An Application of Changepoint Detection Methods to Independence Testing

Robert Chitchian
Armimpexbank, Automation and Banking Technologies Department
V. Sargssyan 2
Yerevan, Armenia
chirona@mail.ru

Irina Safaryan
Institute for Informatics and Automation Problems of the
National Academy of Sciences of Armenia and of the Yerevan State University
P. Sevak 1
Yerevan, Armenia
evhar@ipia.sci.am

The problem of nonparametric testing the hypothesis of independence between two random variables (r.v.) X and Y under alternative of unusual kind dependence is investigated. Let the common bivariate distribution function (d.f.) of vector (X, Y) is $H(x, y)$ having continuous marginal df's $F(x)$ and $G(y)$. Denote $F_1(x|y) = Pr\{X \le x|By\}$ and $F_2(x|y) = Pr\{X \le x|\overline{By}\}$, where $By = \{Y \in My\}$ and $My = (-\infty, y]$. We define the null hypothesis of independence and alternative hypothesis as follows

$$H_0 : F_1(x|y) = F_2(x|y) \;\; for \quad all \;\; y \in \mathcal{R}$$

$$H_1 : \exists! \mu \in \mathcal{R} \;\; such \;\; that \;\; F_1(x|y \le \mu) = F_1(x|\mu) \ne F_2(x|\mu) = F_2(x|y > \mu)$$

The df's $F(x), G(y), F_1(x|y)$ and $F_2(x|y)$ as well as the μ quantile level defined by relationship $q = G(\mu)$ is supposed unknown. It is known only that for $0 < \Delta < 1/2$ sufficiently small the inequality $\Delta < q < 1 - \Delta$ holds. Our aim is testing under these conditions H_0 hypothesis versus alternative H_1 and in case of its rejection to obtain a consistent and unbiased estimate of unknown quantile μ.

The problem is typical for medical, ecological and technical applications. In these application it is frequently required to test homogenity of the distribution of r.v. X for arbitrary value r.v.Y or to establish the existence of some theshold value of r.v.Y after achieving of which d.f. of X is changed. In this connection the distribution of survival time in respect to the age of patient is studied in [1].

It is easy to receive that under H_0 the d.f. is of the form $H(x, y) = F(x)G(y)$ and under H_1 this d.f. may be expressed as

$$H(x, y) = \begin{cases} F(x)G(y) + (F_1(x|\mu) - F_2(x|\mu))G(y)(1 - G(\mu)), & for \;\; y \le \mu \\ F(x)G(y) + (F_1(x|\mu) - F_2(x|\mu))G(\mu)(1 - G(y)), & for \;\; y > \mu \end{cases}$$

Thus, the linear rank statistics most commonly used to test the independence [2] can be applied in this case, but they do not allow to estimate value of μ.

For this purpose the changepoint detection approach is applied in this work. Let the sequence $\{X'_n\}_{n=1}^N$ is a permutation of observations $X_1, \cdots, X_N$ placed in the order corresponding the order of increase for value of $Y_1, \cdots, Y_N$. The following theorem is prove.

Theorem. Under H_0 members of sequence $\{X'_n\}_{n=1}^N$ are independent and distributed according $F(x)$. Under H_1 for $[\Delta N] \le n \le [(1-\Delta)N]$ and $N \to \infty$ r.v. X'_n are asymptotically distributed as sequence of independent r.v. Z_n which distribution function is $F_1(x|\mu)$ for $n \le \tilde{n} = [qN]$ and $F_2(x|\mu)$ for $n \ge \tilde{n}+1$.

Number $\tilde{n}$ is called a changepoint for the sequence $\{Z_n\}_{n=1}^N$ and according [3] its consistent estimate is the number $\hat{n} = [\hat{t}N]$, where

$$\hat{t} = arg \min_{\Delta < t < 1-\Delta} W_N(t)$$

and $W_N(t)$ is the sequence of two-sample rank statistics rejected hypothesis of homogenity the sequence $\{X'_n\}_{n=1}^N$.

Then we shall reject the independence hypothesis if in the sequence $\{X'_n\}_{n=1}^N$ a changepoint exists. From above mentioned theorem we obtain that consistent and unbiased estimate for unknown parameter μ is the order statistics with the number $\hat{n}$ of the sample $(Y_1, \cdots, Y_N)$ from d.f. $G(y)$, namely $\hat{\mu} = Y_{(\hat{n})}$.

The suggested method of independence detection is compared on model examples with traditional ones based on linear rank statistics of independence, such as correlation coefficient of Spearman.

A new approach to independence testing combining both mentioned methods is proposed.

REFERENCES

[1] Contal C., O'Quigly I. (1999). An application of changepoint methods in studying the effect of age on survival in breast cancer. *Computational Statistics And Data Analysis.* **30**. 253-270.

[2] Hajek J., Sidak Z. (1967). Theory of rank tests. *Academic Press. New York.*

[3] Haroutunian E. A., Safaryan I. A. (1997) Nonparametric consistent estimation of the change moment of random sequence properties. *Mathematical Problems of Computer Science. Yerevan.* **17**. 76-85. (in Russian)

RESUME

Changepoint detection methods to independence testing under alternative of unusual kind dependence is suggested.