Journal of Kerbala University , Vol. 12 No.4 Scientific . 2014

Copula Functions and correlations Personal Contribution

دوال الكوبله ومعاملات الارتباط مساهمه شخصيه

Ahmed Al-Adilee
Department of Computer Science
Faculty of Science
Kerbala University
07816885554
ahmeda.aladilee@uokufa.edu.iq

Karrar Dheiaa. Mohammed
Department of Mathematics
Faculty of math.a CS
University of Kufa
07816048659
karrard.alsabti@uokufa.edu.iq

Abstract

This study concerned with comparing correlation coefficient (correlation) to copula functions (copulas) throughout their correlations, which are defined by means of their copulas . Indeed, it is an attempt to answer the following question, why copula functions (copula) is preferred in modern statistical inference to correlation?. We will discuss each topic individually, to make a right decision about the question above. For making our decision clear we have recalled some of the basic concepts related to correlation and copulas so that we can show the characteristics and differences of using each one.

المستخلص

هذه الدراسه تهتم بمقارنة معامل الارتباط مع دوال الكوبله من خلال العلاقات المتعلقه بهذه الدوال والمعرفه بدلالتها. في الحقيقه, انها محاولة لاجابة السؤال التالي, لماذا دوال الكوبله في الاستدلال الاحصائي الحديث تكون مفضله على معامل الارتباط؟. سنناقش كل موضوع بشكل مستقل لغرض الحصول على جواب مضبوط عن السؤال اعلاه. لجعل قرارنا واضح قمنا باستدعاء بعض المفاهيم الاساسيه المتعلقه بالارتباط ودوال الكوبله لكي نستطيع تبين الخواص والاختلافات في استخدام كل واحد.

1. Introduction

There are two main parts that this study deal with. First part is the correlation topic, while the second one is known by copula functions topic. Statistical inference is one of the most classical, and important approaches that used to studying those topics. It is widely used in various fields of scientific studies, for example, mechanical engineering, floods, biomedical studies, risk management, and other interesting fields. This means that these topics are the tools that we employed in statistical inference.

Whatever, each random experiment needs statistical inference means that we need to interpreted and describe relationships among random variables related to that random experiment. This can be done by correlation or what is well-known as correlation [5]. Indeed, correlation represents a dependence measure between two or multi variables. Each dependence structure consist of random variables that we need to measure their dependency. Precisely, this exactly means that correlation can evaluate the strength and direction of two random variables. What it is necessary to be mentioned in here that correlation only represents a linear relationship among random variables [1].

By going to talk about the beginning of correlation, so we are talking not only about correlation, but also about regression concepts. This fact follows from the strong relationship between correlation, and linear regression approach. From a historical point of view, we could say that first notion of correlation was introduced by Galton in 1885. He defined, and derived most of the regression equations and mentioned what is called the bivariate correlation. After that this topic was developed and modified by Pearson in 1895. He found what is nowadays called Pearson's correlation and denoted by ρ.

While the historical description about copula is not so old. We begin with a brief historical review. The word copula borrowed from the Latin language, and it means the "link", or "join". Indeed, copula is a powerful tool and play an important role in modern statistics. Implicitly, copulas were mentioned in several statistical studies without referring to it by name copula, for example, Hoeffding (1940, 1941). In 1959 Sklar was the first author who use the terminology copula, and involves it in his literatures.

The central role of the well-known theorem of Sklar in 1954 represents the rigid foundation that all the structure, and concepts of copulas lied on. There are several authors, who studied copula functions, and demonstrated it after Sklar, but the most important one, who left a huge impact in the literatures about copula was Nelsen in (1993). He builds and collects most of copula concepts, definitions, theorems, properties, applications and examples in his very well-known book entitled "an introduction to copulas" in (1998). Honestly, we should refer to some other names, who have a huge impact on the studies of copulas like Sampson (1975), Galambos (1978), Schweizer (1991), Marshall (1988), Olkin (1988), and Joe (1993,1997).

Eventually, we should say that in order to present a sufficient decision about the comparison of correlation approach to copula approach, and answer our above question, we need to set some basic concepts about both approaches. We suppose that reader is familiar with some basic statistical concepts such as random variables, distribution functions, mathematical expectation, marginal distribution functions, copulas, and other statistical concepts.

2. Correlation and Linear Relationships.

We are obliged to present some basic concepts related to correlation and linear relationships so that we could find out the points of weakness and strength of that topic. We begin with the definition of correlation.

Definition 1:- A correlation denoted by ρ of two random variables X, Y is the linear relationship that measure both strength and direction between those random variables and defined by the following formula.

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \qquad (1)$$

where, $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$, and provided that $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$.

Indeed, the correlationρ is well-known as Pearson's correlation. According to definition 1 we are also able to define the sample correlation as follows [5]:

Definition 2:- Let the order pairs$(X_i, Y_i), i = 1, \dots, n$ be a random sample from a bivariate normal distribution. Then the correlation denoted by r is

$$r = \frac{Cov(X_i,Y_i)}{\sqrt{Var(X_i)Var(Y_i)}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \qquad (2)$$

where$\bar{X}, \bar{Y}$are the sample mean of $X_i, Y_i, i = 1,2, \dots, n$, respectively.

In general, there are three main notions that we could figure out from the above definitions. First notion that correlation is in fact a scale to measure the strength and direction between two or multi variables of their linear relationship. The second notion is that correlation is useful when the random variables are normally distributed. The last notion is that correlation of random variables is a measure of them only when their margins are elliptically distributed.

Geometrically, we can also interpreted the role of correlation under regression by the following graphs shown in Fig.1. We see in the graphs below how the elliptical margins convert to linear relation between the two random variables X, and Y.

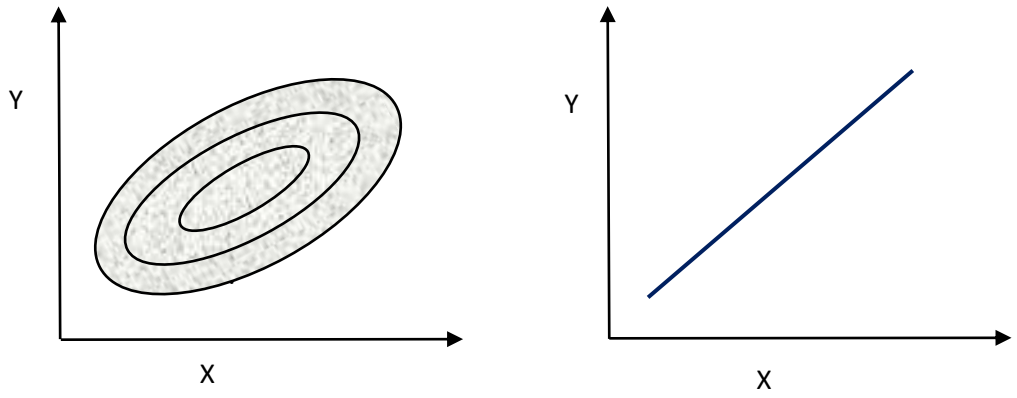

Fig.1. Elliptical margins with perfect correlation.

Moreover, we could list the following important properties of correlation ρ as follows

- $-1 \leq \rho \leq 1$;
- When $\rho = 0$, there is no information about the correlation between the random variables;
- $0 < \rho \leq 1$ is called perfect positively correlation;
- $-1 \leq \rho < 0$ is called perfect negatively correlation;
- If X, Y are two random variables, and $(X - \mu_1) < 0$, and$(Y - \mu_2) < 0$, then $\rho > 0$; whereμ_1, μ_2 are the mean of X, Y, respectively;
- If X, Y are two random variables, and $(X - \mu_1) < 0$, and$(Y - \mu_2) > 0$, then $\rho < 0$;

Note that, when $\rho = 0$ this does not mean we have independent random variable. In other words, two or multi-random variables are not independent even the correlation between them is equal to zero. From the last two properties above, we should know that $\mathrm{Cov}(X, Y)$ is positive when $(X - \mu_1) < 0$, and$(Y - \mu_2) < 0$, while $\mathrm{Cov}(X, Y)$is negative when$(X - \mu_1) < 0$, and$(Y - \mu_2) > 0$.

Furthermore, looking again at the properties of correlation above lead us to conclude that correlation with numbers 1, or -1 is called the perfect linear relationship between two or multi random variables. Counting on this fact of correlation we are able to compose the following proposition:

Proposition1:- Let X, Y be two random variables. The correlation ρis ± 1, if and only if, X, and Y are linearly correlated.

As we have mentioned before, we note that ρ is a perfect tool when the random variables are under the elliptical world, and give a good description about the dependence structure between the relevant random variables. But what, if we are investigating the dependence structure between two or multi random variables beyond the elliptical world, or studying random variables that are not normally distributed?. Indeed, the answer is that correlation is simply failed to be founded, or may have no statistical meaning. On the other hand, we can again see that correlation is not sufficient if the marginal distribution functions of the random variables are not inside the elliptical representation, or not normally distributed.

Here, we could list some disadvantages of the correlation as a measure of the linear relationship of random variables:

1. Invariant correlation under strictly transformation of some random variables does not exist. For instance, when the correlation of the two random variables X, and Y exists, so the correlation under $logX$, and $logY$is not the same correlation of the original random variables.

2. Correlation cannot tell us all the information we need about the whole dependence structure. It is only describe the linear relationship between random variables without telling us any further information about other properties of the dependence structure.
3. Correlation cannot be defined, and used as a measure tool when the variances are approximated to infinity.

Furthermore, it is well-known fact that correlation and regression have an important relationship, but they are not similar. Regression or linear regression used to fit data in an linear equation, while correlation follow after regression in order to measure and find any dependence relationship between random variables related to that linear equation. This means correlation is used to determine how good is the linear regression equation of the fitted data.

3. Copula Functions (Copulas)

Now, we turn to our second part, which is called copulas. Copulas are a powerful statistical tool and modern phenomenon in the world of statistics. Indeed, copulas are a more general approach than correlation, and have a very rigid theory. First of all, we should recall some basic concepts of copulas so that we can compare them to correlation. We will only use the notion of what is known by the bivariate copula.

Definition 4:- Let $B = [X_1, X_2] \times [Y_1, Y_2]$ be a rectangle region whose vertices are the end points of intervals$[X_1, X_2], [Y_1, Y_2]$, respectively. Then H-volume, denoted by$V_H(B)$ over the rectangle region B is given by

$$V_H(B) = \Delta_{Y_1}^{Y_2}\Delta_{X_1}^{X_2}H(X,Y) \geq 0 \qquad (4)$$

According to equation (4), H is called a 2-increasing property. This definition is very necessary to define copula.

Definition 5 [4]:- Let $I = [0,1]$. A bivariate copula, denoted by C is a function $C: I^2 \to I$ with the following properties

1. For every $u, v \in I, C(u,0) = C(0,v) = 0$ (grounded);
2. For every $u, v \in I, C(u,1) = u$, and $C(1,v) = v$, where 1 is the greatest element;
3. Let $u_1 \leq u_2, v_1 \leq v_2$,where $u_1, u_2, v_1, v_2 \in I$. Then

$$C(u_1, v_1) + C\big(u_2, v_2\big) \geq C(u_1, v_2) + C(u_2, v_1) \qquad \text{(2-increasing).}$$

Note that copulas are not exist for less than two dimensional function. Also, each function satisfies the three properties above is called copula, otherwise it is not a copula. Copula is a modern approach relevant to correlation. It could be employed to describe everything we need about the dependence structure. The most important benefit of using copula in application follows from its flexibility as a scale used to measure the relationship between variables, and because it has no strict limits. Indeed, copula was firstly defined to join or couple the joint distribution function to its marginal distribution functions. In other words, we describe the joint distribution by means of its marginal distribution functions. This assertion can be shown by the central theorem of Sklar in (1954).

(Sklar'sTheorm):- Let F be a joint distribution function of the two random variables X, Y, and let $F_1(x) = u, F_2(y) = v$ be the marginal distribution functions of F. Then there exist a copula C such that $\forall\, x = X, y = Y \in [-\infty, \infty]$, respectively

$$F(x,y) = C\big(F_1(x), F_2(y)\big) = C(u,v) \qquad (5)$$

and, if$F_1, F_2$are continuous margins then the copula C is unique. This theorem can also be restated by means of copula and using the converse direction [3].
One of the result of Sklar's theorem can be shown by the following proposition

Proposition 1:- LetF_1^{-1}, F_2^{-1}be the inverse of F_1, F_2, respectively, such that $F_1^{-1}(F_1(x) = u) = x, F_2^{-1}(F_2(y) = v) = y$. Then

$$C(u,v) = F(F_1^{-1}(u), F_2^{-1}) \quad (6)$$

There are many types of copulas that we could refer to them. In particular, let us list the following three well-known types.

i. The product copula.

This type of copula is only exist when random variables are independent and denoted by$\prod$. Its formula is given by

$$\prod(u,v) = u.v \quad (7)$$

One of the important notion related to product copula can be shown by the following proposition.

Proposition 2:- Two random variables are independent if and only if their copula is the product copula ($\prod(u, v) = u. v$).

ii. The Fréchet-Hoeffding bounds

Frechet-Hoeffding bounds are knowing by a Frechet-Hoeffding upper, and lower bound. For any two random variables there is a copula C with an upper bound copula denoted by $W(u,v)$, lower bound copula denoted by $M(u,v)$, and satisfies the following property

$$W(u,v) \leq C(u,v) \leq M(u,v) \quad (8)$$

where $W(u,v) = max(u+v-1,0)$, and $M(u,v) = min(u,v)$.

From the Fréchet-Hoffding bound definition, we can see that $F(x,y)$is also lied between the upper, and lower bounds. For survival copula there is a very important notion associated with conditional distributions and using it to generate various types of random variables[4]. There are more details about most of the copula types in [3].

iii. The Survival Copula

One of the most important types of copulas is called the survival copula. For testing the lifetime of components we can see that survival copula is a good tool used for analyzing and test the surviving of item with respect to time t. We firstly need to express the meaning of survival distribution function, then use its formula to define survival copula. If $F(x,y)$is the distribution function of the random variables X, Y, with its marginal distribution functions $F_1(x), F_2(y)$. Then the joint survival function of these two random variables is

$$G(x,y) = 1 - F_1(x) - F_2(y) + F(x,y) \quad (9)$$

According to equation (9), the survival copula, denoted by $\hat{C}$is given by

$$\hat{C}(u,v) = u + v + C(1-u, 1-v) - 1 = G(x,y) \quad (10)$$

We again emphasize that each type of copula involving the three above should satisfy the essential properties of being copula.

As we have mentioned before, correlation is only available under the elliptical world as a scale to measure the dependence structure, but it will fail when the margins are beyond that. By comparing correlation concepts to copula concepts we can see that copulas avoid such problems because of their flexible forms that allow them to measure the dependence structure whether margins are normally distributed or not.

Furthermore, there is another important concept relevant to copulas. This concept is called the families of copulas. In fact, each copula should belong to one of the copula families. Indeed, there are many of them that we will only refer to some interesting of them according to their margins.

3.1 Archimedean Family of copula

This family of copula classified as a non-elliptical family (not normally distributed), and it can be divided to two types depending on its parameters. We can show the family of each copula with its generator function by the following two tables.

Family of copula	$C_\theta(u,v)$	$\varphi_\theta(t)$
Ali-Mikhail-Haq	$\frac{uv}{1-\theta(1-u)(1-v)}$	$ln\frac{1-\theta(1-t)}{t}$
Gumbel	$Exp\left(-\left[(-ln\,u)^\theta+(-\ln v)^\theta\right]^{\frac{1}{\theta}}\right)$	$(-lnt)^\theta$
Gumbel-Hougaard	$uve^{(-\theta\ln u\ln v)}$	$ln(1-\theta\ln t)$

Table 1.Archimedean families with one parameter

Family of copula	$C_\theta(u,v)$	$\varphi_\theta(t),\alpha=1$
Ali-Mikhail-Haq	$C_{\theta,\alpha}(u,v)=\frac{uv}{\left[1-\theta\left(1-u^{\frac{1}{\alpha}}\right)\left(1-v^{\frac{1}{\alpha}}\right)\right]^\alpha}$	$\ln\left(\frac{[1-\theta(1-t)]}{t}\right)$

Table 2.Archimedean families with two parameters

3.2 The Bivariate Normal Family

This type of families is also called the Gaussian copula (normally distributed), and it is belong to distributions, that are elliptically distributed. In fact, it is defined under distributions, that has a symmetrical tail. Indeed, Gaussian family has only one specific formula that is defined by, [4]

$$\mathrm{C(u,v)}=\Phi_\rho\left(\Phi^{-1}(\mathrm{u}),\Phi^{-1}(\mathrm{v})\right)=\frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\Phi^{-1}(\mathrm{u})}\int_{-\infty}^{\Phi^{-1}(\mathrm{v})}\mathrm{e}^{\left[\frac{-(x^2-2\rho xy+y^2)}{2(1-\rho^2)}\right]}\,\mathrm{dxdy}$$

where, $-1\le\rho\le1$, andΦ_ρ is the standard two-dimensional normal distribution with $\mu_1,\mu_2=0,\sigma_1,\sigma_2=1$.

Note that ρ is a correlation, and the parameter of dependency between the two random variables X, and Y.Also, one of the most important applications of using Guassian copula is risk management and financial analysis of assets, and portfolio.

Now, we are in position to present a brief discussion about new types of correlations relevant to copulas. These types are known by Kendall's tau and Spearman's rho.

4. Kendall's Tau and Spearman's Rho [4]

Based on the concepts of copulas and their properties, there are two modern correlations that can be defined by the following ways

Definition 6:- Let X, and Y be two absolutely continuous random variables with their copula C, and concordance function Q ($Q = P[X_1Y_1 + X_2Y_2 > X_1Y_2 + X_2Y_1] - P[X_1Y_1 + X_2Y_2 < X_1Y_2 + X_2Y_1]$). Then there exists a correlation called Kendall's tau, denoted by τ_C, and defined by the following form

$$\tau_C = Q(X,Y) = 4E(C) - 1 = 4\int_0^1\int_0^1 C(u,v)dC - 1 \qquad (11)$$

where $u = F_1(x), v = F_2(y)$, and $dC = \frac{\partial^2 C}{\partial u\,\partial v}dudv$.

Definition 7:-Let X, and Y be two absolutely continuous random variables with their copula C, and concordance function Q. Then there exists a correlation called Spearman's rho, denoted byρ_C, and defined by the following form

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

One of the important notion about the correlations τ_C, ρ_Cthat we should refer to that those two types of correlations might be failed to be evaluated when their copula is singular [4]. This led to rewrite the formula of equation (11)by means of generator function [4]. That is

$$\tau_C = 1 + 4\int_0^1 W(t)dt, \qquad (13)$$

where $W(t) = \frac{\varphi(t)}{\dot{\varphi}(t)}$, and $\varphi(t)$is the generator function of copula families.Now, let us illustrate the following example to show how this generator can be applied to determine, for example, the Kendall's tau.

Example 1:-Recall Ali-Mikhail-Haq family with one parameter that is given by

$$C(u,v) = \frac{uv}{1-\theta(1-u)(1-v)} \qquad (14)$$

In order to determine the Kendall's tau τ_C, we recall the generator of Ali-Mikhail-Haq family, that is given by,

$$\varphi(t) = \ln\big(1-\theta(1-t)\big) - \ln t \qquad (15)$$

Differentiate $\varphi(t)$ with respect to t, we obtain that

$$\dot{\varphi}(t) = \frac{-(1-\theta)}{t[1-\theta(1-t)]} \qquad |(16)$$

From equation (15), and (16), it is easy to show that W can be written by the following formula

$$W(t) = \frac{-\ln\frac{1-\theta(1-t)}{t}\left[t\left(1-\theta(1-t)\right)\right]}{1-\theta} \tag{17}$$

Now use the formula of Kendall's tau in equation (13) with respect to the above $W(t)$, obtain that

$$\tau_C = 1 + 4\int_0^1 \frac{-\ln\frac{1-\theta(1-t)}{t}\left[t\left(1-\theta(1-t)\right)\right]}{1-\theta}\,dt \tag{18}$$

We can determine the integral in equation (18) using by part method, and this will lead us to the following result

$$\tau_C = 1 + 4\left[\frac{-1}{2(1-\theta)} - \frac{1-\theta}{2}\left(\frac{1}{\theta} + \frac{1-\theta}{\theta^2}\ln(1-\theta)\right) - \frac{\theta}{3}\left(\frac{1}{2\theta} - \frac{1-\theta}{\theta^2} - \frac{(1-\theta)^2}{\theta^3}\ln(1-\theta)\right)\right]$$

After some arrangements of the right hand side of the equation above, we obtain that Kendall's tau correlation for Ali-Mikhail-Haq is

$$\tau_C = \frac{3\theta-2}{3\theta} - \frac{2(1-\theta)^2}{3\theta^2}\ln(1-\theta) \tag{19}$$

Also, we can show an example of illustrating the Spearman's rho as follows

Example 2:- Let C be a copula such that $C(u,v) = uv^{1-\theta}$, where $\theta \in [0,1]$. Then $\rho_C$of this copula can be determined by using equation (12)

$$\rho_C = 12\int_0^1\int_0^1 uv^{1-\theta}\,dudv - 3 \tag{20}$$

By evaluating the right hand side of equation (20), which involve double integral. Then we obtain the following value of the Spearman's rho

$$\rho_C = \frac{3\theta}{2-\theta} \tag{21}$$

Moreover, it is a well-known fact that copulas has no restrictions and that may lead to problems in the evaluations of correlations, but this can be solved it by evaluating the correlations with respect to the restrictions of the parameters. In other words, the restrictions of copula family is defined by the range of the parameters of that family.

5. Correlation (Pearson) Against Kendall's Tau and Spearman's Rho

This part devoted to introduce a summary, and a comparison between the ordinary correlation known by Pearson(ρ)from one side and the modern correlations that we have mentioned in the previous section, which are called Kendall's tau, and Spearman's rho from the another. According to these concepts that we have investigated our comparison can be summarized by the following table

ρ	τ_C,ρ_C
Its formula is $\rho = \frac{Cov\,(X,Y)}{\sqrt{Var\,X\,Var\,Y}}$	Their formulas are $\tau_C = 4E(C) - 1, \rho_C = 12\iint_{I^2} CdC - 3$
it is a measure of the strength and direction of the linear relationship between random variables.	They are a measure of the dependency and interrelated of random variables whether they have a linear or non-linear relationship.
It can only tell us information about the Relationship between random variables, but not the whole dependence structure.	They describe the whole dependence structure and Tell us everything we need about the relationship between the random variables.
Its values always between -1, and 1.	Their values depend on the range of the parameters under copula, for example, if θbetween -1, 1, and the value of Kendall's tau is $\frac{2\theta}{9}$, then $\tau_C \in \left[\frac{-2}{9}, \frac{2}{9}\right]$.
It is only defined for random variables, whom are normally distributed, and it will have problems and difficulties to evaluated when random variables are not under the elliptical world.	They can be determined whether random variables are normally distributed, or not, and even when the marginal distribution functions are not elliptically distributed.
It is a special case of copula.	It is a more general construction than correlation.
Correlation ρ is only associated with linear regression that used to estimate parameter and then fit data in linear equation.	These correlation are relevant to whether linear or non-linear regression and has more complicated estimation forms.
Correlation is not invariant under transformation of random variables with respect to some special function like logarithmic functions.	They are invariant correlations, whatever, the type of transformation of the random variables.

Table 3. Comparison between correlation and the new τ_C, ρ_C types of correlations

Note that the comparison above can be extended to involve the study of correlations with n-dimensional random variables.

6. Conclusion

This study leads us to conclude that copulas are a much better, and more flexible structure than the ordinary correlation. Also, they have no restrictions which leads to better calculations. Copulas are a more general forms than correlation. Correlation is always symmetric because it is a scale measure with respect to symmetric tail, while Kendall's tau, or spearman's rho not need to be symmetric. Correlation is an old approach and insufficient to describe the whole dependence structure of random variables, while copula are a modern approach with rigid structure and can describe not only the relationship between random variable, but even any other properties relevant to the whole structure. Finally, we should say that copulas still a fresh approach with many open problems that need to be investigated.

References

[1]Habiboellah, F. Copulas, modeling dependencies in Financial Risk Management, BMI,university of Amesterdam, (2007).

[2] Klement, E. P.,Mesiar, R.,and Pap, E. Archimax copulas and invariance under transformations. C.R. Math. Acad. Sci. Paris *340*, 755-758, (2005).

[3]Nelsen, R. B. On measures of association as measures of positive dependence. Statist. Probab. Lett. 14, 269-274, (1992).

[4]Nelsen, R.B. An introduction to copulas, Springer-New York, (2006).

[5]Ramachandran, K.M., Tsokos, C.P. Mathematical statistics with applications, Elsevier-California, (2009).