + C_2(u, v)I\{n > n_\lambda\},\ n = \overline{1, N}. \tag{2}$$

Recall that the copula of two random variables (RVs) $X$ and $Y$ with a joint distribution function $F(x, y)$ is a function $C(u, v)$, defined by the relation

$$C(F_X(x), F_Y(y)) = F(x, y),$$

or

$$C(u, v) = F(F^{-1}(u), G^{-1}(v)),$$

where $F_X(x)$ and $F_Y(Y)$ are marginal distribution functions and $F^{-1}$ and $G^{-1}$ are quasi inverse functions defined as $F^{-1}(u) = \inf\{x : F(x) > u\}$. If the marginal distributions are continuous, then this representation is unique [1]. Expediency of application of copulas in the discrete case is discussed in the article of Blagoveschensky [6], where the basic elements of the theory of copulas are also presented. The maximum likelihood estimator for the copula function change-point are obtained by Dias and Embrechts in [7]. In the article by Brodsky *et al* [8], an estimate of the change-point of 7-dimensional copula is obtained on the basis of multivariate modification of the Kolmogorov-Smirnov statistic. Unfortunately, this statistic is not very convenient for practical calculations, and also efficiency of a Kolmogorov-Smirnov statistic with respect to the statistic of rank score even in one-dimensional case is equal to zero.

In this paper a heuristic algorithm is proposed that allows to reduce the investigation of changes in the joint distribution of multivariate random sequence, ordered chronologically, or according to some other categorical variable, to the examination of changes in the corresponding one-dimensional sequance of conditional distributions with the application of appropriately selected rank scores statistics. The algorithm was first introduced in [9].

## 2. The Rescaling Procedure and Steps of the Algorithm

### 2.1 Copulas of Weak Dependence

Now we formalize the notion of weak dependence in terms of the copula. Let $X$ and $Y$ be RVs with continuous marginal distribution functions $F(x)$ and $G(y)$, the joint distribution function $F(x, y)$ and the corresponding copula $C(u, v)$. Each copula is a surface in the unit cube, so each distance between surfaces $z_0 = uv$ and $z_1 = C(u, v)$ should yield a measure of dependence between $X$ and $Y$. The most famous of them are the Pearson correlation coefficient $r$ and the Spearman rank correlation coefficient $\rho$:

$$r(X, Y) = \frac{1}{D(X)D(Y)} \int_0^1 \int_0^1 (C(u, v) - uv) dF^{-1}(u) dG^{-1}(v), \tag{3}$$

where $D$ stands for a standard deviation and

$$\rho(X, Y) = 12 \int_0^1 \int_0^1 (C(u, v) - uv) du dv. \tag{4}$$

The smaller is Spearman $\rho(X, Y)$, the weaker is the dependence [3]. Below some examples of weak dependence of copulas are given, which are relevant for the considered applications.

*Case 1.* Two families of one-parameter copulas describing the relatively weak dependence between the random variables X and Y are presented by Nelsen [1]. This is the Farlie-Gumbel-Morgenstern (FGM) copula:

$$C_\theta(u, v) = uv + \theta uv(1 - u)(1 - v), \quad \theta \in [-1, 1],$$

and the Ali-Michail-Haq copula (AMH):

$$C_\theta(u, v) = \frac{uv}{1 + \theta uv(1 - u)(1 - v)}, \quad \theta \in [-1, 1].$$

Coefficients (3) and (4) for copulas FMG and AMH change within the limits $[-1/3, 1/3]$.

*Case 2.* The second variant relates to the fact that the predictor can also be a grouping variable, i.e., contain one or more cut-points. In this case, the high correlation coefficient between the predictor and the response cannot be interpreted as a sign of dependence of one to the other. This dependence in the monograph of the Blagoveschensky [10] is named false or spurious and some examples of why such a spurious dependence may arise are given. In [11] the false dependence is defined as a threshold dependence. Here we remind the definition of homogeneity of an RV with respect to another, which is equivalent to the definition of independence.

We call an RV $Y$ homogeneous with respect to RV $X$ if for all pairs $(x, y)$ on the plane the following conditional probabilities are equal:

$$Pr(Y \le y / X \le x) = Pr(Y \le y / X > x). \tag{5}$$

If there exists a unique value of $x = \mu$ such that for all $y \in R$,

$$Pr(Y \le y / X \le x) = Pr(Y \le y / X \le \mu), \qquad for \quad x \le \mu, \tag{6}$$

$$Pr(Y \le y / X > x) = Pr(Y \le y / X > \mu), \qquad for \quad x > \mu, \tag{7}$$

and

$$Pr(Y \le y/X > x) \ne Pr(Y \le y/X > \mu), \tag{8}$$

then the statistical dependence between $X$ and $Y$ is called **one-threshold** and the value $\mu$ is called a **threshold**.

Actually one-threshold dependence is an example of dependence in single point $\mu$. In a similar way $M$- threshold dependence with $M > 1$ can be defined.

*Case 3.* Dependence arising in the time means that a dependence between $X$ and $Y$ occurs at some unknown moment of time, i. e., in relation (2) $C_1(u, v) = uv$. A practical example of dependence arising in the time given by Stakheev [12] is dedicated to earthquke geochemical precursors.

## 2.2 Problem Formulation for Some Heterogeneity Models

We solve the above stated problem about the determination of change moment of copula function without fixing in advance some kind of weak dependence, since if $X$ and $Y$ are weakly dependent like in Cases 1 or 3, then they verify (8). It is assumed that there is some unobserved random variable $Z^1$, which changes the copula function in some threshold point, then as it is shown in [13], $X$ is heterogeneous with respect to $Y$, and vice versa. Taking one of the variables for the base, we find the threshold value, which is change moment of the copula function. The approach to cut-point detection using the change-point identification techniques is introduced in [14].

## 2.3 Steps of Algorithm

**Step1. Selection of a base variable.**

Let $R_{X_i} = \#(X_n : X_n \le X_i,\ n = \overline{1, N})$ and $R_{Y_i} = \#(Y_n : Y_n \le Y_i,\ n = \overline{1, N})$, $i = \overline{1, N}$ be ranks of sequences $\{X_n\}_{n=1}^N$ and $\{Y_n\}_{n=1}^N$.

We define two sequences of **rank score statistic** as follows:

$$W_N^i(n) = n/(N-n)(T_J^i(n) - A(J)), \quad n = \overline{1, N}, \tag{9}$$

where

$$T_J^i(n) = 1/n \sum_{i=1}^{n} J(R_i/(N+1)), \tag{10}$$

and

$$A(J) = \int_0^1 (J(u)du), \tag{11}$$

where $R_i$ is $R_{X_i}$ in the first case or is $R_{Y_i}$ in the second case and $J(u)$ is a score function [15]. The **change points by time** are

$$\hat{n}_i = \arg \min_{[\Delta N] \le n \le [(1-\Delta)N]} W_N^i(n), \quad 0 < \Delta < 1/2, \quad i = 1, 2. \tag{12}$$

We define the standardized statistic

$$W_N^{*i} = \sqrt{N(1 - \frac{n}{N})\frac{n}{N}S(J)}W_N^i(n),$$

where $S(J) = \int_0^1 J^2(u)du - (\int_0^1 J(u)du)^2$. If for both variables $W_N^{*i} \le z_\alpha$, where $z_\alpha$ is the quantile of the level $\alpha$ for standard normal distribution, is detected and we choose the base

variable for which the extremum is observed earlier. For other cases, the choice of base variable is arbitrary.

**Step 2. Estimation of the moment homogeneity violation**

In what follows we denote the base variables by $X$, ranks of variable $Y$ **rearranged** in order of increasing of the base variable by $R^X_{Y_n}$, $n = \overline{1,N}$, and rank score statistic calculated with rearranged ranks by $W^X_N(n)$, $n = \overline{1,N}$. The moment $n_\mu$ of the violation of homogeneity of response variable relative to the base variable is defined by (12), where $W^i_N(n)$ is replaced by $W^X_N(n)$.

The founded moment of change is a certain rank of base variable, and corresponding to this rank the value of variable is the **cut-point**. Index of the moment in chronologically ordered sequence corresponding to cut-point is the moment of change of the copula. Unfortunately, in the real data quite a lot of such coinciding values are encountered. Different ranks correspond to them since in accordance with our ranking system, the smaller rank receives a chronologically earlier observation.

**Step 3. Correction of the moment of change of the copula visually using graphical representations**

For graundwaters it is presented on Fig. 2. by scatterplots. For survival time data the copula density histograms are obtained but not included in the paper for restriction of the volume of the text.

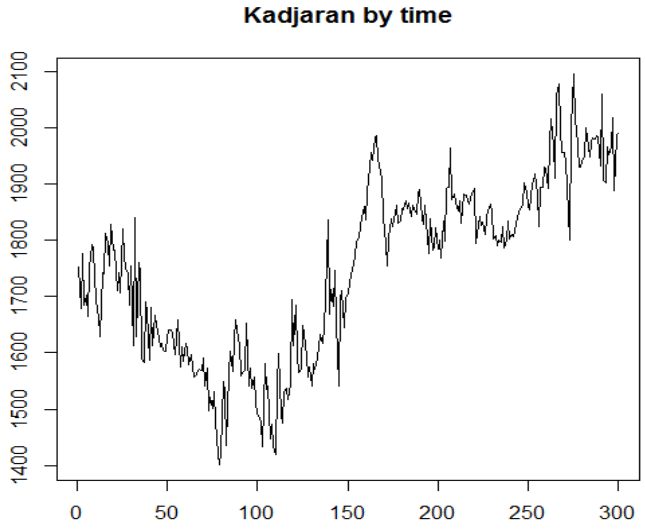
## 3. Analysis of Real Data

### 3.1 Comparison of Variations of the Components of Groundwater

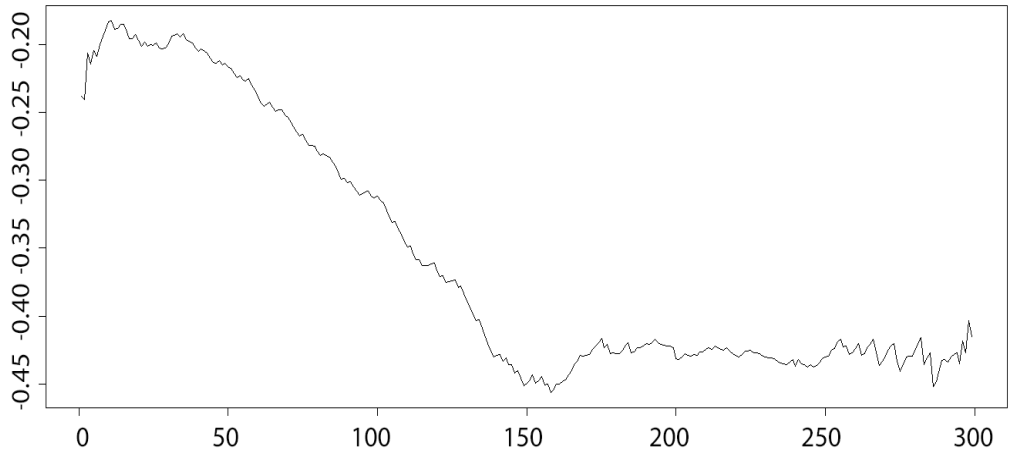
The concept of seismographic geochemical anomaly is not clearly defined. Usually, the sequence of observations between the two earthquakes contains two specific points: the start of the accumulation of changes and release to the level of quasi-permanent ("geochemical quiescency"). Thus, the observations in the period prior to the earthquake can be considered as a chronologically ordered random sequence with two unknown points of the "disorder". Processing of the data obtained on a number of stations of the hydrogeochemical observation network of the National Service for Seismic Protection (NSSP) of Armenia, using nonparametric algorithms, designed to changes in the statistical properties of the random sequences, confirmed this model of anomalies. The present example is related to the determination of the time of changing the structure of two-dimensional relationships for the three observation stations on the eve of the helium content of the Spitak earthquake. Note that the correlation analysis presented in the book of Petrosian [16] did not determine a significant statistical dependence on helium ($He$) between the observations of two stations.The actual data in the example is the content of helium in the groundwater on the eve of major seismic events. Definition of change moments of these two sequences was carried out in a known manner by means of the Wilcoxon statistic.

If we choose $\Delta = 0.1$, then the left 30 values of the Wilcoxon statistic will be cut to both sides and the minimum value outside the critical border -1.96 comes to the point 150. Thus, data from the monitoring station Ararat, having a rank 150, corresponds to the time of the change of homogeneity of helium in station Kajaran with respect to the station Ararat. Theoretically, this means some form of weak dependence (case 1-3) between the values of the indicators of these two stations. The rank 150 corresponds to the value of cut-point $\mu = 284$ and time 06.02.88 (2/6/88). We present a two-dimensional scatterplot until the

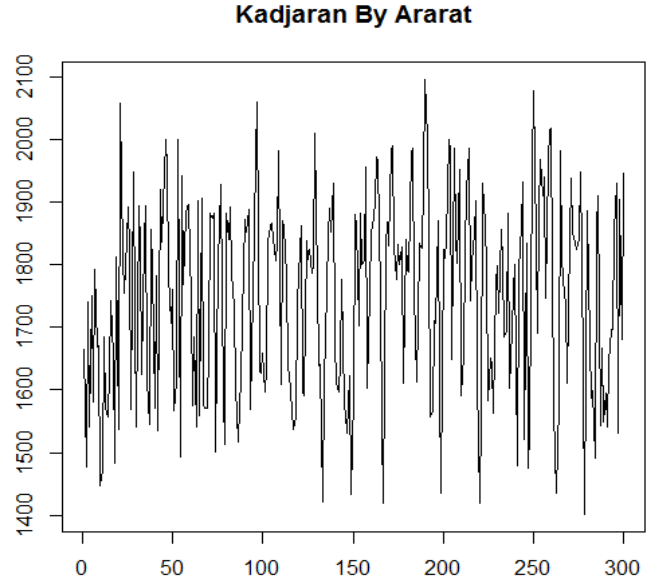
time moment 06.02.88 by one color, and after it with another.



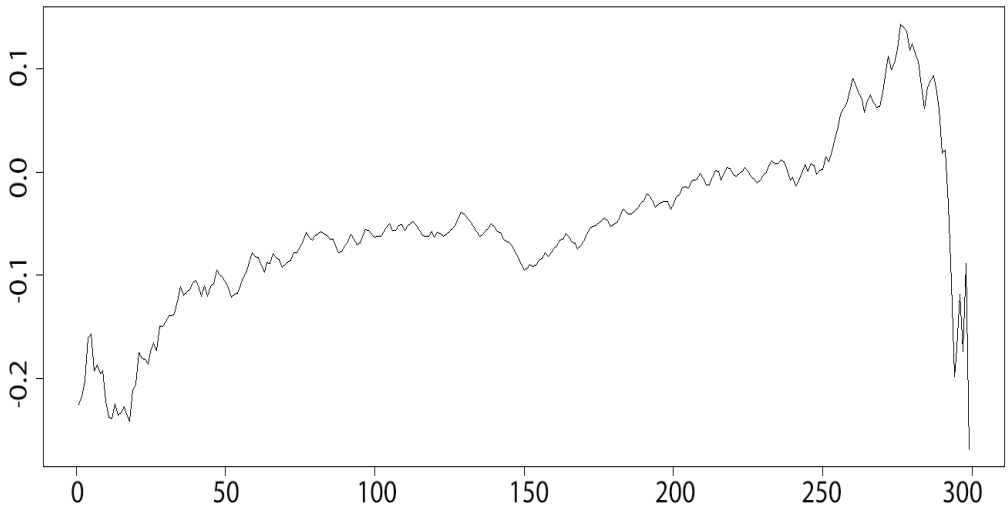

(a)




(b)




(c)




(d)

Fig. 1. (a)data of helium at the station Kajaran are on the vertical axis, the horizontal axis is for the number of observations over time. (b) Wilcoxon statistic over time, (c) data for station Kajaran ordered according to increasing of values of the index of helium on Ararat station are on the vertical axis, the horizontal axis is for Rank values of the helium in station Ararat. (d) Wilcoxon statistic for Kajaran by Ararat.

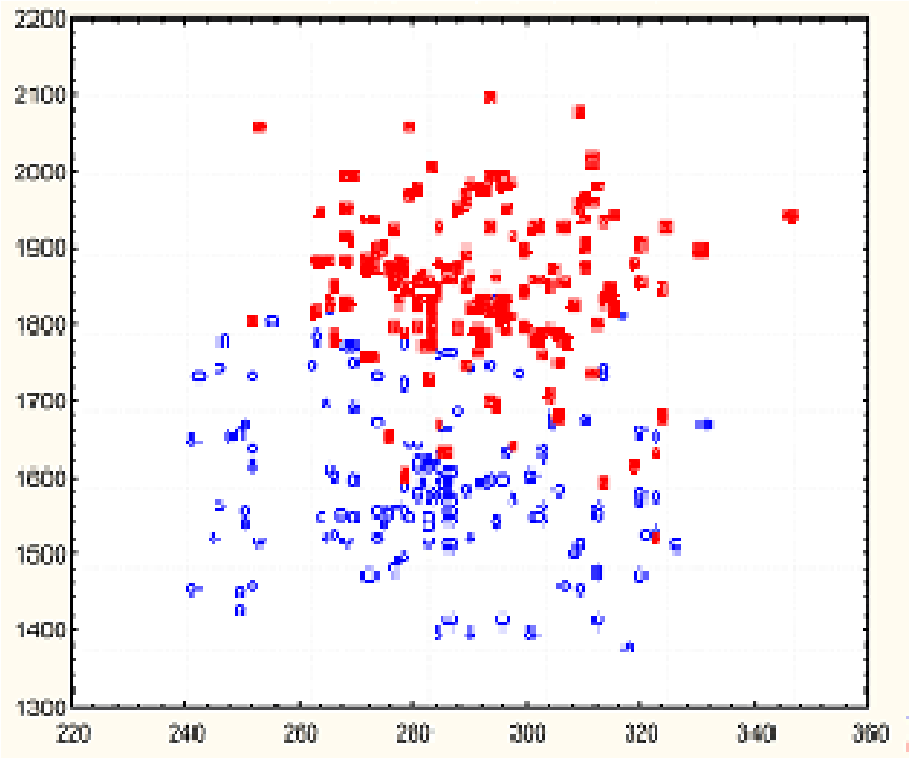



Fig. 2. Two-dimentional data of helium classified over time (Kajaran -Ararat).

We see that the two-dimensional observations are sufficiently well separated in time. However, the following ranks: 144(29.09.87), 145 (29.11.87), 146 (01.12.87), 147(11.12.87), 149(07.01.88) 150(06.02.88), 151(28.04.88), 152(30.04.88) correspond to the value 284. We constructed two-dimensional scatter plots for all of moments corresponding to the cut-point $\mu = 284$ and made sure that the best division by time corresponded to 02.06.88.

### 3.2 Analysis of Survival Function of Patients with Myeloid Leukemia

Recently a lot of works have been devoted to the study of two-dimensional survival functions which include also applications describing algorithms and programs in R codes such as in [17,18]. The same algorithm was applied to the analysis of two-dimensional life time data of expectancy of patients with myeloid leukemia. The sequence $\{X_n\}_{n=1}^N$ is the life time data denoting the date of diagnosis, and the sequence $\{Y_n\}_{n=1}^N$ denotes the life time from the point of therapeutic treatment. Categorizing variable $Z$ is the age of patient, the sample size is 413 patients. In this example, the variables are strongly dependent, however,even in this case, the proposed algorithm works. The cut-point is found: 60 years of age, after which the copula function is changed. For two-dimensional histogram of the density copula such separation is visible even to the eye. The main conclusion is that after 60 years, the life time is only slightly dependent on the therapeutic treatment.

## 4. Conclusion

The presented algorithm shows the prospects of application of threshold copula methods and mixed sampling to determination of anomalies in multidimensional hydrogeochemical data occurring prior to earthquakes and for spatially correlated survival data. Further theoretical elaboration and implementation of programs for their realization are admissable. It is desirable to fitting copulas $C_1(u, v)$ and $C_2(u, v)$ for using the goodness-of-fit tests.

## Acknowledgement

This work was supported in part by SCS of MES of RA under Thematic Program No SCS 13–1A295.

## References

[1] R. V. Nelsen, *An Introduction to Copulas*, Springer, New York, 2006.

[2] D. D. Mari, S. Kotz, *Correlation and Dependence*, London Imperial College Press. 2004.

[3] A. A. Borovkov, *Mathematical Statistics*, (in Russian), M., "Nauka", 2007.

[4] B. E. Brodsky, B.S. Darkhovsky, *Nonparametric Methods in Change-Point Problems.* Kluver, Dordrecht, 1993.

[5] I. A. Safaryan, *Nonparametric algorithms for monitoring of time series,* PhD Thesis (in Russian), Yerevan, 1998.

[6] U. N. Blagoveschensky, "The main elements of the theory of copulas", (in Russian), *Applied Econometrics, N2*, pp. 113-131, 2012.

[7] A. Dias and P. Embrechts, "Change-point analysis in finance and insurance", *In New Risk Mesures in Investment and Regulation, Willey, New York*,pp. 2003.

[8] B. E. Brodsky, G.I. Penikas and I.A. Safaryan, "Detecting structural changes in the copula models", (in Russian), *Applied Econometrics, 4(16)*, pp. 3-16, 2009.

[9] E. A. Haroutunian, I.A. Safaryan, H.M. Petrosyan and A. R. Gevorkian, "On identification of anomalies in multidimensional hydrogeochemical data as earthquake precursors", *Mathematical Problems of Computer Science,* vol. 40, p. 76, 2013.

[10] U. N. Blagoveschensky, *Secrets of the correlations,* (in Russian), M ., "Nauchnaya kniga", 2008.

[11] E. A. Haroutunian and I. A. Safaryan, "Copulas of two-dimensional threshold models", *Mathematical Problems of Computer Science*, vol. 31, pp. 40-48, 2008.

[12] U.I. Staheev, "Geochemical precursors of earthquakes", (in Russian), *Russian Chemical Journal,* vol. XLIX, no. 4, pp. 110-119, 2005.

[13] E. A. Haroutunian and I. A. Safaryan, "On estimation of threshold parameter in three-dimensional copulas model", *Abstracts of International Conference on Computer Science and Information Technologies, Yerevan*, pp. 129-131. 2011.

[14] B. Lausen and M. Schumacher, "Maximally selected rank statistics". *Biometrics,* vol. 48, pp. 73-85, 1992.

[15] E. A. Haroutunian and I.A. Safaryan, "Nonparametric consistent estimation of random sequence change moment", (in Russian), *Mathematical Problems of Computer Science*, vol. 17, pp. 76-85, 1997.

[16] H. M. Petrosyan, *Precursors and Prognosis of Earthquakes on the Territory of the Republic of Armenia*, (in Russian), Yerevan, 2009

[17] J. Paik and Zh. Ying, "A composite likelihood approach for spatially correlated survival data", *Computational Statistics and Data Analysis*, vol. 56, pp. 209-216, 2012.

[18] P. F. Su and Ch.-I. Li, Yu. Shyr, "Sample size determination for paired right-censored data based on the difference of Kaplan-Meier estimates", *Computational Statistics and Data Analysis*, vol. 74, pp. 39-51, 2014.

**Submitted 22.08.2014, accepted 10.11.2014.**

# Եռաչափ տվյալների հաջորդականությունների անհամասեռության հայտնաբերում. ալգորիթմ և կիրառություններ

Ե. Հարությունյան, Ի. Սաֆարյան, Ա. Նազարյան և Ն. Հարությունյան

## Ամփոփում

Ներկայացվում է ոչ պարամետրական ալգորիթմ, որը թույլ է տալիս բազմաչափ պատահական հաջորդականության համատեղ բաշխման հետազոտումը հանգեցնել որոշ միաչափ պայմանական բաշխումների հետազոտմանը: Ալգորիթմը իրականացված է ծրագրման վիճակագրական R լեզվի միջոցով: Ծրագրի գործելիությունը ցույց է տրվել երկու կիրառությունների վրա՝ սեյսմիկ և բժշկական:

# Обнаружение неоднородности в трехразмерных последовательностях данных: алгоритм и приложения

Е. Арутюнян, И. Сафарян, А. Назарян и Н. Арутюнян

## Аннотация

Мы представляем непараметрический алгоритм, позволяющий сводить исследование изменений совместного распределения многомерной случайной последовательности к исследованию некоторых одномерных условных распределений. Алгоритм реализован в среде статистического языка R. Действие программы показано на двух приложениях: сейсмологическом и медицинском.