Mathematical Problems of Computer Science 31, 40–48, 2008.

Copulas of Two-Dimensional Threshold Models

Evgueni A. Haroutunian and Irina A. Safaryan

Institute for Informatics and Automation Problems of NAS of RA.
E-mail: evhar@ipia.sci.am

Abstract

A representation of two-dimensional random vector bivariate distribution by copula is proposed for the case when one of components is categorizing for the other. The regression function and the bounds for the Spearman rank correlation coefficient are derived.

1 Introduction

Definition and comparison of several nonparametric measures of dependence between random vector components is one of the important problems of the risk theory in financial mathematics and medicine. Recently the model of multivariate distribution function, presented by copula, has been used for this goal by many authors. In mathematical terms a copula $C(u_1, ..., u_k)$ is any multivariate distribution with K uniformly distributed marginals on [0,1]. The representation of multidimensional distribution function of a random vector by K-copulas allows to define a joint dependence measure between its component which is free from influence of one-dimensional marginal distributions. As it was shown in monograph of Joe [1] as well as by Schweizer and Wolff in [2] such measures can be presented in the form of various functionals from the difference $C(u_1, ..., u_k) - \Pi_{k=1}^{K} u_k$. An interesting task for practical applications is refinement of these measures in so called threshold models where K-variate distribution function is assumed heterogeneous, which means that, it changes after attaintion by categorizing variable of certain specific threshold. The works of Embrechts, McNeil and Strauman [3], Lausen and Schumacher [4], Safaryan, Haroutunian and Manasyan [?] and others can be mentioned in this direction. We study the copula of a bivariate vector (X, Y) for cases when the categorizing variable X can have one or two thresholds. The regression function $E(Y|X = x)$ and bounds for Spearman's rank correlation coefficient will be derived too.

2 Definition of Threshold Dependence Models

Let (X, Y) be a random vector with a joint distribution function (DF) $F_{XY}(x, y)$ and continuous marginal distributions $F_X(x)$ and $F_Y(y)$.

Definition 1. We call random variable (RV) Y homogeneous with respect to RV X, if for all pairs (x, y) on the plane the following conditional probabilities are equal:

$$\Pr(Y \leq y|X \leq x) = \Pr(Y \leq y|X > x). \qquad (1)$$

It can be shown [4] that the homogeneity and statistical independence of RVs X and Y are equivalent notions. Otherwise Y is called heterogeneous with respect to X, which in that case is called a separating variable for Y. The violation of (1) leads to different models of bivariate dependence. We will consider the simplest cases, when the violation of relationship (1) occurs in one or two points.

Definition 2. If there exist a unique value $x = \mu$ such that for all $y \in R$,

$$\Pr(Y \leq y|X \leq x) = \Pr(Y \leq y|X \leq \mu) \quad \text{for} \quad x \leq \mu, \tag{2}$$

$$\Pr(Y \leq y|X > x) = \Pr(Y \leq y|X > \mu) \quad \text{for} \quad x > \mu, \tag{3}$$

and

$$\Pr(Y \leq y|X \leq \mu) \neq \Pr(Y \leq y|X > \mu), \tag{4}$$

then the statistical dependence between X and Y is called one- threshold, and the value μ is called the threshold.

Definition 3. If there exist no more than two values $x = \mu_1$ and $x = \mu_2$ ($\mu_1 < \mu_2$) such that for all $y \in R$

$$\Pr(Y \leq y|X \leq x) = \Pr(Y \leq y|X \leq \mu_1) \quad \text{for} \quad x \leq \mu_1, \tag{5}$$

$$\Pr(Y \leq y|\mu_1 < X \leq x) = \Pr(Y \leq y|\mu_1 < X \leq \mu_2) \quad \text{for} \quad \mu_1 < x \leq \mu_2, \tag{6}$$

$$\Pr(Y \leq y|X > x) = \Pr(Y \leq y|X > \mu_2) \quad \text{for} \quad x > \mu_2, \tag{7}$$

and

$$\Pr(Y \leq y|\mu_1 < X \leq x) \neq \Pr(Y \leq y|\mu_1 < X \leq \mu_2), \tag{8}$$

$$\Pr(Y \leq y|\mu_1 < X \leq \mu_2) \neq \Pr(Y \leq y|X > \mu_2), \tag{9}$$

then the statistical dependence between X and Y is called two-threshold and values μ_1 and μ_2 are called thresholds.

In similar way M-threshold dependence with $M > 2$ can be define . Let us note that if $Y = g(X)$, where g is some continuous function, then the equality (1) is violated for any point (x, y) on the plane.

3 Representation of Threshold Dependence Structure by Copula

It is known [2] that for arbitrary bivariate DF $F_{X,Y}(x,y)$ with continuous marginal DFs $F_X(x)$ and $F_Y(y)$ there exists unique copula $C_{X,Y}$ such that

$$F_{X,Y}(x,y) = F_{X,Y}(F_X^{-1}(F_X(x), F_Y^{-1}(F_Y(y)) = C_{X,Y}(F_X(x), F_Y(y)).$$

Definition 4. One say that two random vectors (X, Y) and (X', Y') have the same dependence structure if their bivariate distributions corresponds to the same copula.

By substituting different pairs of marginal DFs in some definite copula $C(u, v)$ we can obtain different bivariate DFs with the same dependence structure.

Bellow we will obtain the representation by copula of one-threshold and two-threshold dependence structures.

One-threshold case.

Consider the following conditional DFs $F_{Y|X=x}(y) = \Pr(Y \leq y|X = x)$, $F_{Y|X\leq x}(y) = \Pr(Y \leq y|X \leq x)$ and $F_{Y|X>x}(y) = \Pr(Y \leq y|X > x)$. And let $p = F_X(\mu)$ be the level corresponding to the quantile μ of RV X.

Theorem 1: If RVs X and Y satisfy one-threshold dependence conditions (2)–(4) under threshold $x = \mu$ then their joint DF can be presented as follows:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) + (F_{Y|X\le\mu}(y) - F_{Y|X>\mu}(y))(\min(F_X(x),p) - pF_X(x)). \quad (10)$$

This formula allows to derive the equation to which corresponding copula meets.

Proof: The conditional DFs $F_{Y|X\le x}(y)$ and $F_{Y|X>x}(y)$ can be expressed in terms of joint DF $F_{X,Y}(x,y)$ and marginal DFs $F_X(x)$ and $F_Y(y)$ in the following way:

$$F_{Y|X\le x}(y) = F_{X,Y}(x,y)/F_X(x), \quad (11)$$

$$F_{Y|X>x}(y) = (F_Y(y) - F_{X,Y}(x,y))/(1 - F_X(x)). \quad (12)$$

According to Definition 2 $F_{Y|X\le x}(y) = F_{Y|X\le\mu}(y)$ for all $x \le \mu$. Let us derive the expression for $F_{Y|X\le x}(y)$ for the case $x > \mu$. By substituting (12) into (11) and taking into account (3) we obtain that

$$F_{Y|X\le x}(y) = F_{X,Y}(\mu,y)F_X(\mu)/F_X(x)F_X(\mu)+$$

$$+(F_{X,Y}(x,y) - F_Y(y) + F_Y(y) - F_{X,Y}(\mu,y))/F_X(x) =$$

$$F_{Y|X\le\mu}(y)\frac{F_X(\mu)}{F_X(x)} + \frac{(F_{X,Y}(x,y) - F_Y(y))(1 - F_X(x))}{(1 - F_X(x))F_X(x)} +$$

$$+\frac{(F_Y(y) - F_{X,Y}(\mu,y))(1 - F_X(\mu))}{(1 - F_X(x))F_X(x)} =$$

$$F_{Y|X\le\mu}(y)\frac{F_X(\mu)}{F_X(x)} - F_{Y|X>\mu}(y)\frac{1 - F_X(x)}{F_X(x)} + F_{Y|X>\mu}(y)\frac{1 - F_X(\mu)}{F_X(x)} =$$

$$= F_{Y|X>\mu}(y) + (F_{Y|X\le\mu}(y) - F_{Y|X>\mu}(y))F_X(\mu)/F_X(x). \quad (13)$$

If the components of vector (X,Y) satisfy conditions (2)–(4), then, as it was remarked in [6], the marginal DF of Y can be represented as a mixture of two conditional DFs:

$$F_Y(y) = F_{Y|X\le\mu}(y)p + F_{Y|X>\mu}(y)(1 - p). \quad (14)$$

By substituting (13) and (14) into (11) we obtain that

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) + \begin{cases} (F_{Y|X\le\mu}(y) - F_{Y|X>\mu}(y))(1-p)F_X(x), & x \le \mu, \\ (F_{Y|X\le\mu}(y) - F_{Y|X>\mu}(y))(1 - F_X(x))p, & x > \mu. \end{cases}$$

The last expression is equivalent to (10) and so the Theorem 1 is proved.

It is by Schweizer and Wollf [2] that for any copula $C(u,v)$ including the independence copula following inequalities hold:

$$C^-(u,v) \le C(u,v) \le C^+(u,v). \quad (15)$$

We shale use following notations for three special copula :
$C^0(u,v) = uv$ the copula of independence, $C^+(u,v) = \min(u,v)$ the upper boundary for copulas, $C^-(u,v) = \max((u+v-1),0)$ the lower boundary for copulas. This inequality provides opportunity to specify the type of random vector (X,Y) dependence structure according to the disposition of its copula and the independence copula.Thus, if the surface $z = C_{X,Y}(u,v)$ is between the surfaces $z = C^0(u,v)$ and $z = C^+(u,v)$ then RV X and Y are called in accord with definition given by H. Joe [1] positive quadrant dependent (PQD).

The following corollary of Theorem 1 allows to obtain the upper and lower bounds for copula of vector (X, Y) with one–threshold dependence.

Corollary 1: The dependence structure of random vector (X, Y) with two-dimensional DF (10) is defined by a copula $C_{X,Y}(u, v)$ satisfying the following equation:

$$C_{X,Y}(u,v) = C^0(u,v) + (C_{X,Y}(p,v) - pv)(C^+(u,p) - pu)/p(1-p). \quad (16)$$

while the upper and lower bounds of such copula are the following:

$$C^+_{X,Y}(u,v) = C^0(u,v) + (C^+(p,v) - pv)(C^+(u,p) - pu))/p(1-p), \quad (17)$$

$$C^-_{X,Y}(u,v) = C^0(u,v) + (C^-(p,v) - pv)(C^+(u,p) - pu))/p(1-p). \quad (18)$$

Proof: Let us remark that the second factor in the right part of (10) is equal to $(C^+(u,p) - pu)$, while the first factor has the representation of the form

$$F_{Y|X\le\mu}(y) - F_{Y|X>\mu}(y) = (F_{X,Y}(\mu, y) - pF_Y(y))p(1-p) =$$

$$= (C_{X,Y}(p,v) - pv)/p(1-p).$$

By substitution of the last expression in (10) we obtain (16). The expressions for the upper and lower bounds for the copula satisfying equality (10) are derived by substitution of (15) into (16).

On Fig.1 the upper bound for the copula of RVs X and Y satisfying one–threshold dependence are presented. On Fig. 2 diagonal sections of surfaces $z = C^+(u,v)$, $z = C^0(u,v)$ and $z = C^+_{X,Y}(u,v)$ presented.

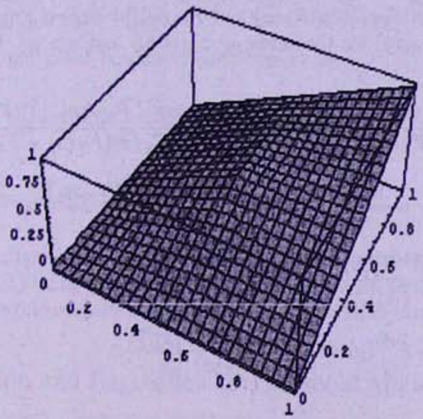

Figure 1: The upper bound for the copula of random vector with one-threshold dependence of components,in case $p = 0.5$

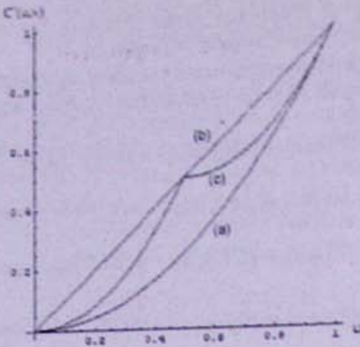

Figure 2: Diagonal sections of independence copula (a), upper bound of copulas (b), and upper bound of one-threshold copulas in case $p = 0.5$ (c).

Two-threshold case.
Let $p_1 = F_X(\mu_1)$ and $p_2 = F_X(\mu_2)$ be two levels corresponding to quantiles $\mu_1 < \mu_2$ of RV X.

Theorem 2: If RVs X and Y satisfy two-threshold dependence conditions (5)–(9) under thresholds $x = \mu_1$ and $x = \mu_2$ ($\mu_1 < \mu_2$), then their joint distribution function can be presented as follows:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) + (\min(F_X(x), p_1) - p_1F_X(x))(F^{(1)}(y) - F^{(2)}(y)) +$$
$$+(\min(F_X(x), p_2) - p_2F_X(x))(F^{(2)}(y) - F^{(3)}(y)), \tag{19}$$

where $F^{(1)}(y) = F_{Y|X\le\mu_1}(y)$, $F^{(2)}(y) = F_{Y|\mu_1<X\le\mu_2}(y)$, $F^{(3)}(y) = F_{Y|X>\mu_2}(y)$.

Proof: Since $F_{X,Y}(x,y) = F_{Y|X\le x}(y)F_X(x)$ and $F_{Y|X\le x}(y) = F_{Y|X\le\mu}(y)$, then we shall derive the expression for $F_{Y|X\le x}(y)$ for the cases $\mu_1 < x \le \mu_2$ and $x > \mu_2$. For $\mu_1 < x \le \mu_2$ we obtain that

$$F_{Y|X\le x}(y) = \frac{F_{X,Y}(x,y)}{F_X(x)} = \frac{F_{X,Y}(\mu_1,y)F_X(\mu_1)}{F_X(x)F_X(\mu_1)} + \frac{(F_{X,Y}(x,y) - F_{X,Y}(\mu_1,y))(F_X(x) - F_X(\mu_1))}{F_X(x)(F_X(x) - F_X(\mu_1))} =$$
$$= \frac{F^{(1)}(y)p_1}{F_X(x)} + \frac{\Pr\{Y \le y|\mu_1 < X \le x\}(F_X(x) - p_1)}{F_X(x)} =$$
$$= \frac{F^{(1)}(y)p_1}{F_X(x)} + \frac{F^{(2)}(y)(F_X(x) - p_1)}{F_X(x)} =$$
$$= F^{(2)}(y) + \frac{(F^{(1)}(y) - F^{(2)}(y))p_1}{F_X(x)}. \tag{20}$$

For $x > \mu_2$ we have that

$$F_{Y|X\le x}(y) = \frac{F_{X,Y}(x,y) + F_{X,Y}(\mu_1,y) - F_{X,Y}(\mu_1,y) + F_{X,Y}(\mu_2,y) - F_{X,Y}(\mu_2,y)}{F_X(x)} =$$

$$= \frac{F^{(1)}(y)p_1 + F^{(2)}(y)(p_2 - p_1) + F^{(3)}(y)(1-p_1) - F^{(3)}(y)(1 - F_X(x))}{F_X(x)} =$$

$$= \frac{F^{(1)}(y)p_1}{F_X(x)} + = \frac{F^{(2)}(y)(p_2 - p_1)}{F_X(x)} - \frac{p_2 F^{(3)}(y)}{F_X(x)} =$$

$$= F^{(3)}(y) + \frac{p_1(F^{(1)}(y) - F^{(2)}(y))}{F_X(x)} + \frac{p_2(F^{(2)}(y) - F^{(1)}(y))}{F_X(x)}. \tag{21}$$

If conditions (5)-(9) holds, then DF of Y can be presented as follows:

$$F_Y(y) = F^{(1)}(y)p_1 + F^{(2)}(y)(p_2 - p_1) + F^{(3)}(y)(1-p_2). \tag{22}$$

Thus from (20), (21) and (22) we obtain that

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)+$$

$$+\begin{cases} (1-p_1)(F^{(1)}(y) - F^{(2)}(y))F_X(x) + (1-p_2)(F^{(2)}(y) - F^{(3)}(y)F_X(x), \; x \le \mu_1, \\ p_1(F^{(1)}(y) - F^{(2)}(y))(1 - F_X(x)) + (1-p_2)(F^{(2)}(y) - F^{(3)}(y))F_X(x), \; \mu_1 < x \le \mu_2, \\ p_1(F^{(1)}(y) - F^{(2)}(y))(1 - F_X(x)) + p_2(F^{(2)}(y) - F^{(3)}(y))(1 - F_X(x)), \; x > \mu_2. \end{cases} \tag{23}$$

The statement of the theorem immediately follows from the last express. The partial case of two-threshold model presents interest for medical applications, namely the case $F^{(1)}(y) = F^{(3)}(y)$ called "epidemic". Providing X is the time variable, then thresholds μ_1 and μ_2 correspond to the times of start and end of epidemics.

Corollary 2: The dependence structure of random vector (X, Y) with two-dimensional DF (19) is defined by a copula $C_{X,Y}(u, v)$ satisfying the equation

$$C_{X,Y}(u,v) = uv + (p_2 C_{X,Y}(p_1, v) - p_1 C_{X,Y}(p_2, v))C^+(u,v)/p_1(p_2 - p_1)+$$
$$+(C_{X,Y}(p_2, v)(1-p_2) - v(p_2 - p_1))C^+(u,v)/(p_2 - p_1)(1-p_2) \tag{24}$$

and the upper and lower bounds of such copula are

$$C^+_{X,Y}(u,v) = \frac{C^+(p_2,v)C^+(u,p_1)}{p_1} + \frac{(C^+(p_2,v)(1-p_1) - v(p_2 - p_1))C^+(u_1,p_2)}{(p_2 - p_1)(1-p_2)}. \tag{25}$$

$$C^-_{X,Y}(u,v) = \frac{C^-(p_2,v)C^+(u,p_1)}{p_1} + \frac{(C^-(p_2,v)(1-p_1) - v(p_2 - p_1))C^+(u_1,p_2)}{(p_2 - p_1)(1-p_2)}. \tag{26}$$

The proof is similar to the one of Corrolary 1. One can expect, that under certain conditions the boundary (25) be upper the boundary (12), while the boundary (26) below the boundary (18), i. e. the two threshold dependence is stronger than one-threshold.

4 Correlation and Regression in Threshold Models

Since any copula is some surface in a unique cube, then the measure of deviation from independence, as it was shown in [2] can be expressed with some functional from difference $C_{X,Y}(u,v)$ and $C^0_{X,Y}(u,v)$ having the meaning of distance in three dimensional space. One of such functionals $\rho(X,Y)$ the coefficient of Spearman's rank correlation coefficient is expressed with copula in the following way [2]:

$$\rho(X,Y) = 12\int_0^1\int_0^1 (C_{X,Y}(u,v) - C^0_{X,Y}(u,v))dudv.$$

By integrating (17) and (18) on the interval $[0,1]$ by u and v we come to the following:

Corollary 3: The bounds for the Spearman's rank correlation coefficient in case of one threshold dependence structure (10) are determined by inequalities:

$$-3p(1-p) \le \rho(X,Y) \le 3p(1-p).$$

The maximal value of these bounds is achieved by $p = 0.5$. Hence, for bivariante distribution of form (10) Spearman's correlation coefficient does not exceed in modulus at 0.75. We will prove that the obtained estimates are rather coarse since under conditions of one-threshold dependence (2) – (4) the regressional link between RV X and Y is missed. We will derive hereafter the expression for the regression function of Y to X.

Theorem 3: If conditional DF $F_{Y|X\le x}(y)$ is differentiable by x then the regression function $E(Y|X=x)$ is either constant or step function with a jump on the threshold $x=\mu$.

Proof: Let us prove that the relation between the conditional distributions $F_{Y|X\le x}(y)$ and $F_{Y|X=x}(y)$ is defined by the relation

$$F_{Y|X=x}(y) = F_{Y|X\le x}(y) + F_X(x)(\frac{\partial}{\partial x}F_{Y|X\le x}(y))/f_X(x). \tag{27}$$

We denote two-dimensional density function by

$$f_{X,Y}(x,y) = \frac{\partial F_{X,Y}(x,y)}{\partial x\partial y},$$

then the conditional distribution functions can be expressed as

$$F_{Y|X=x}(y) = \lim_{h\to 0} \Pr(Y\le y|x<X\le x+h) =$$

$$= \lim_{h\to 0}(\int_{-\infty}^{y}\int_{x}^{x+h} f_{X,Y}(t,z)dtdz/\int_{x}^{x+h} f_X(t)dt) =$$

$$= \int_{-\infty}^{y} f_{X,Y}(x,z)dz/\ f_X(x).$$

and

$$F_{Y|X=x}(y) = F_{X,Y}(x,y)/F_X(x) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(t,z)dtdz/F_X(x).$$

By differentiating the last expression on x and substituting it into previous identity we derive (23).

Since for all $x\le\mu$ we have $F_{Y|X\le x}(y) = F_{Y|X\le\mu}(y)$, then $\frac{\partial}{\partial x}F_{Y|X\le x}(y) = 0$ in this interval. Under $x>\mu$ the following equality is true:

$$F_{Y|X\le x}(y) = F_{Y|X>\mu}(y) + (F_{Y|X\le\mu}(y) - F_{Y|X>\mu}(y))F_X(\mu)F_X(x).$$

Consequently

$$\frac{\partial}{\partial x}F_{Y|X\le x}(y) = -(F_{Y|X\le x}(y) - F_{Y|X>\mu}(y))\frac{F_X(\mu)f_X(x)}{f_X^{(}x)}.$$

By substituting expression for $\frac{\partial}{\partial x}F_{Y|X\le x}(y)$ under $x \le \mu$ and $x > \mu$ in (23) we obtain

$$F_{Y|X=x}(y) = \begin{cases} F_{Y|X\le\mu}(y), & x \le \mu, \\ F_{Y|X>\mu}(y), & x > \mu. \end{cases}$$

Thus

$$E(Y|X=x) = \int_{-\infty}^{+\infty} y dF_{Y|X=x}(y) = \begin{cases} \int_{-\infty}^{+\infty} y dF_{Y|X\le\mu}(y) = k_1, & x \le \mu, \\ \int_{-\infty}^{+\infty} y dF_{Y|X>\mu}(y) = k_2, & x > \mu. \end{cases}$$

If constants k_1 and k_2 coincide, then $E(Y|X=x)$ is constant, otherwise $k_1 \neq k_2$, then it is stepwise constant with jump at $x = \mu$.

Let us note, that the coincidence of constants k_1 and k_2 is connected with the type of difference of conditional DFs $F_{Y|X\le\mu}(y)$ and $F_{Y|X>\mu}(y)$. For example in case of difference in shift, i.e. if

$$F_{Y|X>\mu}(y) = F_{Y|X\le\mu}(y-a), \quad a > 0,$$

the inequality $k_1 < k_2$ is true. In case of scale difference $F_{Y|X>\mu}(y) = F_{Y|X\le\mu}((1+a)y)$. $k_1 = k_2$ under condition that $F_{Y|X>\mu}(y)$ and $F_{Y|X\le\mu}(y)$ are symmetrical with respect to origin (i.e. $F_.(y) = 1 - F_.(-y)$) DF.

5 Conclusion

The threshold dependence models are often observed in real research. Meanwhile sampling correlation coefficients of Pearson and Spearman revealed to be sufficiently high and the linear regression built on observed data is significant. It is so called "false" regression. To obtain an adequate regression dependence it is necessary to test the homogeneity of response with respect to predictor. The methodology for such testing is described in [4], [5] and [6].

References

[1] H. Joe, "Multivariate models and dependence concepts", *Chapman and Hall, London*, 1997.

[2] B. Schweizer and E. F. Wolff, "On nonparametric measures of dependence for random variable", *The Annals of Stat.*, vol. 9, num. 6, pp. 879-886, 1981.

[3] P. Embrechts, A. McNeil and D. Strauman, "Correlation and dependence in risk managment. Properties and pitfalls", *In: Risk Management: Value at Risk and Beyond (M Dempster Ed.) Cambridge University Press,*, pp. 176–223, 2002.

[4] B. Lausen and M. Shumacher, "Maximal selected rank statistics", *Biometrics*, vol. 48, pp. 72–85, 1992.

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

[6] Safaryan I., Haroutunian E. and Manasyan A., "Two-dimentsional sequence homogeneity testing against mixture alternative", *Mathematical problems of computer science*, vol. 23, pp. 67–79, 2004.

Երկչափ շեմքային մոդելների կապակցիչը

Ե. Հարությունյան և Ի. Սաֆարյան

Ամփոփում

Առաջարկված է երկչափ պատահական վեկտորի բաշխման ֆունկցիայի կապակցիչի (copula) միջոցով ներկայացումը այն դեպքի համար երբ վեկտորի բաղադրիչներից մեկը պատող է մյուսի համար: Ստացված են ռեգրեսիոն ֆունկցիան և Սպիրմենի կարգային հարաբերակցության գործակցի վերին և ներքին գնահատականները: