Mathematical Problems of Computer Science 23, 2004, 67–79.

Two-dimensional Sequence Homogeneity Testing Against Mixture Alternative*

Irina A. Safaryan, Evgueni A. Haroutunian and Arsen V. Manasyan

Institue for Informatics and Automation Problems of NAS of RA
e-mail evhar@ipia.sci.am

Abstract

The behavior of linear rank statistics is investigated on models in which various subsequences of observations follow different statistical distributions. Such data can be interpreted both as models of a finite number distribution mixtures and as dependence models. We apply data set simulation to obtain estimates of average and variance of used rank statistics. The modeled and asymptotic results are enough close.

1 Introduction

The behavior of a two-dimensional random sequence $(X_n, Y_n), n = 1, 2, ...$ is analyzed. One component, let it be $Y_n, n = 1, 2, ...$, is the source of nonhomogeneity of the other. We suppose that there exists no more than one value of the component Y, called cutpoint, that makes possible classification of observations into two groups with different statistical distributions. In this model the distribution of the dependent variable X is expressed in the form of two distribution mixture with mixture weight coefficient determined by some apriori unknown quantile of independent variable.

To test the homogeneity variable versus mixture alternative we use the approach proposed by B. Lausen and M Schumacher [1] which is based on detection of change point for induced order statistics of response variable. In this work we present a computational verification of the results obtained by E. Haroutunian and I. Safaryan in [2] and [3]. It is shown that the asymptotic average and variance of the mixture weight coefficient are very close to the Monte-Carlo estimates. Moreover, the comparison of different rank tests on the same model demonstrates that the variance of the estimate can be minimized by corresponding choice of the test score function.

2 Rank Test Statistics for Data Homogeneity Testing

Let $\{(X_n, Y_n)\}_{n=1}^N$ be a random sample from two-dimensional random variable (X, Y) with a common distribution function $Q(x, y)$ and continuous marginal distribution functions $F(x)$ and $G(y)$ which are unknown. It is required to test the existence of some threshold value

*The work was partially supported by INTAS, project 00-738.

(cutpoint), of a predictor variable Y such that in pairs (X_n, Y_n) the distribution of each X_n is the same for $Y_n \le \mu$ and other for all $Y_n > \mu$.

We denote conditional probabilities of the event $\{X \le x\}$ under condition $\{Y \le y\}$ by

$$F_{1|y}(x) = \Pr(X \le x | Y \le y)$$

and under condition $\{Y > y\}$ by

$$F_{2|y}(x) = \Pr(X \le x | Y > y).$$

Then the classification problem can be formulated as testing of null hypothesis

$$H_0 : F_{1|y}(x) = F_{2|y}(x), \text{for all } y,$$

versus the alternative hypothesis

$$H_1 : F_{1|y}(x) = F_{1|\mu}(x), \text{for } y \le \mu,$$

$$F_{2|y}(x) = F_{2|\mu}(x), \text{for } y > \mu,$$

$$F_{1|\mu}(x) \ne F_{2|\mu}(x).$$

As it is shown in [2] and [3] the null hypothesis H_0 means the homogeneity of the marginal distribution function $F(x)$ of random variable X as well as independence of X and Y. The alternative hypothesis H_1 means that the marginal distribution function $F(x)$ can be presented as a mixture of two distributions

$$F(x) = \nu F_{1|\mu}(x) + (1 - \nu) F_{2|\mu}(x), \tag{1}$$

where the weight coefficient of the mixture $\nu = G(\mu)$ is the level of an apriori unknown quantile μ. Moreover, under H_1 X and Y are dependent with a common distribution function of the form:

$$Q(x,y) = \begin{cases} F(x)G(y) + (F_{1|\mu}(x) - F_{2|\mu}(x))G(y)(1 - G(\mu)), & \text{for } y \le \mu, \\ F(x)G(y) + (F_{1|\mu}(x) - F_{2|\mu}(x))G(\mu)(1 - G(y)), & \text{for } y > \mu \end{cases} \tag{2}$$

This circumstance allows to suggest a common nonparametric approach to the distributions mixture identification and independence hypothesis testing.

Definition 1: *If the marginal distribution function of predictor is continuous and* $Y_{(1)} < Y_{(2)} \ldots < Y_{(N)}$ *are its order statistics then we define the induced order statistics sequence* $\{X_{nY}\}_{n=1}^{N}$ *as* $X_{nY} = X_i$ *if* $Y_{(n)} = Y_i$, $n = \overline{1,N}$.

It is shown [2] that under H_0 each of induced order statistics X_{nY} is distributed asymptotically by $N \to \infty$ according $F(x)$ and under H_1 has the distribution function $F_{1|\mu}(x)$ if $n \le n(\nu)$ and the distribution function $F_{2|\mu}(x)$ if $n > n(\nu)$. Then this problem may be seen as a generalization of change point problem.

Definition 2: *We will call the number* $n(\nu) = [\nu N]$ *a changepoint of the induced order statistics of sequence* $\{X_{nY}\}_{n=1}^{N}$.

Definition 3: *Ω_J is the set of pair distributions $F_1(x)$ and $F_2(x)$ distinguishable in the sense of some score function $J(u)$, if for each $\lambda \in (0,1)$ $\Omega_J = \Omega_J^- \cup \Omega_J^+$ and sets are defined as follows*

$$\Omega_J^- = \left\{ (F_1, F_2) : \int_{-\infty}^{+\infty} J(\lambda F_1(x) + (1-\lambda)F_2(x)) dF_1(x) < \int_0^1 J(u)du \right\},$$

$$\Omega_J^+ = \left\{ (F_1, F_2) : \int_{-\infty}^{+\infty} J(\lambda F_1(x) + (1-\lambda)F_2(x)) dF_1(x) > \int_0^1 J(u)du \right\}.$$

We define the rank of induced order statistics

$$R_{Nj} = \#\{X_n : X_n \le X_{Nj}, \ n = \overline{1,N}\},$$

the linear rank statistics

$$T_{J,N}(t_n) = \frac{1}{n} \sum_{j=1}^{n} J_N(R_{Nj}/N), \tag{3}$$

where $\lim\limits_{N\to\infty} J_N(u) = J(u)$

$$A(J) = \int_0^1 J(u)du$$

and the function $W_N(t)$ which is constant under $t \in [(n-1)/N, n/N]$ and takes values in points $t_n = n/N$ as follows

$$W_N(t_n) = \frac{1}{1-t_n} \left(T_{J,N}(t_n) - A(J) \right). \tag{4}$$

Then the results obtained in [4] imply the following statements.

1. If $(F_{1|\mu}(x), F_{2|\mu}(x)) \in \Omega_J^-$ then the test with critical region

$$W_N(t) \le C_\alpha^-(t), \quad t \in [\Delta, 1-\Delta], \tag{5}$$

is consistent for H_0 versus H_1,

2. If H_1 is true then the estimate of quantile level ν defined by relationship

$$\hat{\nu} = arg \min_{\Delta \le t \le 1-\Delta} W_N(t) \tag{6}$$

is consistent. Similar statements are valid if $(F_{1|\mu}(x), (F_{1|\mu}(x)) \in \Omega_J^+$.

Let us denote

$$\lambda_1(t,\nu) = \begin{cases} 1, & t \le \nu, \\ \nu/t, & t > \nu, \end{cases}$$

$$\lambda_2(t,\nu) = \begin{cases} (1-\nu)/(1-t), & t \le \nu, \\ 1, & t > \nu, \end{cases}$$

and consider the following distribution functions depending on two parameters t and ν

$$F_1(x,t,\nu) = \lambda_1(t,\nu)F_{1|\mu}(x) + (1-\lambda_1(t,\nu))F_{2|\mu}(x),$$

$$F_2(x,t,\nu) = \lambda_2(t,\nu)F_{2|\mu}(x) + (1-\lambda_1(t,\nu))F_{1|\mu}(x).$$

Then we introduce the following functionals:

$$A_J(t,\nu) = \int_{-\infty}^{\infty} J(F(x))dF_1(x,t,\nu),$$

$$V(F_i,F) = F_i(x,t,\nu)(1-F_i(y,t,\nu))J'(F(x))J'(F(y)), \quad i=1,2,$$

and by analogy with [4] we can obtain that

$$p\lim_{N\to\infty} T_{J,N}(t_n) = A_J(t,\nu)$$

and $T_{J,N}(t_n)$ has asymptotically for $N\to\infty$ normal distribution with expectation

$$\mathrm{E}T_{J,N}(t_n) = A_J(t,\nu)$$

and variance

$$\mathrm{D}T_{J,N}(t_n) = \frac{S_J(t,\nu)}{N} = \frac{2(1-t)}{N}\left\{ \iint\limits_{-\infty<x<y<\infty} V(F_2,F)dF_1(x,t,\nu)dF_1(y,t,\nu) + \right.$$

$$\left. + \frac{1-t}{t} \iint\limits_{-\infty<x<y<\infty} V(F_1,F)dF_2(x,t,\nu)dF_2(y,t,\nu) \right\}.$$

It is easy to obtain that under H_0 : $S_J(t,\nu) = S(J) = \int_0^1 J^2(u)du - \left(\int_0^1 J(u)du\right)^2$.

Thus the $C_\alpha^-(t)$ in (5) is defined as follows:

$$C_\alpha^-(t) = \sqrt{S(J)/Nt(1-t)}\; z_\alpha,$$

where z_α is the quantile of the level α for standard normal distribution. In practical computations instead of $W_N(t)$ standardized statistics are usually used

$$W_N^*(t) = \sqrt{N(1-t)t/S(J)}\; W_N(t) \tag{7}$$

with left side critical region defined by the inequality $W_N^*(t) \le z_\alpha$.
Using this result the following theorem was proved in [2].

Theorem: *If $\nu \in [\Delta, 1-\Delta]$, then the estimate of mixture weight coefficient ν is distributed asymptotically normally with expectation*

$$\mathrm{E}\hat{\nu} = \nu \tag{8}$$

and variance

$$\mathrm{D}\hat{\nu} = \frac{(1-\nu)^2 S_J(\nu,\nu)}{N(A_J(\nu,\nu) - A(J))^2}. \tag{9}$$

Remark: *The critical region for small samples was obtained by T. Hothorn and B. Lausen in [5].*

3 Computational Study and Comparision of the Estimates Properties

We present the results of a series of rank-score tests and estimates in order to compare their asymptotic properties. Following examples were simulated.

Example 1: Shift in mean. The dependent variable was drawn from a standard normal distribution with shifts after changepoint in mean $a = 0.5$ and $a = 2.5$.

Example 2: Shift in scale. The dependent variable was drawn from a standard normal distribution with a shift in scale after changepoint $\sigma = \sqrt{2}$ The cutpoint for dependent variable Y is defined from the relationship $\mu = G^{-1}(\nu)$, where $G(y)$ is the distribution function for Y.

In both examples the distribution functions of the independent variable Y were the following

a) $G(y)$ – uniform distribution,

b) $G(y)$ – standard normal distribution.

Three levels are choosen for the quantiles of independent variable, namely $\nu = 0.25$, $\nu = 0.5$, $\nu = 0.75$. Then, under sample size $N = 300$ changepoints $n(\nu) = [\nu N]$ for simulated sequences of induced order statistics are the following: $n(\nu) = 75$, $n(\nu) = 150$ and $n(\nu) = 225$. For the changepoint detection and estimation we use the follwoing score functions:

$J(u) = u$ (Wilcoxon-score statistics),

$J(u) = u^2$ (Square- score statistics),

$J(u) = (u - 0.5)^2$ (Mood-score statistics).

Hence according to Definition 3 for Example 1 we have that $F_{1|\mu}(x)$ and $F_{2|\mu}(x)$ are distinguishible for statistics with score functions $J(u) = u$ and $J(u) = u^2$, and undistingushible for Mood statistics. For the Example 2 the situation is inverse to the one in Example 1.

As an example in Fig.1, we show that bivariate histograms of simulated realizations are two-vertex. This makes us conclude that the random sample $\{(X_n, Y_n)\}_{n=1}^N$ is drawn from a mixed two-dimensional distribution. This example is built for the case when shift in mean after changepoint is equal to 2.5. In this case, as it is shown on Fig. 2 (a), the change-point can be detected even visually on the linear plot of simulated realization. In Fig. 2(b) we see that the estimate obtained with standardized rank test with Wilcoxon-scores is very close to the real values of the change-point. However, when the shift in mean equals to 0.5 it is impossible to detect the changepoint visually. The behavior of the standardized rank test with Wilcoxon-scores for the shift equal to $a = 2.5$ and $a = 0.5$ is shown on Fig.3 ((b) and (d)).

A Monte-Carlo study for the case $a = 0.5$ was performed for $K = 100$ independent realizations. Let $W^*_{N,k}(n/N)$ be a standartized statistic, defined in (7) and calculated for k-th realization. Then $\hat{n}_k$ defined as

$$\hat{n}_k = arg \min_{[\Delta N] \le n \le [(1-\Delta)N]} W^*_{N,k}(n/N), k = \overline{1, K},$$

is the changepoint estimate for k-th realization. The estimate of the mixture weigth coefficient is the followig

$$\hat{\nu}_k = \hat{n}_k / N$$

and estimates of average and variance for $\hat{\nu}_k$ are

$$\overline{\nu} = \frac{1}{K} \sum_{k}^{K} \hat{\nu}_k, \quad S^2_\nu = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\nu}_k - \overline{\nu})^2.$$

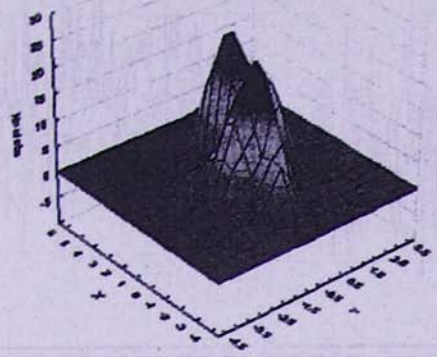
The results of simulation are presented in table 1.

ν	Average	Minimum	Maximum	Variance	Std.Dev.	Confidence(-95%)
0.25	0.251967	0.013333	0.686667	0.008487	0.092123	0.233687
0.5	0.495300	0.066667	0.723333	0.008001	0.089448	0.477552
0.75	0.715400	0.016667	0.913333	0.024067	0.155136	0.684618

Table 1. Monte-Carlo estimates for the average and variance of mixture weight coefficient.

Example 1: Shift in mean

(a)



(b)

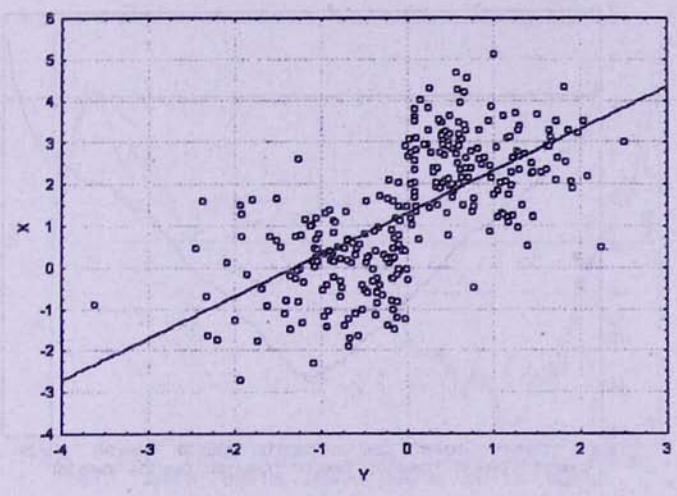
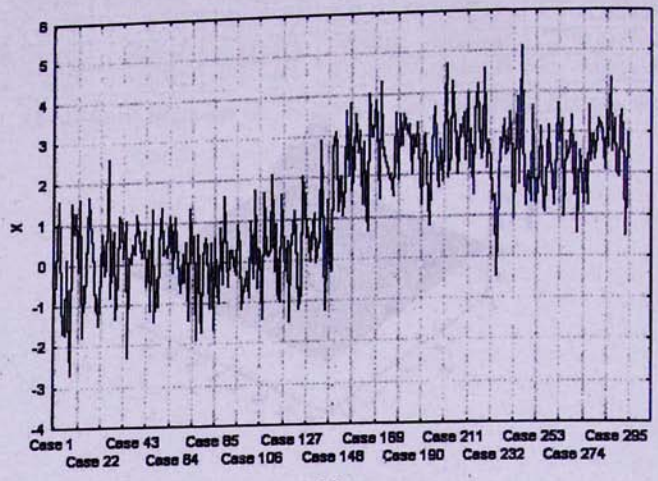


Fig 1. Bivariate histogram (a) and scatterplot of (X, Y) (b). Quantile level $\nu = 0.5$, shift in mean $a = 2.5$

Example 1: Shift in mean

(a)



(b)

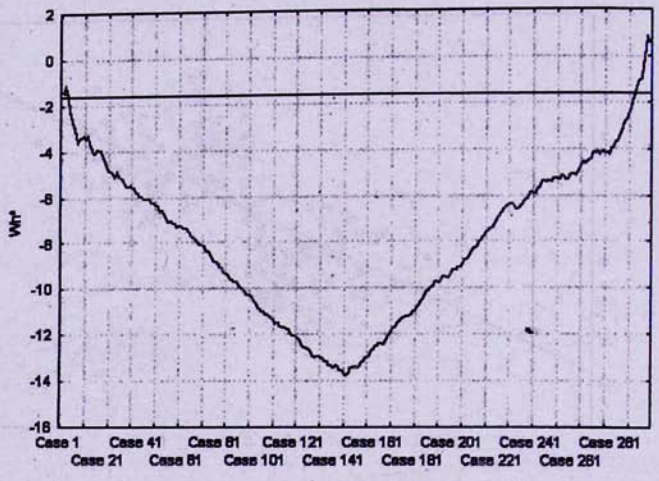


Fig 2. Linear plot of induced order statistics sequence (a) and standartized Wilcoxon statistic (b). The critical level $z_\alpha = -1.64$, the change point estimate $\hat{n} = 147$(real value $n(\nu) = 150$).

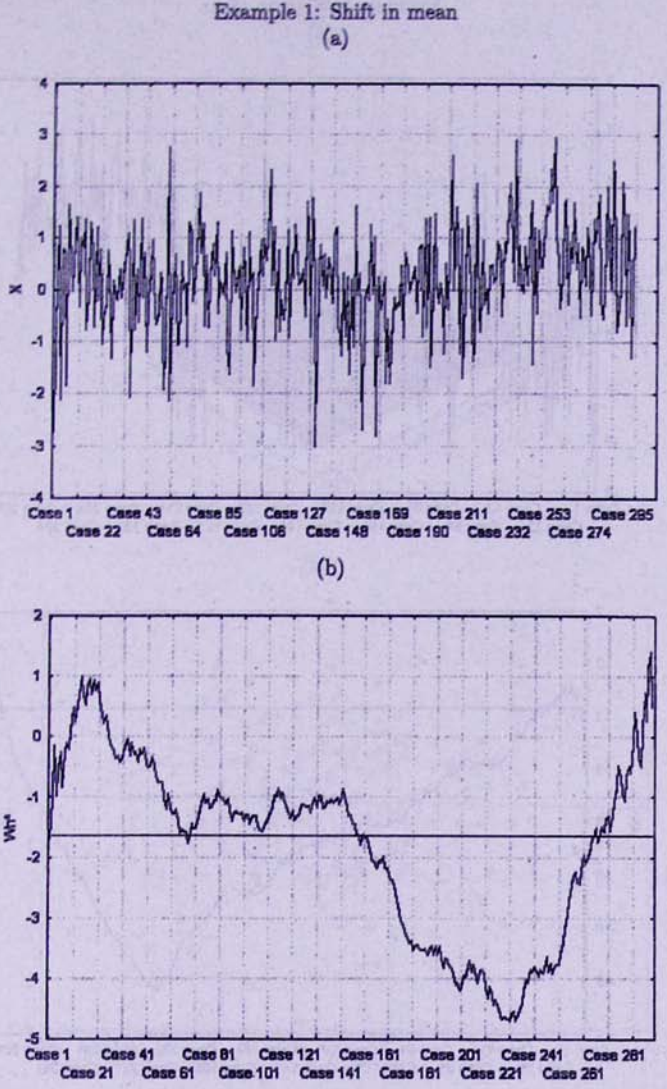
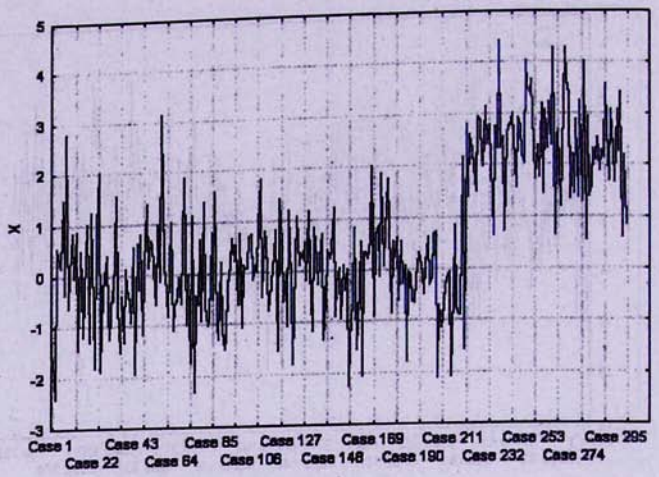


Fig 3. Linear plot of induced order statistics sequence (a) and the behavior of the standardized rank test with Wilcoxon-scores (b). Quantile level $\nu = 0.75$, shift in mean $a = 0.5$.

Example 1: Shift in mean

(c)



(d)

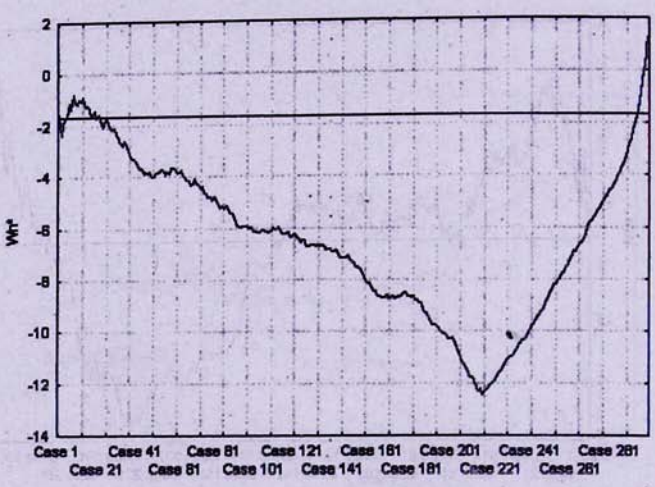
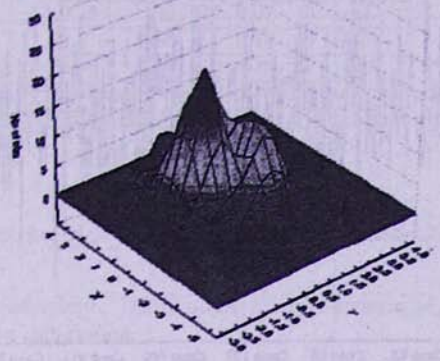


Fig 3. Linear plot of induced order statistics (c), the behavior of the standardized Wilcoxon statistic(d). Quantile level $\nu = 0.75$, shift in mean $a = 2.5$.

Example 2: Shift in scale

(a)



(b)

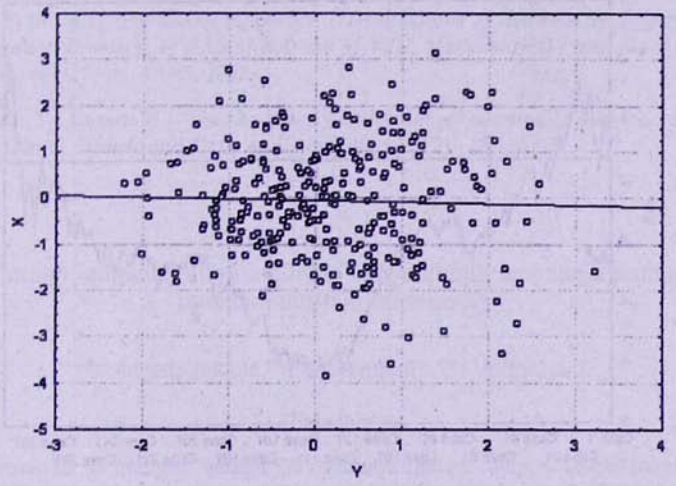
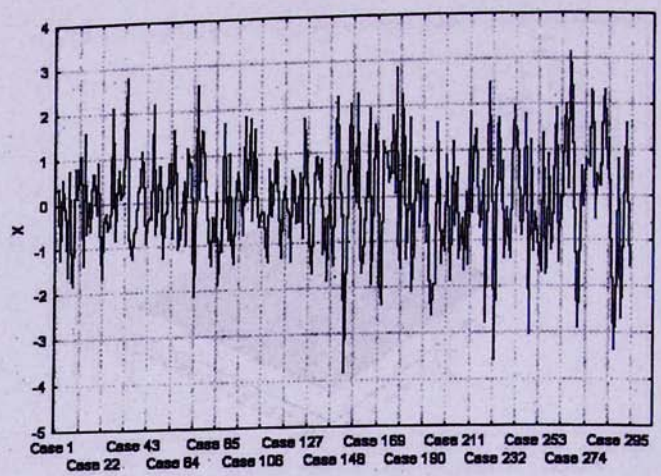


Fig 4. Bivariate histogram (a) and scatterplot of (X, Y) (b).Quantile level $\nu = 0.5$ and scale $\sigma = \sqrt{2}$

Example 2: Shift in scale

(c)



(d)

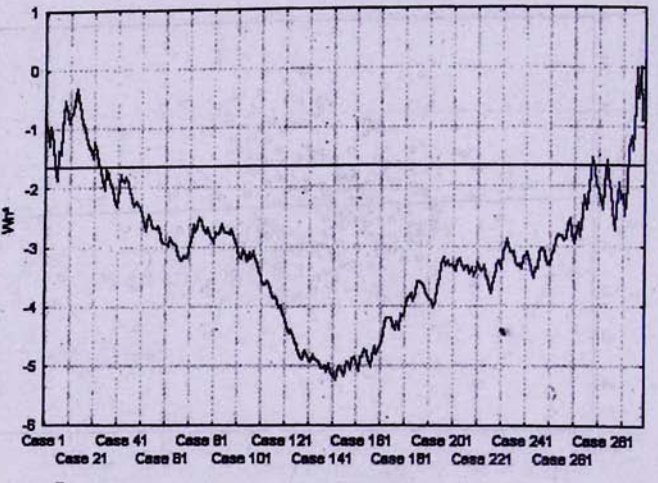


Fig 4. Linear plot of induced order statistics sequence (c) and standardized Mood statistic (d). The critical level $z_\alpha = -1.64$, the change point estimate $\hat{n} = 147$ (real value $n(\nu) = 150$).

In Fig 4. (a), (b), (c), (d) we show bivariate histogram (a) and scatterplot for mixture of two normal distributions (b) of $N(0,1)$ and $N(0,2)$ with mixture weight coefficient $\nu = 0.5$,

induced order statistics sequences (c) and standardized Mood-score statistics (d).

The proposed method allows to analyze the internal structure of data. As it is shown on scatterplots and table 2, variables X and Y have a high correlation coefficient. However the observations can be divided on two conditionally independent groups .

ν	$r(X,Y)$	$r_1(X,Y)$	$r_2(X,Y)$
0.25	0.56	0.07	0.1
0.5	0.65	0.14	-0.04
0.75	0.55	-0.01	-0.13

Table 2. The general correlation coefficient r and correlation coefficients r_1 and r_2 computed before and after the changepoint.

References

[1] Lausen B., Shumacher H., "Maximally selected rank statistics", Biometrics, 48, pp. 73-85, 1992.

[2] Haroutunian E., Safaryan I. "Distributions mixture division with a stratifying parameter", submitted for publication.

[3] Safaryan I., Haroutunian E. "A Common approach to the distributions mixture identification and dependence models analysis", Proceedings of *CSIT 2003*, pp. 184-186.

[4] Haroutunian E. and Safaryan I. "Nonparametric consistent estimation of the change moment of random sequence properties", *Transactions of Institute for Informatics and Automation Problems of NAS of RA and of YSU, Mathematical Problems of Computer Science*, vol. 17, pp. 76-85, 1997.

[5] Hothorn T., Lausen B., "On the exact distribution of maximally selected rank statistics", *Comp. Statist. and Data Anal.*, vol. 43, pp. 121- 137, 2003.

Խառնուրդի երկընտրանքի հանդեպ երկչափանի հաջորդականության համասեռության ստուգումը

Ե. Հարությունյան, Ի. Սաֆարյան և Ա. Մանասյան

Ամփոփում

Հետազոտված է զծային կարգային վիճականիների վարքը մոդելներում, որտեղ դիտարկումների հաջորդականությունները ենթարկվում են տարբեր վիճակագրական բաշխումների: Այդպիսի տվյալները կարելի է մեկնաբանել և որպես վերջավոր թվով բաշխումների խառնուրդ մոդելներ, և որպես կախվածության մոդելներ: Մենք կիրառում ենք տվյալների բազմության մոդելավորում՝ օգտագործված կարգային վիճականիների միջինների և ցրվածքների գնահատականների ստացման համար: Մոդելավորման և ասիմպտոտական արդյունքները բավականաչափ մոտ են: