Using Rank Tests and Threshold Copulas for Classification of Multidimensional Data Sets

Evgueni A. Haroutunian, Irina A. Safaryan and Narine S. Harutyunyan

Institute for Informatics and Automation Problems of
NAS of RA
Yerevan, Armenia

e-mail: eghishe@sci.am,
irinasafaryan@yandex.ru,
narineharutyunyan57@gmail.com

ABSTRACT

This report deals with classification of multidimensional data sets into statistically homogeneous groups and emphasizes the copula representation of the dependence between the components of a random vector. We present a nonparametric algorithm based on rank score test which allows reducing the investigation of changes in the joint distribution of a random vector components to investigation of some one-dimentional conditional distributions. Some applied examples are presented.

Keywords

Change-point problem, change-point detection techniques, threshold dependence, copula, rank score test, categorical variable

1. INTRODUCTION

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. Classification is implemented with respect to some concominant (categorizing) variable which may be discrete, continuous or of non-numerical type.

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 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). Nonparametric methods of detection are presented in the book of Brodsky and Darkhovsky [1]. Theoretical substantiation of nonparametric algorithms based on rank score statistics for detecting change-point in one-dimensional case was obtained by Safaryan [2].

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 number 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 [3].

In the article by Brodsky *et al* [4], 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. Statistics based on rank scores use a priori information on difference between distributions prior and after disorder moment and consequently they are more effective.

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 sequence of conditional distributions with the application of appropriately selected rank scores statistics. The algorithm is based on rank test statistics and is applied to analysis of real data in [5] and [6].

2. STATEMENT OF PROBLEM

Let $\{(X_n, Y_n)\}_{n=1}^{N}$ be a chronologically ordered two-dimensional random sequence. We consider $\{(X_n, Y_n)\}_{n=1}^{N}$, as a random sample of random vector (X, Y) with common distribution function $F(x, y)$ and continuous marginals $F_X(x)$ and $F_Y(y)$.

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

We denote by $C^{(n)}(u, v)$ copula of (X_n, Y_n) and by $n_\lambda = [\lambda N]$ for $\Delta < \lambda \leq 1 - \Delta$, $0 < \Delta < \frac{1}{2}$ - change point. Our aim is to test hypotheses for each $n = \overline{1, N}$

$$H_0 : C^{(n)}(u, v) = C(u, v)$$

under

$$H_1 : C^{(n)}(u, v) = I\{n \leq n_\lambda\}C_1(u, v) + I\{n > n_\lambda\}C_2(u, v)$$

where $I\{A\}$ is the indicator of the event A and

$$C_1(u, v) \neq C_2(u, v).$$

3. CHANGE MOMENT DETECTION UNDER SOME ASSUMPTIONS

In practical tasks it is often necessary not only detect the existence of change in dependencies between RVs X and Y but also specify in some sense the extent of such change. a) The copulas $C_1(u, v)$ and $C_2(u, v)$ belong to the same family of one-parametric copulas and differ only in the parameter values, such as the Farlie-Gumbel-Morgenstern families presented by Nelsen [3].

$$C(u, v, \theta) = uv + \theta uv(1 - u)(1 - v), \quad \theta \in [-1, 1], \quad (3)$$

b) The copulas $C_1(u, v)$ and $C_2(u, v)$ belong to different one-parametric families and differ both in the functional type and in parameter values. $C_1(u, v)$ is Farlie-Gumbel-Morgenstern and $C_2(u, v)$ is

$$C(u, v, \theta) = \frac{uv}{1 + \theta uv(1 - u)(1 - v)}, \quad \theta \in [-1, 1]. \quad (4)$$

For cases a) and b) we suggested an algorithm that allows reducing the detection of changes in copula function occurring in unknown change-point n_λ to testing homogeneity of one RV, for instance Y, with respect to another RV X [7]. In [7] it is proved that homogeneity of RV Y in relation to RV X is equal to independence of RV's X and Y. Violation of homogeneity at least in one point is defined as a threshold or very weak dependence.

4. SOME NUMERICAL EXAMPLES

In the report, model examples are given for the case a) when the suggested method cannot be applied under some value of parameter λ. In the case b) the method can be applied provided the copula $C_1(u, v)$ corresponds to a relatively weak dependence, while the copula $C_2(u, v)$ expresses a stronger dependence (the Frank copula, for example). Calculations are made with programming system R.
On Fig. 1. the modelled copula (3) of weak dependence is presented. We see the absence of points an the main diagonal.

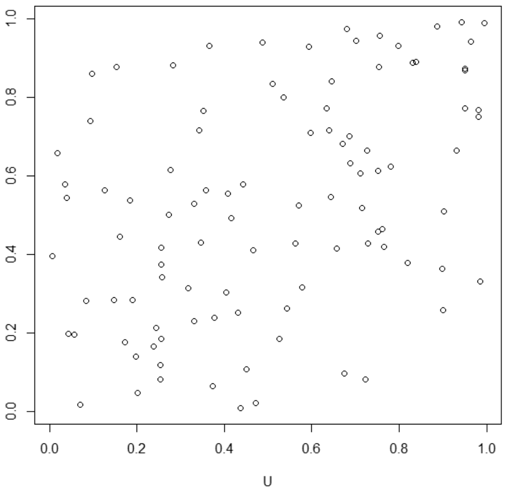

Figure 1.

5. CONCLUSION

To summarize we can emphasize that the choice of an adequate model for real problem presents some difficulty. As it is mentioned by Blagoveschensky: "Using copulas in different statistical problems only starts" [8]. The relatively small or moderate dependence is of interest in seismological applications.

6. ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

[4] B. E. Brodsky, G. I. Penikas and I. A. Safaryan, "Detecting structural changes in the copula models", (in Russian), *Applied Econometrics,* , vol. 4, no. 16, pp. 3-16, 2009.

[5] 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, pp. 76-84, 2013.

[6] E. A. Haroutunian, I. A. Safaryan, A. Nazaryan and N. Harutyunyan, "Detection of heterogeneity on three-dimensional data sequences: algorithm and applications", *Mathematical Problems of Computer Science,* vol. 42, pp. 63-72, 2014.

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

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