PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Krzysztof Jajuga
Wroclaw University of Economics

STATISTICAL PARAMETERS BASED ON COPULA FUNCTIONS

1. Introduction

Analysis of multivariate data is one of the most important practical tasks, where statistical methods are used. When one assumes stochastic approach, then the most common theoretical framework of analysis is multivariate distribution. The most often studied multivariate distribution is, of course, multivariate normal distribution.

From the point of view of practitioners, important part of data analysis is the determination of different statistical parameters characterizing data set. In the univariate case, the analyzed parameters are: location parameter (mean, median, etc), scale parameter (standard deviation, interquartile range, etc.), skewness parameter and kurtosis parameter.

If we move to multivariate data set, then the main parameters are: location vector (for multivariate normal distribution: mean vector) and scatter matrix (for multivariate normal distribution: covariance matrix). This approach is a classical one in the analysis of multivariate distribution, where off-diagonal parameters of scatter matrix contain "joint" information about scale and dependence. For example, covariance between two variables is a product of standard deviation of the first variable, standard deviation of the second variable and correlation coefficient between two variables.

In this paper we present another approach, based on the so called copula functions. The key point of this approach lies in the fact that dependence parameter is treated separately, rather than being linked with scale parameters, as it is in scatter matrix. This approach has been used rather rarely in practice. After giving some basic results, we will review several groups of parameters, derived from copula approach.

2. Copula analysis – main theoretical results

The main idea behind the application of copula functions in the analysis of multivariate data lies in the separation of the analysis of the marginal univariate distributions from the analysis of the dependence between univariate components of random vector. This idea is reflected in the so called Sklar theorem, given in the following representation of the multivariate distribution function [Sklar 1959]:

$$F(x_1, ..., x_m) = C(F_1(x_1), ..., F_m(x_m)), \tag{1}$$

where: F – distribution function of a random vector;
F_i – distribution function of the *ith* component of a random vector;
C – copula function.

As one can see from (1), copula function is a distribution function of multivariate uniform distribution. On the other hand, it is also multivariate distribution function defined for quantiles of univariate marginals, as in the following form:

$$C(u_1, ..., u_n) = F(F_1^{-1}(u_1), ..., F_n^{-1}(u_n)). \tag{2}$$

One can also notice that in (1) marginal univariate distributions are "separated" from dependence function, given as copula function.

A closely related notion is copula density function being the *mth* derivative of copula function:

$$c(u_1, ..., u_m) = \partial^m C(u_1, ..., u_m). \tag{3}$$

This allows representing multivariate density function in the following form:

$$f(x_1, ..., x_m) = c(F_1(x_1), ..., F_m(x_m)) \cdot f_1(x_1) \cdot ... \cdot f_m(x_m) \tag{4}$$

where: f – density function of a random vector;
f_i – density function of the *ith* component of a random vector;
c – copula density function.

Copula function can be interpreted through the notions of probability. Similar interpretation can be given to the related notion, copula survival function. We present this for bivariate case:

$$P(X_1 \le x_1, X_2 \le x_2) = C(F_1(x_1), F_2(x_2)), \tag{5}$$

$$P(X_1 > x_1, X_2 > x_2) = \bar{C}(F_1(x_1), F_2(x_2)), \tag{6}$$

$$\bar{C}(u_1, u_2) = 1 - u_1 - u_2 + C(u_1, u_2), \tag{7}$$

where: $\bar{C}$ – copula survival function.

There is infinite number of possible copula functions. One of the most often studied is normal (Gaussian) copula, given by the following formula:

$$C(u_1, \ldots, u_m) = \Phi^m(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m)), \tag{8}$$

where: Φ^m – distribution function of multivariate normal distribution;

Φ – distribution function of univariate normal distribution.

So we can see that multivariate normal distribution can be obtained by "linking" univariate normal distributions through normal copula. This leads to two important statements, the first one being sometimes explored while teaching multivariate statistics:

- If one applies copula function – different than normal copula – to univariate normal distributions, then resulting multivariate distribution is not multivariate normal.
- If one applies normal copula to univariate distributions – different from normal distribution – then resulting multivariate distribution is not multivariate normal.

In the case of bivariate normal distribution, formula (8) can be expressed in the following form:

$$C(u_1,u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) dxdy, \tag{9}$$

where: ρ – correlation coefficient between components of bivariate random vector.

The very well studied family of copula functions is the family of so called Archimedean copulas. We restrict here the presentation to bivariate case. There, Archimedean copulas are defined for strictly decreasing and convex function, called generator (e.g. [Nelsen 1999]):

$$\begin{aligned} &C(u_1,u_2) = \psi^{-1}(\psi(u_1) + \psi(u_2)), \\ &\psi : [0;1] \to [0;\infty), \\ &\psi(1) = 0, \end{aligned} \tag{10}$$

where: ψ – generator.

Among the members of this family of copulas, the most often studied are:

- Gumbel copula, where:

$$\psi(t) = -(\log(t))^\theta, \ \theta \in [1;\infty). \tag{11}$$

- Clayton copula, where:

$$\psi(t) = \begin{cases} (t^{-\theta} - 1)/\theta, & \theta \geq -1, \theta \neq 0 \\ -\log(t), & \theta = 0 \end{cases}. \tag{12}$$

- Frank copula, where:

$$\psi(t)=\begin{cases}-\log\left(\dfrac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right) & \theta\neq 0\\ -\log(t), & \theta=0\end{cases}. \tag{13}$$

- Ali – Mikhail – Haq copula, where:

$$\psi(t)=\log\left(\frac{1-\theta(1-t)}{t}\right),\ \theta\in[-1;1]. \tag{14}$$

What is important for Archimedean copulas, is the fact that (even in multivariate case) they are one parameter functions. This parameter, denoted by θ, can be interpreted as dependence parameter.

Now we review the most important theoretical properties of copula functions, which are useful for analysis of multivariate data.

3. Copula functions and the measures of dependence

As we pointed out in the previous chapter, the parameters of copula functions can be treated as measures of dependence. Moreover, the values of copula function itself can serve as a guideline for the measurement of dependence. This comes from the following three properties:

1. If the variables are independent, then the copula function is given as:

$$C(u_1,\ ...,\ u_m)=C^{\perp}(u_1,\ ...,\ u_m)=u_1\ ...\ u_m. \tag{15}$$

2. The lower limit for copula function is given as:

$$C^{-}(u_1,\ ...,\ u_m)=\max(u_1+\ ...\ +u_m-m+1;0). \tag{16}$$

3. The upper limit for copula function is given as:

$$C^{+}(u_1,\ ...,\ u_m)=\min(u_1,\ ...,\ u_m). \tag{17}$$

Therefore, the copula functions allow comparing the strength and direction of dependence between variables, not being restricted to linear dependence.

It is worth to mention, that for normal copula, the mentioned three situations corresponds to three particular values of correlation coefficient:

- independent case – correlation coefficient equal to 0;
- lower limit – correlation coefficient equal to –1;
- upper limit – correlation coefficient equal to 1.

The next important property is related to the fact that three well known coefficients used to measure the dependence between two variables, can be presented through copula functions. These are:

- Spearman coefficient, known as correlation coefficients between the values of distribution functions, presented as:

$$\rho_S = 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3. \tag{18}$$

- Kendall coefficient, presented as:

$$\rho_T = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1. \tag{19}$$

- Gini coefficient, presented as:

$$\rho_G = 2 \int_0^1 \int_0^1 (|u_1 + u_2 - 1| - |u_1 - u_2| dC(u_1, u_2). \tag{20}$$

This confirms the role of copula functions in the analysis of dependence.

4. Copula functions and tail dependence

In the previous chapter we discussed measures of dependence. In some applications, the important notion is tail dependence. This is the dependence between variables where the values from the tails of the distributions of these variables are taken into account. The most common tool used here is tail dependence coefficient (coefficients).

Tail dependence coefficients are defined for the bivariate case. There are two such coefficients:

- Lower tail dependence coefficient, defined as:

$$\lambda_L = \lim_{u \to 0} P(X_2 \le F_2^{-1}(u) \Big| X_1 \le F_1^{-1}(u)). \tag{21}$$

- Upper tail dependence coefficient, defined as:

$$\lambda_U = \lim_{u \to 1} P(X_2 > F_2^{-1}(u) \Big| X_1 > F_1^{-1}(u)). \tag{22}$$

Lower tail dependence coefficient is defined as limiting probability that one variable takes value from the lower tail, given that the other variable takes value from the lower tail. Upper tail dependence coefficient is defined as limiting probability that one variable takes value from the upper tail, given that the other variable takes value from the upper tail. As one can see, these notions are defined through limiting case, where limit is taken with respect to lower or upper tail.

Both coefficients can take value from the interval [0;1]. There are two possibilities:

- If tail dependence coefficient is equal to 0, then two considered variables are asymptotically independent;
- If tail dependence coefficient is higher than 0, then two considered variables are asymptotically dependent.

These coefficients can be represented through copula functions. It is given in the following formulas:

$$\lambda_L = \lim_{u \to 0}[C(u,u)/u], \tag{23}$$

$$\lambda_U = \lim_{u \to 1}[(1-2u+C(u,u))/(1-u)]. \tag{24}$$

It is worth to mention that normal copula is asymptotically independent, if the correlation coefficient is different from +1. Therefore normal copula is not appropriate for modeling variables which exhibit dependence in tails.

On the other hand, we have for Gumbel copula:

$$\lambda_L = 0, \tag{25}$$

$$\lambda_U = 2 - 2^{1/\theta}, \quad \theta > 1. \tag{26}$$

So Gumbel copula is appropriate when variables are upper tail dependent and lower tail independent.

The presented tail dependence coefficients are defined for the bivariate case. It might be interesting task, however, to consider more general case. Now we give simple proposals to extend tail dependence coefficients to multivariate case.

In the definitions given above, tail dependence coefficient is understood as the conditional probability that one variable takes value from the tail given the other variable takes value from the tail. It is then natural approach to consider the probability that each of many, say, $n-1$, variables take values from the tails, given that the remaining, say, n-th variable, takes value from the tail. Therefore we propose the following definition in trivariate case:

- Lower tail dependence coefficient:

$$\lambda_L = \lim_{u \to 0} P(X_2 \le F_2^{-1}(u), X_3 \le F_3^{-1}(u) \,|\, X_1 \le F_1^{-1}(u)). \tag{27}$$

- Upper tail dependence coefficient:

$$\lambda_U = \lim_{u \to 1} P(X_2 > F_2^{-1}(u), X_3 > F_3^{-1}(u) \,|\, X_1 > F_1^{-1}(u)). \tag{28}$$

It can be proved that these tail dependence coefficients can be represented through copula functions and copula survival functions, using the following formulas:

$$\lambda_L = \lim_{u \to 0}[C(u,u,u)/u], \tag{29}$$

$$\lambda_U = \lim_{u \to 1}[\bar{C}(u,u,u)/(1-u)]. \tag{30}$$

The generalization of this approach to higher number of dimensions is straightforward.

5. Copula functions and extreme values

Copula functions can also be applied in multivariate extreme value analysis. This analysis can be performed by studying the distribution function for a vector of maxima taken componentwise:

$$F_{n:n}(\mathbf{x}) = P(X_{1,n:n} \le x_1, \ ..., \ X_{m,n:n} \le x_m). \tag{31}$$

If one analyzes the limiting distribution, by allowing n to go to infinity, then we get the multivariate generalization of the Fisher-Tippett theorem, derived for univariate case. Here the main result is given for the limiting distribution of normalized maxima, given as:

$$\lim_{n \to \infty} P\left(\frac{X_{1,n:n} - b_{1,n}}{a_{1,n}} \le x_1, \ ..., \ \frac{X_{m,n:n} - b_{m,n}}{a_{m,n}} \le x_m \right) = G(x_1, \ ..., \ x_m) = G(\mathbf{x}). \tag{32}$$

Here, limiting distribution is a so called, Multivariate Extreme Value Distribution, being the generalization of univariate extreme value distribution. As it is well known, that the family of univariate extreme value distributions contains just three distributions: Gumbel, Fréchet and Weibull distribution.

Unfortunately, in multivariate case there is no parametric representation of this family. However, there are two properties of Multivariate Extreme Value Distribution, which are useful in practice:

1. Marginal distributions of Multivariate Extreme Value Distribution are univariate extreme value distributions (Gumbel, Fréchet or Weibull distribution).

2. In the copula representation of Multivariate Extreme Value Distribution, copula function is a so called Extreme Value Copula. It satisfies the following relationship:

$$C(u_1^t, \ ..., \ u_m^t) = C^t(u_1, \ ..., \ u_m). \tag{33}$$

Therefore, to construct Multivariate Extreme Value Distribution one has to apply Extreme Value Copula to marginal distributions being univariate extreme value distributions.

The most often studies copula functions belonging to Extreme Value Copula family are (for simplicity we restrict ourselves to bivariate case):

– Gumbel copula, given as:

$$\begin{aligned} &C(u_1,u_2) = \exp[-(\log u_1^{\ \theta} + \log u_2^{\ \theta})^{1/\theta}] \\ &\theta \in [1;\infty) \end{aligned} \tag{34}$$

– Gumbel II copula, given as:

$$C(u_1,u_2)=u_1u_2\exp[\theta(\log u_1\log u_2)/(\log u_1+\log u_2)]$$
$$\theta\in[0;1] \tag{35}$$

– Galambos copula, given as:

$$C(u_1,u_2)=u_1u_2\exp[((\log u_1)^{-\theta}+(\log u_2)^{-\theta})^{-1/\theta}]$$
$$\theta\in[0;\infty) \tag{36}$$

These results simplify the analysis of multivariate extreme values.

6. Copula functions in time series analysis

The main idea behind the use of copula in modeling relations in univariate time series comes from the important results obtained by Darsow, Nguyen and Olsen [1992].

There are two important and often discussed copula functions for time series: Brownian copula and Ornstein-Uhlenbeck copula. They correspond to the two continuous time stochastic processes, known under the same names.

1. Brownian copula. It is given as:

$$C_{s,t}(u_1,u_2)=\int_0^{u_1}\left(\Phi\left(\frac{\sqrt{t}\Phi^{-1}(u_2)-\sqrt{s}\Phi^{-1}(u)}{\sqrt{t-s}}\right)\right)du. \tag{37}$$

The most important properties of Brownian copula are:

– Brownian copula is normal copula with parameter:

$$\rho=\sqrt{t-s}. \tag{38}$$

– If the marginal distributions are normal distributions, then applying Brownian copula leads to the stochastic process, which is geometric Brownian motion.

2. Ornstein-Uhlenbeck copula.

It is given as:

$$C_{s,t}(u_1,u_2)=\int_0^{u_1}\left(\Phi\left(\frac{h(0,s,t)\Phi^{-1}(u_2)-h(0,s,s)\Phi^{-1}(u)}{h(s,s,t)}\right)\right)du. \tag{39}$$

$$h(t0,s,t)=\sqrt{e^{2a(t-s)}-e^{2a(s-t0)}}. \tag{40}$$

The most important properties of Ornstein-Uhlenbeck copula are:

– Ornstein-Uhlenbeck copula is normal copula with parameter:

$$\rho = e^{-a(t-s)}\sqrt{\frac{1-e^{-2as}}{1-e^{-2at}}}. \tag{41}$$

- If the marginal distributions are normal distributions, then applying Ornstein-Uhlenbeck copula leads to the stochastic process, which is Ornstein-Uhlenbeck process.
- The parameter a – being also the mean-reverting coefficient of Ornstein-Uhlenbeck process, known as speed of reversion – can be interpreted as the parameter of the dependence between random variables being the components of stochastic process – the larger this coefficient, the less dependence between random variables.

References

Darsow W.F., Nguyen B., Olsen E.T. (1992), *Copulas and Markov Processes*, „Illinois Journal of Mathematics", 36, p. 600-642.

Nelsen J. (1999), *Introduction to Copulas*, Springer, New York.

Sklar A. (1959), *Fonctions de repartition à n dimensions et leurs marges*, Publications de l'Institut de Statistique de l'Université de Paris, 8, p. 229-231.

PARAMETRY STATYSTYCZNE MAJĄCE U PODSTAW FUNKCJE POŁĄCZEŃ

Streszczenie

Artykuł przedstawia zastosowanie funkcji kopuli (połączeń) w określeniu parametrów zbioru danych wielowymiarowych. Na wstępie przedstawiono podstawy analizy połączeń, a następnie omówiono takie grupy parametrów, jak: miary zależności, współczynniki zależności w ogonie, parametry wielowymiarowych wartości ekstremalnych, miary zależności w szeregach czasowych.