Research Paper

Volume : 2 | Issue : 12 | December 2013 • ISSN No 2277 - 8179

Copulas: As a Measure of Dependence

Statistics

KEYWORDS : Copula, joint distribution, marginal distribution, conditional distributions, correlation.

Jugal Gogoi — Department of Mathematics, Dibrugarh University, Dibrugarh-786004, Assam

Navajyoti Tamuli — Department of Statistics, Gargaon College, Simaluguri-785686. Assam

ABSTRACT

The study of the relationship between two or more random variables remains an important problem in Statistical Science. Copulas provide a convenient way to express joint distributions of two or more random variables. With a copula we can separate the joint distribution into two contributions: the marginal distributions of each variable by itself, and the copula that combines these into a joint distribution. One basic result is that any joint distribution can be expressed in this manner. Another convenience is that the conditional distributions can be readily ex-pressed using the copula. Copulas are a useful tool for understanding relationships among multivariate variables, and are important tools for describing the dependence structure between random variables, with different copulas representing different dependencies. Since the study of dependence is critically important in Statistics in order to carry out reliable analyses, understanding and applying results from copulas can be very beneficial. The main object of this paper is that the advantages of using a copula over correlation were to model dependencies between two or more variables.

Introduction:

The study of the relationship between two or more random variables remains an important problem in statistical science. For example, when two lives are subject to failure, such as under a joint life insurance or annuity policy, actuaries generally concerned with joint distribution of lifetimes. One another example, when we simulate the distribution of a scenario that arises out of a financial security system, we need to understand the distribution of several variables interacting simultaneously, not individually. Suppose we have two dependent random variables X & Y and we know the marginal distributions of them and also we know the Pearson's linear correlation coefficient of X & Y. now the question is "Do we have enough information to describe the joint behavior of X & Y?" the answer is no (although we can fit bivariate normal distribution if marginal's of both X & Y are normal). Since most of actuarial or other real life data applications of normal distribution does not provide an adequate approximation. A suitable way to express joint distributions of two or more random variable is provided by copula. Actually, a copula separates the joint distribution into two aids: the marginal distribution if the individual variables and the interdependency of the probabilities. In the present study, it is attempt to derive bivarate copula over correlation coefficient and show how it can be used in practical situation.

The history of copulas may be said to begin with ([1], [2], [3]) Fréchet (1951). Fréchet's problem: given the distribution functions $F_j(j = 1, 2, \dots, d)$ of d r.v.'s $X_1, X_2, \dots, X_d$ defined on the same probability space (Ω, F, P), what can be said about the set $F(F_1, F_2, \dots, F_d)$ of the d–dimensional d.f.'s whose marginals are the given $F_j(j = 1, 2, \dots, d)$. $H \in F(F_1, \dots, F_d) \Leftrightarrow H(+\infty, \dots, +\infty, t, +\infty, \dots, +\infty) = F_j(t)$ The set $F(F_1, F_2, \dots, F_d)$ is called the Fréchet class of the F_j's. Notice $(F_1, F_2, \dots, F_d) \neq \emptyset$; since, if $X_1, X_2, \dots, X_d$ are independent, then $H(X_1, X_2, \dots, X_d) = \prod_{j=1}^{d} F_j(x_j)$ But, it was not clear which the other elements of $F(F_1, F_2, \dots, F_d)$ were. In 1959, [4] Sklar obtained the most important result in this respect, by introducing the notion, and the name, of a copula, and proving the theorem that now bears his name. At end of the nineties, the notion of copulas became increasingly popular. Two books about copulas appeared and were to become the standard references for the following decade. In 1997 Joe published his book on multivariate models, with a great part devoted to copulas and families of copulas. In 1999 Nelsen published the first edition of his introduction to copulas (reprinted with some new results in 2006). But, the main reason of this increased interest has to be found in the discovery of the notion of copulas by researchers in several applied field, like finance.

Dependence measure and its properties:

If X and Y are any random variables with joint distribution function H and their marginal's are F and G respectively. Then we say that X and Y are dependent if

$$H(x, y) - F(x)G(y) \neq 0 \qquad \forall\, x, y \in R$$

Linear correlation or simply correlation (r) between X and Y is only one particular measure of stochastic dependence among many dependence measures. When variances of X and Y are not finite the linear correlation is not defined. Independence of two random variables implies they are uncorrelated (linear

correlation equal to zero) but zero correlation does not imply independence. It is not invariant under non-linear strictly increasing transformation.

Karl Pearson correlation coefficient assumes that the parent population from which sample observations are drawn is normally distributed. If the assumption is violated, then we need a measure, which is distribution free (or non parametric).

Nonparametric correlations often used are the Spearman's rank correlation ρ and Kendall's rank correlation τ. Let $(X_1, Y_1), (X_2, Y_2) and\ (X_3, Y_3)$ are three independent random vectors with a common joint distribution function H. Let us consider the vector $(X_1, Y_1)\ and (X_2, Y_2)$. Then the Spearman's rank correlation ρ associated to a pair (X, Y), distributed ascending to H, is defined as

$$\rho_{xy} = 3P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\}$$

Kendal's rank correlation can be defined as the difference between the probabilities of concordance and discordance for two independent pairs $(X_1, Y_1)\ and (X_2, Y_2)$ each with joint distribution H i.e.,

$$\tau_{XY} = P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\}$$

Both ρ and τ measure the degrees of monotonic dependence X and , whereas linear correlation r measures the degrees of linear dependence only.

The dependence measure should possess the following desired properties:

1. Symmetry: $d(X, Y) = d(Y, X)$ and
2. Normalization: $-1 \leq d(X, Y) \leq +1$

Where $d(\dots, \dots)$ is a dependence measure which assigns a real number to any pair of real valued random variables X and Y.

Both ρ and τ fulfill above two properties and also they fulfill following properties:

3. (i)$d(X, Y) = +1 \iff X, Y$ co-monotonic
   (ii)$d(X, Y) = -1 \iff X, Y$ counter-monotonic
4. For a transformation $T: R \to R$ strictly monotonic on the range of X.
   (i)$d(T(X), Y) = d(X, Y)$, if T is increasing
   (ii)$d(T(X), Y) = -d(X, Y)$, if T is decreasing
5. Another property we must desire is
   $d(X, Y) = 0 \iff X, Y$ are independent

Unfortunately, this contradicts property 4. There is no dependence measure satisfying property 4 and property 5. If we require property 5, then we can consider dependence measures with amended properties.

P1. $0 \leq d(X, Y) \leq 1$
P2.$d(X, Y) = 1 \iff X, Y$ Co-monotonic
P3. For a $T: R \to R$ strictly monotonic $d(T(X), Y) = d(X, Y)$

A measure which satisfies all the above five property (with the exception of the implication $d(X, Y) = 1 \iff X, Y$ co-monotonic) is monotone correlation.

$$d(X, Y) = \underbrace{sup}_{f,g} r[f(X), g(Y)]$$

Where r represents linear correlation and the supreme is taken over all monotonic functions f and g such that $0 \leq \sigma_x^2, \sigma_y^2 < \infty$ (Kimeldorf and Sampson 1989) the disadvantage of all these measures is that they are constrained to give non-negative values and as such cannot differentiate between positive and negative dependence. It is often not clear how to estimate them. An overview of dependence measures and their statistical estimation is given by Tjostheim (1996). Scoweizer and

Volume : 2 | Issue : 12 | December 2013 • ISSN No 2277 - 8179

Wolff (1981) used distance criterion for measuring dependence and proposed

$$d_1(X,Y) = 12 \int_0^1 \int |C(u,v) - uv|\, dudv$$

$$d_2(X,Y) = (90 \int_0^1 \int |C(u,v) - uv|^2 dudv)^{1/2}$$

$$d_3(X,Y) = 4 \underbrace{sup}_{u,v\in[0,1]} |C(u,v) - uv|$$

Where $C(u,v)$ is the joint distribution function of $F(X)$ and $G(Y)$ called copula of random variables X and Y or bivariate distribution $H(X,Y)$.

Copula Function

The term copula originates from Latin, and means "a link, tie, bond" and is referred to joining together. The phrase 'copula' was first used in 1959 by Abe Sklar fifty-two years ago, but traces of copula theory can be found in Hoeffding's work during 1940's [5]. The mathematical meaning of the term copula, namely the copula function is defined as- *"A copula function links n univariate marginal distributions to a full multivariate distribution resulting in a joint distribution function of n standard uniform random variables"*. The copula function actually 'couples' the marginal distributions together to form a joint distribution. Assume that for two random variables *(X, Y),* the standard formulation is: *H (x, y) = C (F (x), G (y)),* where *C (u, v)* is the copula, *F* and *G* are marginal distribution functions, and *H* is the joint cumulative distribution function. The information of the marginal distributions is retained in *F(x)* and *G(y),* and the dependence information is summarized by *C (u, v)*. The dependence relationship is entirely determined by the copula, while the scaling and the shape (e.g., the mean, the standard deviation, the skewness, and the kurtosis) are entirely determined by the marginals.

Definition1: Technically a copula is a method of maping $C:(0,1)$ from the unit 'hypercube' to the unit interval, with marginals that are uniformly distributed on the interval $(0,1)$. The formal definition of copula is as follows,

Definition2: An n-dimensional copula is a function $C: [0,1]^n \to [0,1]$, with the following properties:

1. C is grounded, it means that for every $u = (u_1, u_2, \dots, u_n) \in [0,1]^n, C(u) = 0$ if at least one coordinate u_i is zero, $i = 1, 2, \dots, n$,
2. C is n-increasing, it means that for every $u \in [0,1]^n$ such that $u \le v$, the C-volume $V_C([u,v])$ of the box $[u,v]$ is non-negative.
3. $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$, for all $u_i \in [0,1], i = 1,2, \dots, n$.

For $n = 2$ this definition is reduced to the Bivariate Copula, which is easy to deal with.

Bivariate Copula: A two dimensional (bivariate) copula is a function $C: [0,1]^2 \to [0,1]$, with the following properties:

1. C is grounded: for all $u, v \in [0,1], C(u,0)$ and $C(0,v) = 0$
2. C is 2-increasing: for all $u_1, u_2, v_1, v_2 \in [0,1]$ such that $u_1 \le u_2$ and $v_1 \le v_2$,
   $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \ge 0$
3. For all $u, v \in [0,1], C(u,1) = u$ and $C(1,v) = v$.

Copulas are functions that join or couple multivariate distribution functions to their one dimensional marginal distribution functions. For two random variables X and Y with marginal distributions F(x) and G(y) respectively, their joint behavior is described by joint distribution H(x, y), then the joint distribution for every $(u,v) \in [0,1]^2$ can be expressed by copula function as

$$C(u,v) = P[F(x) \le u, G(y) \le v]$$

$$= P(X \le F^{-1}(u), Y \le G^{-1}(v))$$
$$= H\,[\,F^{-1}(u), G^{-1}(v)$$

Where $F^{-1}(u)$ and $G^{-1}(v)$ are the quantile functions.

If $F(x)$ and $G(y)$ are continuous, then $C(u,v)$ is unique (By Sklar theorem given in the section 2.2). An important feature of copulas is that any choice of marginal distribution can be used. Hence copulas are constructed based on the assumption that marginal distribution functions are known. The two standard nonparametric rank correlations, Kendall's τ and Spearman's ρ are expressed in copula form as:

$$\tau = 4\iint_0^1 C(u,v)dC(u,v) - 1$$

$$\rho = \iint_0^1 C(u,v)dudv - 3$$

The explicit expression of τ for Archimedian copulas are considered here.

Sklar's theorem: Let $F: \Re^n \mapsto (0,1)$ be a joint distribution function with margins $X_1, X_2, \ldots, X_n$. then there exists a copula $C: (0,1)^n \to (0,1)$ such that for all $x \in \Re^n, u \in (0,1)^n$

$$F(x) = C\{F_1(x_1), \ldots, F_n(x_n)\} = C(u)$$

Conversely, if $C: (0,1)^n \to (0,1)$ is a copula and $F_1, F_2, \ldots, F_n$ are distribution functions, then there exists a joint distribution function F with margins $F_1, F_2, \ldots, F_n$ such that for all $x \in \Re^n$, $u \in (0,1)^n$

$$F(x) = F\{F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\} = C(u)$$

furthermore, if $F, F_1, F_2, \ldots, F_n$ are continuous, then the copula C is unique.

Construction of Copulas:

There are several methods of constructing copulas or specifying families of copulas. Some of the most widely used methods are -Inversion Method, Archimedean Approach and Compound Method. The inversion method for constructing bivariate copula is based on the Sklar (1959) theorem [7], where copulas are constructed directly from the joint distribution function. Archimedean approach is a general method of constructing both bivariate and multivariate copulas, which are easier to construct. Marshal and Olkin (1988)[5] suggested the method of construction of copulas which involves the Laplace transform and its inverse function, which is known as compound method.

In the present study, we consider three one parameter (θ) Archimedean bivariate copulas namely, Ali-Mikhail-Haq copula, Cook-Jhonson copula and Gumbel-Hougaard copula for the analysis. These bivariate copulas find a wide range of applications for: (1) it can be easily constructed, (2) the great variety of families of copulas which belong to Archimedean class and (3) the many nice properties possessed by the members of Archimedean class [5].The parameter θ in each case measures the degree of dependence and controls the association between the two variables. When $\theta \to 0$, there is no dependence and if $\theta \to \infty$, there is perfect dependence. The dependence parameter θ which characterizes each family of Archimedean copulas can be related to Kendall's τ .This property is used to empirically determine the applicable copula form

In general, a bivariate Archimedean copula can be defined as (Nelsen, 1999):

$$C_\theta(u_1, u_2) = \phi^{-1}\{\phi(u_1) + \phi(u_2)\}$$

Volume : 2 | Issue : 12 | December 2013 • ISSN No 2277 - 8179

where subscript θ of copula C is the parameter hidden of the generating function Φ . Φ is a continuous function, called generator, strictly decreasing and convex from $I = [0,1]$ to $[0, \Phi(0)]$. The mathematical expressions of some single-parameter bivariate Archimedean copulas and their fundamental properties are listed in **Table 1**.

Table1. One Parameter Family of Bivariate Archimedean Copulas

Sl No.	Family	General Form of Copula $[C_\theta(u,v)]$	$\theta \in$	Generating $Function^{(\alpha)}$ $[\Phi(t)]$	$\tau = 1 + 4\int_0^1 \frac{\Phi(t)}{\Phi'(t)}$
1	Independent	uv	0	lnt	0
2	Ali-Mikhail-Haq Family	$\frac{uv}{[1-\theta(1-u)(1-v)]}$	$[-1,1)$	$ln\left\{\frac{[1-\theta(1-t)]}{t}\right\}$	$\left[\frac{(3\theta-2)}{\theta}\right] - \left[\frac{2}{3}(1-\theta^{-1}\ln(1-\theta)\right]$
3	Cook-Jhonson Family	$\{\max[u^{-\theta}+v^{-\theta}-1,0]\}^{-1/\theta}$; $\theta \geq 0$	$[-1,\infty)\{0\}$	$\frac{[(t)^{-\theta}-1]}{\theta}$	$\frac{\theta}{\theta+2}$
4	Gumbel-Hougaard Family	$exp\left\{-[(-lnu)^\theta + (-lnv)^\theta]^{1/\theta}\right\}$	$[1,\infty)$	$(-lnt)^\theta$	$(1-\theta^{-1})$

(α) $t = u$ *or* v

In these copula functions, the parameter θ synthesizes the dependence strength among the dependent random variables. For each bivariate Archimedean copulas, value of θ can be obtained by considering mathematical relationship (Nelsen, 1999) between Kendall's coefficient of correlation (τ) and generating function is $\Phi(t)$, which is given by $\tau = 1 + 4\int_0^1 \frac{\Phi(t)}{\Phi'(t)}$

Where $t = u$ *or* v (as shown in the last column of **Table 1**).

Determination of Generating Function and Resulting Copula:

The first step in determining a copula is to obtain its generating function from observed data. The procedure to obtain the generating function and the resulting copula is described by Genest and Rivest (1993). It assumes that for a random sample of bivariate observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ the underlying distribution function $H_{XY}(x,y)$ has an associated Archimedean copula C_θ which also can be regarded as an alternative expression of the joint CDF. The procedure involves the following steps:

1. Determine Kendall's τ from the observed data,

2. Determine the parametric or nonparametric marginal distributions for each variable under study.
3. Determine the copula parameter, θ from the above value of τ for Gumbel Hougaard, Cook-Jhonson and Ali- Mikhail-Haq families of copulas given in **Table 1.**
4. Obtain the generating function, Φ , of each copula;
5. Obtain the copula from its generating function.

Illustration: Shear Strength and Age of Propellant Data

For illustration, we consider data given in the book entitled "Introduction to Linear Regression Analysis" written by Montgomery, D.C., Peck, A.E. , Vining, G.G.(2001) of Shear strength and age of propellant data. For the purpose of illustration we analyze data on 20 shear strength and age of propellant as given in the book. Let the random variable X denote the shear strength (psi) and Y the age of propellant (weeks). Here both the variable X and Y are well fitted by lognormal distribution with least standard error of parameter estimates. We apply simulation technique to generate the data set with same distribution and parameters. The variables under study are summarised in statistical sense and are given in the **Table2.** below.

Table2. Summary Statistics of data under study:

Sl No	Descriptive Statistics		Variable	
			Shear Strength	Age of Propellant
1	Range	Statistic	976.05	23
		Minimum	1678.2	2
		Maximum	2654.2	25
2	Mean		2131.4	13.363
3	Standard Deviation		298.57	7.6315
4	Coefficient of Skewness		-0.72315	0.66341
5	Kurtosis		-1.0554	-1.3561

To obtain generating function and copula, the first step is to find the nonparametric rank correlation Kendal's τ. The **Table3.** shows the parametric and nonparametric coefficient of correlation, which are all within the range and not widely varying, which confirms the adequacy of the modeling the dependence structures of Shear Strength and Age of Propellant variables using Archimedean copulas.

Table3. Correlation Coefficient of Shear Strength and Age of Propellant variables

Characteristics	Kendall's coefficient of correlation	Pearson's coefficient of correlation	Spearman's coefficient of correlation
Shear Strength – Age of Propellant	-0.82	-0.949	-0.86

The Ali-Mikhail-Haq copula parameter for 'Shear Strength and Age of

Volume : 2 | Issue : 12 | December 2013 • ISSN No 2277 - 8179

Propellant' is obtained by solving the equation given below

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

and is obtained 0.454.

The plots of marginal distribution function of X and Y simulated data are given below:

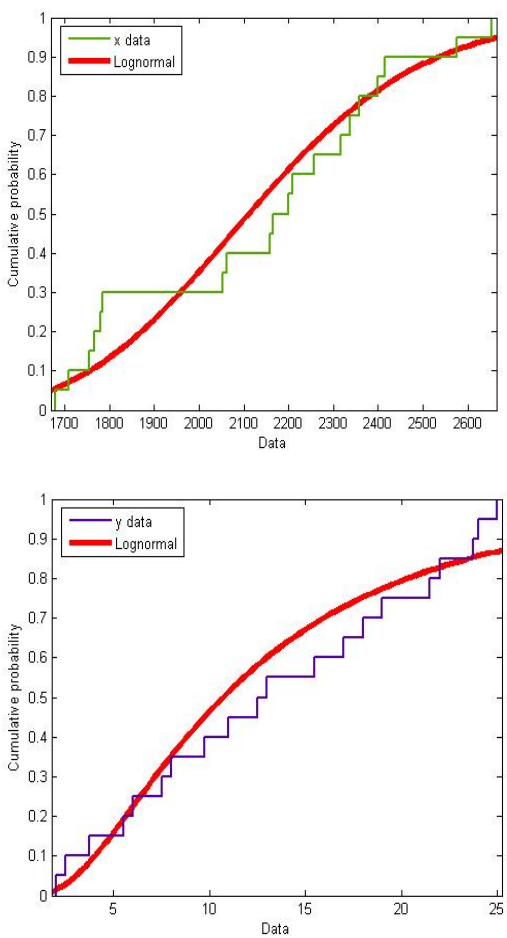

Figure 1 Marginal CDF of age of propellant and Shear strength

After selecting marginal distributions using parametric method, a bivariate joint distributions for shear strength and age of Propellant is determined using the concept of bivariate Archimedean copula. We display the graphical form of all the joint distribution functions for bivariate combinations of shear strength and age of Propellant characteristics mentioned above using Ali-Mikhail-Haq copula in **Figure.2**

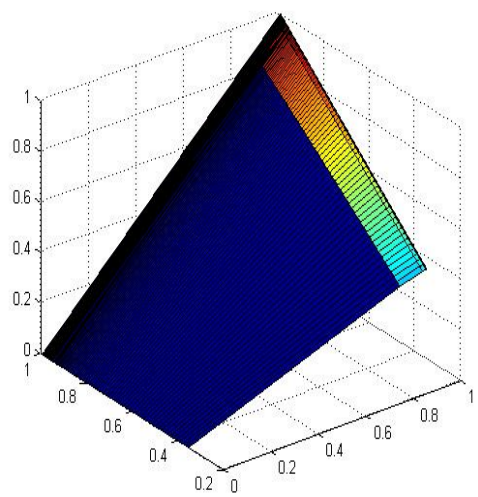

Figure 2 Ali-Mikhail-Haq-Copula

Conclusions

From this study, it is understood that the correlation coefficient is not a complete description of dependence structure between two random variables. An alternative to model the dependence structure is using copulas which overcome the limitations of the correlation. Copulas are functions that join or couple multivariate distribution functions to their one dimensional marginal distribution functions. Copulas allow modeling both linear and nonlinear dependence. Copulas offer an intuitively appealing structure, first for investigating univariate distributions and second for specifying a dependence structure.

REFERENCE

[1] M. Fréchet, Sur les tableaux de corrélation dont les marges sont donnés, Ann. Univ. Lyon, Science, 4, 13–84 (1951) | [2] G. Dall'Aglio, Fréchet classes and compatibility of distribution functions, Symposia Math., 9, 131–150 (1972) | [3] Carlo Sempi, An introduction to Copulas, The 33rd Finnish Summer School on Probability Theory and Statistics, June 6th–10th, 2011. | [4] Sklar, A. (1959): "Fonctions de r´epartition `a n dimensions et leurs marges," Publications de l'Institut de Statistique de l'Universit´e de Paris, 8, 229–231. | [5] Nelsen, RB: (1999): An Introduction To Copulas; Springer Series in Statistics; IInd Edition; Springer- Verlag, Newyork, Inc. (Internet Edition)