On Certain Threshold Copulas Estimators

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

Irina Safaryan
Institute for Informatics and Automation Problems
of NAS of RA
e-mail: safari@ipia.sci.am

ABSTRACT

A method of construction of empirical associate measures for components of two-dimensional random vector in threshold copula models is proposed. It is shown that for such models Spearman's rank correlation coefficient and Kendall's concordance coefficient are the linear functions of two-sample Wilcoxon statistics.

Keywords

Copula, empirical associate measures, threshold dependence model, Kendall's concordance coefficient, Spearman's rank correlation

1. INTRODUCTION

In two-dimensional data analysis situations arise when rather high values of certain empirical measures of association such as Spearman's rank correlation and Kendall's concordance cannot be interpreted as monotone dependence between components of two-dimensional random vector. In particular, this refers to threshold dependence structures in which one of the components serves as categorizing variable for the other. In case under consideration the observed sequence can be divided on two or more groups concentrating along some line on plane, while correlation between components inside each group is missed.

The separation of two, or many groups is based on quantiles of the categorizing variable, called thresholds which are not known in general. To obtain consistent estimates of unknown threshold in one-threshold model Safaryan, Haroutunian and Manasyan in [1] applied the change-point detection technique. The representation of dependence of two dimensional random vector components by copulas was derived in [2]. We propose a method of construction of empirical measures for two-dimensional dependence structures with applying the results obtained in [2] as well as the function of concordance between two copulas defined by Nelsen in [3] and empirical copula notion introduced by Fermanian, Radulovic and Wegkamp in [4].

2. ASSOCIATION MEASURES FOR ONE-THRESHOLD COPULA MODELS

Let (X, Y) be a random vector with two-dimensional distribution function (DF) $F(x, y)$, continuous marginal DF's $F_X(x)$ and $F_Y(y)$ and corresponding copula $C(u, v)$. We remind that copula $C(u, v)$ is a function, which connects two-dimensional DF with marginal DF's by relationship

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

The concordance function $K(C_1, C_2)$, defined by Nelsen in [3], denotes the difference between the probabilities of concordance and disconcordance of random vectors (X_1, Y_1) and (X_2, Y_2) with copulas $C_1(u, v), C_2(u, v)$, i.e.

$$K(C_1, C_2) = \Pr((X_1 - X_2)(Y_1 - Y_2) > 0) -$$

$$-\Pr((X_1 - X_2)(Y_1 - Y_2) < 0).$$

It can be presented in the form

$$K(C_1, C_2) = 4 \int_D \int C_2(u, v) dC_1(u, v) - 1, \qquad (1)$$

or in an equivalent form as

$$K(C_1, C_2) = 1 - 4 \int_D \int \frac{\partial C_2(u, v)}{\partial u} \frac{\partial C_1(u, v)}{\partial u} dudv, \qquad (2)$$

where $D = [0, 1] \times [0, 1]$ is the unique square.

Then, as it is shown in [3] many measures of association between random variables (RV's) X and Y whose copula is C can be defined in terms of concordance function. For instance, Spearmans correlation coefficient ρ_C is proportional to concordance function with arguments $C_1 = C$ and $C_2 = \Pi = uv$, that is

$$\rho_C = 3K(C, \Pi).$$

and Kendall's concordance coefficient τ_C is equal to $K(C, C)$ Thus taking in account (1) well-known representation for ρ_C and τ_C can be obtained, namely,

$$\rho_C = 12 \int_0^1 \int_0^1 (C(u, v) - uv) dudv \qquad (3)$$

and

$$\tau_C = 12 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1. \qquad (4)$$

We introduce notion of threshold copulas and derive corresponding expressions for ρ_C and τ_C.

Definition 1: A copula $C_p(u, v)$ depending on scalar parameter $p \in (0, 1)$, is called one–threshold if $u = p$ is a single point for which the following relations hold

$$\frac{C_p(u, v)}{v} = \frac{C_p(p, v)}{p}, \qquad u \leq p,$$

$$\frac{v - C_p(u, v)}{1 - u} = \frac{v - C_p(p, v)}{1 - p}, \qquad u > p,$$

and

$$C_p(u, v) \neq pv.$$

Theorem 1: *Spearman's correlation coefficient ρ_{C_p} for one-threshold copula C_p is the following:*

$$\rho_{C_p} = 6\int\limits_0^1 C_p(p,v)dv - 3p. \qquad (5)$$

For Kendall's concordance coefficient the following relation is true

$$\tau_{C_p} = \frac{2}{3}\rho_{C_p}.$$

Proof: We note that definition of one-threshold copula completely coincide with definitions of the threshold dependence between RV's X and Y expressed in terms of conditional distributions of RV Y under conditions $\{X \le \mu\}$, $\{X > \mu\}$, brought in [3] if $p = F_X(\mu)$. Consequently, C_p can be represented as follows

$$C_p(u,v) = uv + \frac{1}{p(1-p)}(C(p,v) - pv)(\min(u,p) - pu). \quad (6)$$

The further proof immediately follows by substitution of (6) in (3) and (4) and integration of the derived quintile.

The obtained expressions for Spearmann's and Kendall's coefficients allow to construct estimators for τ_C and ρ_C which are based only on ranks of Y and propose a new algorithm for the estimation of parameter p .

3. ESTIMATORS FOR ONE-THRESHOLD DEPENDENCE CASE

The sample versions of measures of association can be described by empirical copula function $\hat{C}_N(u,v)$, which has been established by Fermanian, Radulovic and Wegkamp in [4].

Let $\{(X_n,Y_n)_{n=1}^N\}$ be a random sample from RV (X,Y), with DF $F(x,y)$, marginal DF's $F_X(x)$ and $F_Y(y)$ and copula $C(u,v)$. Consider the following empirical functions:

$$\hat{F}_{N,X}(x) = \#\{X_n : X_n \le x\} = \frac{1}{N}\sum_{n=1}^N \mathbf{1}\{X_n \le x\},$$

where $\mathbf{1}$ $\{\mathcal{A}\}$ is the indicator of event $\mathcal{A}$. Let

$$\hat{F}_{N,Y}(y) = \frac{1}{N}\sum_{n=1}^N \mathbf{1}\{Y_n \le y\},$$

$$\hat{F}_N(x,y) = \#\{(X_n,Y_n) : X_n \le x, Y_n \le y\} =$$

$$= \frac{1}{N}\sum_{n=1}^N \mathbf{1}\{X_n \le x\} \times \mathbf{1}\{Y_n \le y\}.$$

The cadlag version of empirical copula, according to [4] is defined as follows

$$\hat{C}_N(u,v) = \frac{1}{N}\sum_{n=1}^N \mathbf{1}\{\hat{F}_{N,X}(X_n) \le u, \hat{F}_{N,Y}(Y_n) \le v\}. \quad (7)$$

A representation of empirical copula suggested in [5] is the following

$$\hat{C}_N(u,v) = \frac{1}{N}\sum_{n=1}^N \mathbf{1}\{\frac{R_{X_n}}{N+1} \le u\} \times \mathbf{1}\{\frac{R_{Y_n}}{N+1} \le v\}, \quad (8)$$

where R_{X_n} and R_{Y_n} are ranks of RV's X_n and Y_n in sequences $\{X_n\}_{n=1}^N$ and $\{Y_n\}_{n=1}^N$ respectively. After replacing $C(u,v)$ in (3) with $\hat{C}_N(u,v)$ given in (8) we obtain Spearman's rank correlation coefficient between RV's X and Y, in the form

$$\hat{\rho}_C = \frac{12}{N(N+1)^2}\sum_{n=1}^N (R_{X_n} - \frac{N+1}{2})(R_{Y_n} - \frac{N+1}{2}).$$

For the threshold models we obtain sampling estimates of ρ_{C_p} by substituting (8) in (5). Then the following theorem holds.

Theorem 2: *Expression of Spearman's rank coefficients for one-threshold model are the following:*

$$\hat{\rho}_{C_p} = \frac{6}{(N+1)N}\sum_{n=1}^{[Np]} (N+1-R_{Y_n'}) - 3p, \qquad (9)$$

Where $\{Y_n^{'}, \}_{n=1}^N$ is the of induced order statistics sequence, (i.e. for $X_{(1)} < X_{(2)} < \ < X_{(N)}$ induced statistic $Y_n^{'}$ is defined as $Y_n^{'} = Y_i$, if $X_{(n)} = X_i$.

Proof:

$$\hat{\rho}_{c_p} = 6\int\limits_0^1 \hat{C}_N(p,v)dv - 3p =$$

$$\frac{6}{N(N+1)}\sum_{n=1}^N\sum_{j=1}^N \mathbf{1}\{R_{X_n} \le p\} \times \mathbf{1}\{R_{Y_n} \le \frac{R_{Y_j}}{N+1}\} - 3p =$$

$$= \frac{6}{N(N+1)}\sum_{n=1}^{[Np]}\sum_{j=1}^N \mathbf{1}\{R_{Y_n'} \le R_{Y_j}\} - 3p.$$

The last expression proves (9).

Let

$$W_N(\frac{n}{N}) = \frac{N}{N-n}(\frac{1}{n}\sum_{i=1}^n \frac{R_{Y_i'}}{N+1} - \frac{1}{2}), \ n = \overline{1, N-1}, \quad (10)$$

is the sequence of Wilcoxon statistic, which tests homogeinety of two samples, $\{Y_i^{'}\}_{i=1}^n$ and $\{Y_i^{'}\}_{i=n+1}^N$.
Then from Theorem 2 can be deduced the following consequences:

Corollary 1: Rank correlation coefficient of Spearman is the linear function from Wilcoxon statistic $W_N(\frac{n(p)}{N}), n(p) = [pN]$ defined by relation

$$\hat{\rho}_{c_p} = -\frac{6(N-n(p))n(p)}{N^2}W_N(\frac{n(p)}{N}) + 3(\frac{n(p)-Np}{N}). \quad (11)$$

We use (11) to estimate ρ_{c_p} if $n(p)$ is known, otherwise a change point technique proposed in [1] can be applied

Let

$$\hat{n} = arg \max_{0<n<N} |W_N(\frac{n}{N})|$$

and

$$W_N^* = W_N(\frac{\hat{n}}{N})\sqrt{12(1-\frac{\hat{n}}{N})\hat{n}}.$$

Then, a consisent estimate of ρ_{c_p} is defined in the following

Corollary 2: If $W_N^* > z_\alpha$, or $W_N^* < 1 - z_\alpha$, where z_α is quantile of level α of RVZ distributed as $\mathcal{N}(0,1)$, then the consistent estimate of p is defined by

$$\hat{p} = \frac{\hat{n}}{N},$$

and consistent estimator of ρ_{C_p} is the following

$$\hat{\rho}_{C_p} = -6(1-\hat{p})\hat{p}W_N(\hat{p}).$$

Proof: The induced order statistic sequence $\{Y_{n,X}\}_{n=1}^{N}$ can be viewed as a random sample from conditional DF $F(y|x) = Pr\{Y \leq y|X = x\}$. As it follows from Definition 1 that probabilities

$$Pr\{Y_n^{'} \leq y|X = F^{-1}(p)\}, \quad n = \overline{1,N},$$

are the same if $n \leq [Np]$ and the other if $n > [Np]$ then the number $n(p) = [Np]$ can be called the shangepoint for the sequence $\{Y_n^{'}\}_{n=1}^{N}$.

4. ESTIMATORS FOR TWO-THRESHOLD DEPENDENCE CASE

Similar results can be obtained for cases of two and more thresholds.

Definition 2: A copula $C_{\mathbf{p}}$ depending on vector parameter $\mathbf{p} = (p_1, p_2)$, $0 < p_1 < p_21 < 1$ is called two-threshold is there exist exactly two values $u = p_1$ and $u = u_2$ such that relations holds

$$\frac{C_{\mathbf{p}}(u,v)}{u} = \frac{C_{\mathbf{p}}(p_1,v)}{p_1}, \quad for \quad u \leq p_1,$$

$$\frac{C_{\mathbf{p}}(u,v) - C_{\mathbf{p}}(p_1,v)}{u - p_1} = \frac{C_{\mathbf{p}}(p_2,v) - C_{\mathbf{p}}(p_1,v)}{p_2 - p_1},$$

$$for \quad p_1 < u \leq p_2,$$

$$\frac{v - C_{\mathbf{p}}(u,v)}{1-u} = \frac{v - C_{\mathbf{p}}(p_2,v)}{1-p_2}, \; for \; u > p_2,$$

and

$$p_2 C_{\mathbf{p}}(p_1,v) \neq p_1 C_{\mathbf{p}}(p_2,v),$$

$$(p_2 - p_1) \neq p_1 C_{\mathbf{p}}(p_2,v).$$

Theorem 3: Spearman's correlation coefficient ρ_{c_p} for two-threshold copula $C_{\mathbf{p}}$ is equal to

$$\rho_{c_p} = 6\int\limits_0^1 p_2 C_{\mathbf{p}}(p_1,v) + (1-p_1)C_{\mathbf{p}}(p_2,v) - 3p_2.$$

For Kendall's concordance coefficients the following relation is true

$$\tau_{c_{\mathbf{p}}} = \frac{2}{3}\rho_{c_{\mathbf{p}}}.$$

Proof: We note that dependence between RV's X and Y expressed in terms of conditional distributions of RV Y under conditions $\{X \leq \mu_1\}$, $\{X > \mu_2\}$, $\{\mu_1 < X \leq \mu_2\}$ brougth in [3], if $p_1 = F_X(\mu_1)$, $p_2 = F_X(\mu_2)$. can be represented by $C_{\mathbf{p}}$ as follows

$$C_p(u,v) = uv + \Delta_1(v,\mathbf{p})(\min(u,p_1) - p_1 u) +$$

$$\Delta_1(v,\mathbf{p})(\min(u,p_2) - p_2 u),$$

where

$$\Delta_1 = \frac{1}{p_1(p_2 - p_1)}(p_2(C_p(p_1,v) - p_1 v) - p_1(C_p(p_2,v) - p_2 v)),$$

and

$$\Delta_2 = \frac{(1-p_1)(C_p(p_1, -p_2 v) - (1-p_2)(C_p(p_1,v) - p_1 v))}{(1-p_2)(p_2-p_1)}$$

The proof is similar of the proof of Theorem 1.

Theorem 4. Expression of Spearman's rank coefficient for two-threshold model is the following:

$$\hat{\rho}_{c_{\mathbf{p}}} = \frac{6}{(N+1)N}(\sum_{n=1}^{[Np_1]} p_2(N+1-R_{Y_n'}) +$$

$$+ \sum_{n=1}^{[Np_2]} (1-p_1)(N+1-R_{Y_n'} - 3p_2)$$

The further extension of the noted results are connected with empirical threshold copulas of the K-dimensional random vector $X^{(1)}, ..., X^{(K)}$ in the case, when RV $X^{(1)}$ serves categorizing variable for $X^{(2)}, ..., X^{(K)}$.

REFERENCES

[1] E. Haroutunian, I. Safaryan and A. Manasyan, "Two-dimentsional sequence homogeneity testing against mixture alternative", *Mathematical Problems of Computer Science*, vol. 23, pp. 67–79, 2004.

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

[3] B. R. Nelsen,"Concordance and copulas", *The Annals of Statistics*, vol. 9, no. 6, pp. 879–886, 1981.

[4] J. D. Fermanian, D. Radulovic and M. Wegkamp, "Weak convergence of empirical copula processes", *Bernoulli*, vol. 10, no. 5, pp. 847–860, 2004.

[5] B. Remillard and O. Scaillet, "Testing for equality between two copulas", *Journal of Multivariate Analysis*, vol. 100, no. 3, pp. 377– 386, 2008.