Applied Mathematical Sciences, Vol. 4, 2010, no. 14, 657 - 666

Probability Distributions and Estimation of Ali-Mikhail-Haq Copula

Pranesh Kumar

Mathematics Department
University of Northern British Columbia
Prince George, BC V2N 4Z9, Canada
kumarp@unbc.ca

Abstract

The Pearson product-moment correlation commonly used as statistical dependence measure was developed assuming normal marginal and addresses only linear dependence. In most applications, the distribution is assumed to be a multivariate normal or log- normal for tractable calculus even if the assumption may not be appropriate. A copula based approach couples marginal distributions to form flexible multivariate distribution functions. The appeal of copula approach lies in the fact that it eliminates the implied reliance on the multivariate normal or the assumption that dimensions are independent. We present Ali-Mikhail-Haq (AMH) copula and its statistical properties to show that AMH copula could be extensively used in data analysis.

Mathematics Subject Classification 62H20, 62F40, 62E10

Keywords and phrases: Dependence measure, Tail dependence, Copulas, Distribution function, Simulation

1. INTRODUCTION

Copulas express joint distributions of random variables. With a copula we can separate the joint distribution into marginal distributions of each variable. One basic result is that any joint distribution can be expressed in this manner. Another advantage is that the conditional distributions can be readily expressed using the copula. Sklar's theorem (1959) states that any multivariate distribution can be expressed as the copula function $C(u_1, ..., u_i, ..., u_k)$ evaluated at each of the marginal distributions. Using probability integral transform, each continuous marginal $u_i = F_i(x_i)$ has a uniform distribution on $I \in [0, 1]$ where $F_i(x_i)$ is the cumulative integral of $f_i(x_i)$ for the random variable X_i where X_i assume values on the extended real line $[-\infty, \infty]$.

The k-dimensional probability distribution function F has a unique copula representation

$$F(x_1, x_2, ..., x_k) = C(F_1(x_1), F_2(x_2), ..., F_k(x_k)) = C(u_1, u_2, ..., u_k). \quad (1.1)$$

The joint probability density is written as

$$f(x_1, x_2, ..., x_k) = \Pi_{i=1}^{k} f_i(x_i) \times C(u_1, u_2, ..., u_k), \quad (1.2)$$

where $f_i(x_i)$ is each marginal density and coupling is provided by the copula probability density

$$f(u_1, u_2, ..., u_k) = \frac{\partial^k C(u_1, u_2, ..., u_k)}{\partial u_1 \partial u_2 \ ... \ \partial u_k}. \quad (1.3)$$

When random variables are independent, $C(u_1, u_2, ..., u_k)$ is identically equal to one. The importance of equation (1.2) is that the independent portion, expressed as the product of the marginal, can be separated from the function $C(u_1, u_2, ..., u_k)$ describing the dependence structure or shape. The simplest copula is $C(u_1, u_2, ..., u_k) = u_1 u_2 ... u_k$ with the uniform density for independent random variables. Three famous measures of concordance namely Kendall's τ, Spearman's ρ and Gini's index γ, could be expressed in terms of copulas (Schweizer and Wolff, 1981)

$$\tau = 4 \int \int_{I^2} C(u_1, u_2) dC(u_1, u_2) - 1, \quad (1.4)$$

$$\rho = 12 \int \int_{I^2} u_1 u_2 dC(u_1, u_2) - 3, \quad (1.5)$$

$$\gamma = 2 \int \int_{I^2} (|u_1 + u_2 - 1| - |u_1 - u_2|) dC(u_1, u_2). \quad (1.6)$$

It may however be noted that the Pearson's linear correlation coefficient r can not be expressed in terms of copula. The choice of copula can be made using information criteria such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) or the Schwartz information criterion (SIC). Both AIC and BIC penalize the negative maximum log-likelihood of the estimated model by the number of parameters in the model. These criteria are AIC = -2 log (maximum likelihood) + 2 (number of parameters) and BIC = -2 log (maximum likelihood) + (number of parameters)(log of the sample size). A smaller relative AIC or BIC represents a better model fit.

In this paper, we attempt to show that Ali-Mikhail-Haq (AMH) copula have nice statistical properties and tractable results. Hence AMH copula can be quite useful in statistical data analysis. Paper is organized as follows. In section two, we summarize class of Archimedean copulas. AMH copula and probability distributions are discussed in section three. Problem of statistical inference is considered in section four. We illustrate application of copula method in section five. Section six deals with discussion and future research.

2. Archimedean Family of Copulas

An important class of copulas is known as Archimedean copulas. Like a copula, a triangle norm or t-norm maps $[0,1]^p$ to $[0,1]$ and joins distribution functions. Some t-norms (exactly those which are 1-Lipschitz) are copulas and vice versa; some copulas (exactly those which are associative) are t-norms. The Archimedean t-norms which are also copulas are called Archimedean copulas. The Archimedean representation allows to reduce the study of a multivariate copula to a single univariate function.

2.1. Archimedean copulas

Genest and Mackay (1986) define Archimedean copulas as

$$C(u_1,u_2,...,u_k)=\begin{cases}\phi^{-1}\left(\phi(u_1)+...+\phi(u_k)\right), & if\ \sum_{i=1}^{k}\ \phi(u_i)\leq\phi(0)\\ 0, & otherwise\end{cases}, \quad (2.1)$$

where $\phi(u)$ is a C^2 function with $\phi(1)=0$, $\phi^{'}(u)<0$, and $\phi^{''}(u)>0$ for all $u\in[0,1]$. The function $\phi(u)$ is called the *generator* of the copula. Archimedean copulas play an important role because they possess several desired properties of dependence measure like symmetry, associative etc. For any constant $s>0$, the function $s\phi$ is also a generator of C. Archimedean copulas (which are always 2-copulas) as p-ary operators need not be p-copulas. A *necessary* and *sufficient* condition for an Archimedean copula to be p-copula for each $p\geq 2$ is the total *monotonicity* of the function ϕ^{-1} (Nelson, 2006). Different choices of the generator function ϕ yield different copulas. For an Archimedean copula, the Kendall's rank correlation τ can be evaluated directly from the *generator* of the copula (Genest and MacKay, 1986)

$$\tau=1+4\int_0^1\frac{\phi(t)}{\phi^{'}(t)}dt. \quad (2.2)$$

2.2. Tail dependence

The notion of tail dependence is more interesting. Tail dependence describes the limiting proportion that one margin exceeds a certain *threshold* given that the other margin has already exceeded that *threshold*. Joe (1997) defines:

Definition 1. If a bivariate copula $C(u_1,u_2)$ is such that

$$\lim_{u\to 1}\frac{[1-2u+C(u,u)]}{(1-u)}=\lambda_U, \quad (2.3)$$

exists, then $C(u_1,u_2)$ has upper tail dependence for $\lambda_U\in(0,1]$ and no upper tail dependence for $\lambda_U=0$. The measure is extensively used in extreme value theory. It is the probability that one variable is extreme given that the other is extreme, i.e., $\lambda_U=\Pr(U_1>u|U_2>u)$. Thus λ_U can be viewed as a *quantile*-dependent measure of dependence (Coles, Currie and Tawn, 1999). Similarly

lower tail dependence is defined as $\lambda_L = \Pr(U_1 < u | U_2 < u)$, $\lambda_L \in (0, 1]$ and is expressed in terms of copula

$$\lim_{u \to 0} \frac{C(u, u)}{u} = \lambda_L. \quad (2.4)$$

Copula has lower tail dependence for $\lambda_L \in (0, 1]$ and no lower tail dependence for $\lambda_L = 0$.

The one-parameter (θ) families of Archimedean copulas are tabulated in Nelson (2006). In what follows now, we will consider the bivariate case ($k = 2$) for tractable calculus results.

3. Ali-Mikhail-Haq copula and probability distributions

Gumbels's bivariate logistic distribution (Gumbel, 1960) for random variables X_1 and X_2 is given by

$$H(x_1, x_2) = \left(1 + e^{-x_1} + e^{-x_2}\right)^{-1}. \quad (3.1)$$

This joint distribution function $H(x_1, x_2)$ suffers from the defect that it lacks a parameter which limits its usefulness in applications. Ali, Mikhail and Haq (1978) corrected it by defining the joint distribution as

$$H_\theta(x_1, x_2) = [1 + e^{-x_1} + e^{-x_2} + (1 - \theta)e^{-x_1 - x_2}]^{-1}, \quad (3.2)$$

where $\theta \in [-1, 1]$.

By using the probability transform and algebraic method, AMH copula is derived. Alternatively, by considering the generator function

$$\phi(t) = \ln[1 - \theta(1 - t)]/t, \quad (3.3)$$

and from (2.1), AMH copula is defined by

$$C(u_1, u_2) = \frac{u_1 u_2}{1 - \theta(1 - u_1)(1 - u_2)}, \quad (3.4)$$

where the copula parameter $\theta \in [-1, 1]$. It may be noted that AMH copula is the only one amongst twenty two Archimedean copulas tabulated in Nelson (2006) whose parameter lies on a closed interval between -1 and +1 and measures both, positive and negative, dependence.

We now present the results on the characterization of AMH copula.

Theorem 1 . *The AMH copula parameter θ, Kendall's τ and Spearman's ρ satisfy*

$$\tau = \frac{3\theta - 2}{3\theta} - \frac{2(1 - \theta)^2 \ln(1 - \theta)}{3\theta^2}, \quad (3.5)$$

$$\rho = \frac{12(1 + \theta) di \log(1 - \theta) - 24(1 - \theta) \ln(1 - \theta)}{\theta^2} - \frac{3(\theta + 12)}{\theta}, \quad (3.6)$$

where the dilogarithm function $di\log(x)$ *is*

$$di\log(x) = \int_1^x \frac{\ln t}{1-t} dt.$$

Proof. We consider the relationship between Archimedean copula and Kendall's τ in (2.2) and the generator function for AMH copula in (3.3). Then, we have

$$\begin{aligned} \tau &= 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)} dt \; = 1 + 4\int_0^1 \frac{\ln[1-\theta(1-t)]/t}{(d/dt)\ln[(1-\theta(1-t)]/t} dt \\ &= \frac{3\theta - 2}{3\theta} - \frac{2(1-\theta)^2 \ln(1-\theta)}{3\theta^2}. \end{aligned}$$

The AMH copula parameter θ in (3.5) is computed from the above equation.

To prove the relationship between ρ and θ, we have from (1.5)

$$\rho = 12 \int\int_{I^2} u_1 u_2 dC(u_1, u_2) - 3,$$

or equivalently

$$\begin{aligned} \rho &= 12 \int\int_{I^2} C(u_1, u_2) du_1 du_2 - 3 \\ &= 12 \int\int_{I^2} \frac{u_1 u_2}{1-\theta(1-u_1)(1-u_2)} du_1 du_2 - 3 \\ &= 12 \left(\frac{-3\theta + dilog\,(1-\theta)\,\theta + dilog\,(1-\theta) - 2\ln(1-\theta) + 2\,(\ln(1-\theta))\,\theta}{\theta^2} \right) - 3. \end{aligned}$$

Hence we have (3.6). ■

It may be noted that

$$\tau \in \left[\frac{5 - 8\ln 2}{3}, \frac{1}{3} \right] \cong [-0.1817, 0.3333], \quad (3.7)$$

and

$$\rho \in \left[33 - 48\ln 2,\; 4\pi^2 - 39 \right] \cong [-0.2711, 0.4784]. \quad (3.8)$$

We note that AMH copula parameter θ do not cover the entire range [-1,1] of the association measures however it allows both negative and positive dependence.

The next theorem presents the distribution functions.

Theorem 2 . *The conditional distribution function for U_2 given $U_1 = u_1$ is*

$$f(u_2|u_1) = \frac{u_2[1-\theta(1-u_2)]}{[1-\theta\,(1-u_1)\,(1-u_2)]^2}, \quad (3.9)$$

and the joint distribution function for U_1 and U_2

$$f(u_1,u_2) = \frac{1+\ \theta[(1+u_1)\,(1+u_2)-3]+\theta^2\,(1-u_1)\,(1-u_2)}{[1-\theta\,(1-u_1)\,(1-u_2)]^3}. \quad (3.10)$$

Proof. The conditional distribution function $f(u_2|u_1)$ in (3.9) follows from the derivative of $C(u_1,u_2)$ with respect to u_1, i.e., $\frac{\partial}{\partial u_1}C(u,v)$ and the joint distribution function $f(u_1,u_2)$ from $\frac{\partial^2}{\partial u_1 \partial u_2}C(u_1,u_2)$. ■

Tail dependence properties of AMH copula follows next.

Theorem 3 . *AMH copula exhibits left tail dependence for $\theta = 1$.*

Proof. To prove this result, we evaluate the following limits

$$\begin{aligned}\lambda_L &= \lim_{u\to 0}\frac{C(u,u)}{u} = \lim_{u\to 0}\frac{u}{1-\theta\,(\,1-u)^2} \\ &= \begin{cases} \lim_{u\to 0}\frac{u}{1-\theta(\,1-u)^2} = 0.5, & for\theta = 1, \\ \lim_{u\to 0}\frac{u}{1-\theta(\,1-u)^2} = 0, & for\theta < 1, \end{cases}, \quad (3.11)\end{aligned}$$

and

$$\lambda_U = \lim_{u\to 1}\frac{[1-2u+C(u,u)]}{(1-u)} = \lim_{u\to 1}\left[1-\frac{u\,\{1-\theta(1-u)\}}{1-\theta\,(\,1-u)^2}\right] = 0. \quad (3.12)$$

Since $\lambda_L \neq 0$ for $\theta = 1$, left extreme is asymptotically dependent for $\theta = 1$. For every other $\theta < 1$, extremes are asymptotically independent. Hence the theorem. ■

4. Estimation

Copulas involve several underlying functions like marginal distribution functions and joint distribution function. To estimate copula function, we need to specify how to estimate separately the marginal and joint distributions. Depending on underlying assumptions, some quantities have to be estimated parametrically or semiparametrically or nonparametrically. In case of nonparametric estimation, we choose between the usual method based on empirical counterparts and *smoothing* methods like *kernels*, *wavelets*, *orthogonal polynomials*.Without any valuable *prior* information, nonparametric estimation should be preferred especially for the marginal distribution estimation.

4.1. Estimating marginal and joint distributions

Consider a k- sample $(\mathbf{X}_i), i = 1, ..., k$. These are some realizations of the d-random vector $\mathbf{X} = (X_1, ..., X_d)$. We do not assume that $\mathbf{X}_i = (X_{1i}, ..., X_{di})$ are *mutually* independent. Then the j-th marginal distribution function is empirically estimated by

$$F_j(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}(X_{ji} \leq x), \quad (4.1)$$

$[F_j]^{-1}(u_j)$ is the empirical quantile corresponding to $u_j \in [0, 1]$.

Alternatively, j-th marginal distribution can be estimated by using the function $K : \mathbf{R} \to \mathbf{R}$, $\int K = 1$ by

$$F_j(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{K}(\frac{x - X_{ji}}{h}), \quad (4.2)$$

where $h := h_k$ is a bandwidth sequence such that $h_k > 0$ and $h_k \to 0$ when $k \to \infty$.

Similarly, the joint distribution function can be estimated empirically by

$$F(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}(\mathbf{X}_i \leq \mathbf{x}), \quad (4.3)$$

or by the *kernel* method

$$F(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{K}(\frac{\mathbf{x} - \mathbf{X}_i}{h}), \quad (4.4)$$

with a d-dimensional *kernel*

$$\mathbf{K}(\mathbf{x}) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_d} K, \quad (4.5)$$

for every $\mathbf{x} = (x_1, ..., x_d) \in \mathbf{R}^d$.

Thus, a d-dimensional copula can be estimated by

$$\hat{C}(\mathbf{u}) = F\left([F_1]^{-1}(u_1), ..., [F_d]^{-1}(u_d)\right). \quad (4.6)$$

4.2. Simulation from AMH copula

Simulation in statistics help to investigate properties of the estimator and to understand the underlying joint distributions. In general to generate copula from the k-variate $\mathbf{X} = (X_1, ..., X_k)$ using the conditional distribution method, steps are:

1. Generate a random number $X_1 = x_1$ from the marginal distribution of X_1.

2. Generate a random number $X_2 = x_2$ from the conditional distribution of X_2, given that $X_1 = x_1$.

3. Generate a random number $X_3 = x_3$ from the conditional distribution of X_3, given that $X_2 = x_2$ and $X_1 = x_1$.

4. And so on, for all X_k.

To simulate AMH copula $C(u_1, u_2)$ in (3.4), we thus have the following simplified steps:

1. Generate two *independent* random numbers u_1 and t on (0,1).

2. Let

$$\begin{aligned} a &= 1 - u_1; b = 1 - \theta(1 + 2at) + 2\theta^2 a^2 t; \\ c &= 1 + \theta(2 - 4a + 4at) + \theta^2(1 - 4at + 4a^2 t\ ). \quad (4.7) \end{aligned}$$

3. Set

$$u_2 = \frac{2t(a\theta - 1)^2}{b + \sqrt{c}}. \quad (4.8)$$

3. The desired simulated pair is (x_1, x_2) where

$$x_1 = {F_1}^{-1}(u_1),\ x_2 = {F_2}^{-1}(u_2). \quad (4.9)$$

4.3. Maximum likelihood estimation

In case of uniform marginal, the copula is equivalent to the joint cumulative distribution function. The model parameters can thus be estimated using maximum likelihood method. Let θ be the vector of parameters to be estimated and Θ the parameter space. The likelihood for an observation t,i.e., the probability density of observation t, is function of θ and let it be $L_t(\theta)$. Further, denote by $l_t(\theta)$ the log-likelihood of $L_t(\theta)$. Given n observations, $l(\theta) = \sum_{t=1}^{n} l_t(\theta)$. Then the maximum likelihood estimator $\hat{\theta}_{ML}$ of parameter vector θ satisfies $l(\hat{\theta}_{ML}) \geq l(\theta)$, for all $\theta \in \Theta$. The maximum likelihood estimator $\hat{\theta}_{ML}$ is asymptotically normal and thus $\sqrt{n}\left(\hat{\theta}_{ML} - \theta_0\right) \to N(\mathbf{0}, J^{-1}(\theta_0))$ where $J(\theta_0)$ is the Fisher information matrix.

Applying to the bivariate AMH copula model for a random sample of size n, the log-likelihood for θ is

$$\begin{aligned} l(\theta) &= \sum_{t=1}^{n} \log c(u_{1t}, u_{2t}) \\ &= \sum_{t=1}^{n} \log \frac{1 + \ \theta[(1 + u_{1t})\,(1 + u_{2t}) - 3] + \theta^2\,(1 - u_{1t})\,(1 - u_{2t})}{[1 - \theta\,(1 - u_{1t})\,(1 - u_{2t})]^3}. \quad (4.10) \end{aligned}$$

The maximum likelihood estimator $\hat{\theta}_{ML}$ is the value of θ which maximizes (4.10). The AIC and BIC measures are given by

$$\begin{aligned} AIC &= -2l(\theta) + 2(number of model parameters), \\ BIC &= -2l(\theta) + (\log n)(number of model parameters). \end{aligned}$$

Since, the AMH copula have a single parameter (θ), a comparison of AIC or BIC measures is equivalent to a comparison of their log-likelihoods.

5. Application of AMH copula model

To illustrate the applicability of copula modelling ideas in real situations, we consider a study reported by Morrow *et al.* (1992) which enrolled twenty three patients in a split-mouth trial for the treatment of *gingivitis*. In this trial four sites located either on the left or right side of a patient's mouth were randomly assigned to either the experimental *treatment* (chlorhexidine) or a *control* (saline). Plaque measurements were taken pre-treatment and two weeks after baseline on four sites of the patient's upper jaw. Here we consider modelling the post-treatment proportions of sites exhibiting plaque in *treatment* (X_1) and *control* (X_2) groups at a two-week follow-up visit. Estimated value of Kendall's τ is 0.176. The marginal distributions estimated from the probability plots are: $X_1 \tilde{}$Beta$(66.88, 8.16)$ and $X_2 \tilde{}$Beta$(57.91, 17.13)$. AMH copula parameter θ is estimated from τ and $\theta = 0.\,64813$. Then AMH copula in this case is

$$C(u_1, u_2) = \frac{u_1 u_2}{1 - 0.\,64813(1-u_1)(1-u_2)},\ u_1 u_2 \in [0,1], \quad (5.1)$$

The model for the proportion of sites exhibiting plaque in treatment (X_1) and control (X_2) groups is

$$f(u_1, u_2) = \frac{1 + \ 0.\,64813((1+u_1)\,(1+u_2) - 3) + 0.420\,07\,(1-u_1)\,(1-u_2)}{(1 - 0.\,64813\,(1-u_1)\,(1-u_2))^3}, \quad (5.2)$$

$u_1, u_2 \in [0,1]$ and $x_1 =$Beta$^{-1}(u_1; 66.88, 8.16)$ and $x_2 =$Beta$^{-1}(u_2; 57.91, 17.13)$

The conditional probabilities $P(U_2 = u_2 | U_1 = u_1)$ are computed from

$$f(u_2|u_1) = \frac{u_2[1 - .64813(1-u_2)]}{[1 - .64813\,(1-u_1)\,(1-u_2)]^2}. \quad (5.3)$$

To obtain the maximum likelihood estimate of θ, we have used AMH copula parameter θ estimated from the Kendall's τ and have simulated 1000 pairs of (U_1, U_2). Then, we have numerically evaluated likelihood function in (4.10) for $\theta = -1(\ 0.1)1$. We note that $\theta\ = 0.795$ maximizes the likelihood function. Thus, the maximum likelihood estimate of θ is 0.795 as compared to 0.64813 estimated from Kendall's τ.

6. Discussion

We presented concept of copula functions and AMH copula in particluar. We described how they can be used in estimating the probability models in case of multivariate data. Copulas model basically the dependence structure

of the random variables. They provide a convenient way to model and simulate correlated variables. Several copulas with varying shapes are available for modeling dependence. We have obtained the conditional and joint probability distributions of AMH copula which are expressed in simple algebraic functions. We have also studied tail probabilities in AMH copula framework. Although copula is an old notion, there are many research areas to explore. This paper focussed on bivariate copula but many of the concepts can be generalized to the multivariate case. Parametric estimation may lead to severe underestimation when the parametric models of marginal distributions or copula are misspecified. Nonparametric methods may also result in underestimation when the smoothing methods does not consider potential boundary biases in the tail of the density support.

ACKNOWLEDGMENT

Research was supported by the author's discovery grant from the *Natural Sciences and Engineering Research Council of Canada (NSERC)* which is duly acknowledged.

REFERENCES

[1] Ali, M.M., Mikhail, N.N. and Haq, M.S. (1978). A class of bivariate distributions including the bivariate logistic. *J. Multivariate Anal.* 8, 405-412.

[2] Coles, S., Currie, J. and Tawn, J. (1999). Dependence measures for extreme value analyses. Working paper: Department of Mathematics and Statistics, Lancaster University.

[3] Genest, C. and Mackay, J. (1986). The joys of copulas: Bivariate distributions with uniform marginals. *American Statistician*, 40, 280-283.

[4] Gumbel, E.J. (1960). Bivariate Exponential Distributions. *J. Amer. Stat. Assoc.*, 55, 698-707.

[5] Joe, H. (1997). *Multivariate Models and Dependent Concepts*. New York: Chapman & Hall.

[6] Morrow, D., Wood, D.P. and Speechley, M. (1992). Clinical effect of subgingival chlorhexidine irrigation on gingivitis in adolescent orthodontic patients. *Amer. J. Orthodontics and Dentofacial Orthopedics*, 101, 408-413.

[7] Nelson, R. (2006). *An Introduction to Copulas*. New York: Springer.

[8] Schweizer, B. and Wolff, E. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics*, 9, 879-885.

[9] Sklar, A. (1959). Fonctions de répartition á n dimensional et leurs marges. *Publ. Inst. Stat. Univ. Paris*, 8, 229-231.

Received: June, 2009