# Copula - Overview

Arabin Kumar Dey

Assistant Professor

Department of Mathematics
Indian Institute of Technology Guwahati

**Talk at Indian Institute of Technology Guwahati**

6-th August, 2010

# Outline

# Outline

- The word copula is a Latin noun which means a link, tie or bond, and was first employed in a mathematical or statistical sense by **Abe Sklar.**
- Mathematically, a copula is a function which allows us to combine univariate distributions to obtain a joint distribution with a particular dependence structure.

# Outline

Sklar's theorem, which is the foundation theorem for copulas, states that for a given joint multivariate distribution function and the relevant marginal distributions, there exists a copula function that relates them. In a bi-variate setting:

Let $F_{xy}$ be a joint distribution with margins $F_x$ and $F_y$. Then there exists a function $C : [0,1]^2 \to [0,1]$ such that

$$F_{xy}(x,y) = C(F_X(x), F_Y(y)) \quad (1)$$

If X and Y are continuous, then C is unique; otherwise, C is uniquely determined on the (range of X)$\times$(range of Y). Conversely if C is a copula and $F_x$ and $F_y$ are distribution functions, then the function $F_{xy}$ defined by (1) is a 111,1 distribution with margins $F_x$ and $F_y$.

# Outline

C is a copula if $C : [0,1]^2 \rightarrow [0,1]$ and

1. $C(0,u) = C(v,0) = 0$
2. $C(1,u) = C(u,1) = u$
3. $C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geq 0$ for all $v_1 < v_2$ and $u_1 < u_2$.

If C is differential once in its first argument and once in its second, (iii) is equivalent to

$$\int_{v_1}^{v_2} \int_{u_1}^{u_2} \frac{\delta^2 C}{\delta u \delta v} dudv \geq 0$$

for all $v_1 \leq v_2$, $u_1 \leq u_2$ in the range. From this observation we see that the definition simply states that a copula is itself a distribution function, defined on $[0, 1]^2$, with uniform marginals. Each of the marginal distributions produces a probability of the one dimensional events.

# Example 1

Let the distribution function $H$ be

$$H(x,y) = \begin{cases} \frac{(x+1)(e^y-1)}{x+2e^y-1} & \text{if } (x,y) \in [-1,1] \times [0,\infty] \\ 1 - e^{-y} & \text{if } (x,y) \in (1,\infty] \times [0,\infty] \\ 0, & \text{elsewhere.} \end{cases}$$

with marginals given by

$$F(x) = \begin{cases} 0 & \text{if } x < -1 \\ (x+1)/2, & \text{if } x \in [-1,1] \\ 1, & x > 1 \end{cases}$$

$$G(y) = \begin{cases} 0 & \text{if } y < 0 \\ (1 - e^{-y}), & \text{if } y \geq 0 \end{cases}$$

Quasi-inverses of $F$ and $G$ are given by $F^{-1}(u) = (2u - 1)$ and $G^{-1}(v) = -\ln(1 - v)$ for $u, v$ in some I. Therefore the copula C is given by,

$$C(u, v) = \frac{uv}{(u + v - uv)}.$$

# Example 2

Gumbel's bivariate exponential distribution: Let the $H_\theta$ be the joint distribution function given by

$$H_\theta(x,y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-x-y-\theta xy} & \text{if } x \geq 0,\ y \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\theta$ is the parameter in [0,1]. Then the marginal distribution functions are exponentials, with quasi-inverses $F^{-1}(u) = -\ln(1-u)$ and $G^{-1}(u) = -\ln(1-v)$ for $u, v \in I$. Hence the corresponding copula is

$$C_\theta(u,v) = u + v - 1 + (1-u)(1-v)e^{-\theta \ln(1-u)(1-v)}$$

- The most commonly used copulae are the Gumbel copula for extreme distributions.
- The Gaussian copula for linear correlation, and the Archimedean copula and the t-copula for dependence in the tail.

## Outline

- There are variety of procedures used to generate observations $(x, y)$ of a pair or random variables $(X, Y)$ with a joint distribution function $H$. In this section we will focus on using the copula as a tool.
- By virtue of Sklar's theorem, we need only generate a pair $(u, v)$ of observations of uniform (0,1) random variables $(U, V)$ whose joint distribution function is $C$, the copula of $X$ and $Y$, and then transform those uniform variates in the following manner.
- One procedure for generating such a pair $(u, v)$ of uniform $(0, 1)$ variates is the conditional distribution method. For this method, we need the conditional distribution function for $V$ given $U = u$, which we denote $c_u(v)$: $c_u(v) = P(V \leq v|U = u) = \lim_{\Delta u \to 0} \frac{C(u+\Delta u, v) - C(u, v)}{\Delta u} = \frac{\delta C(u, v)}{\delta u}$

# Algorithm

- Generate two independent uniform (0,1) variates $u$ and $t$.
- Set $v = c_u^{-1}(t)$, where $c_u^{-1}$ denotes a quasi-inverse of $c_u$.
- The desired pair is $(u, v)$.

# Example

- Let $X$ and $Y$ be random variables whose joint distribution function $H$ is

$$H(x,y) = \begin{cases} \frac{(x+1)(e^y-1)}{x+2e^y-1} & \text{if } (x,y) \in [-1,1] \times [0,\infty] \\ 1 - e^{-y} & \text{if } (x,y) \in (1,\infty] \times [0,\infty] \\ 0, & \text{elsewhere.} \end{cases}$$

# Example Continued

- The copula C is given by,

$$C(u,v) = \frac{uv}{(u+v-uv)}.$$

- The conditional distribution function $c_u$ and its inverse $c_u^{(-1)}$ are given by $c_u(v) = \frac{\delta C(u,v)}{\delta u} = \left(\frac{v}{u+v-uv}\right)^2$ and $c_u^{-1}(t) = \frac{u\sqrt{t}}{1-(1-u)\sqrt{t}}$.

An algorithm to generate random variable (x,y) is :

- Generate two independent uniform (0,1) variates u and t;
- Set
$$v = \frac{u\sqrt{t}}{1-(1-u)\sqrt{t}}$$
- Set $x = 2u - 1$ and $y = -\ln(1-v)$
- The desired pair is (x,y).

# Outline

- **Ali- Mikhail-Haq :** $c(u,v) = uv[1-\alpha(1-u)(1-v)]^{-1}$
- **Cook-Johnson :** $c(u,v) = [u^{-\alpha} + v^{-\alpha} - 1]^{-\frac{1}{\alpha}}, \qquad \alpha \geq 0$
- **Farlie-Gumbel-Morgenstern:** $c(u,v) = uv[1+\alpha(1-u)(1-v)], \qquad -1 < \alpha < 1.$
- **Frank:** $c(u,v) = -\frac{1}{\alpha}\ln[1 + \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1}] \qquad \alpha \neq 0$

- **Gumbel-Hougaard :**
  $c(u,v) = exp{-}[(-\ln u)^{\alpha} + (-\ln v)^{\alpha}]^{\frac{1}{\alpha}} \qquad \alpha \geq 1$
- **Clayton Copula :**
  $c(u,v) = max[-(-logu)^{1/\beta} + (-logv)^{1/\beta^{\beta}}, 0]$
- **Normal :** $c(u,v) = H[\Phi^{-1}(u), \Phi^{-1}(v)], \qquad -1 \leq \alpha \leq 1,$
  where $\Phi^{-1}(u)$ is the inverse of the univariate standard normal distribution and $H(x,y)$ is a standard normal distribution function with correlation $\alpha$.

# Outline

# Kendal's Tau

Let $(x_i, y_i)$ and $(x_j, y_j)$ denote two observations from a vector $(X, Y)$ of continuous random variables.

- We say that $(x_i, yi)$ and $(x_j, y_j)$ are concordant if $x_i < x_j$ and $y_i < y_j$, or if $x_i > x_j$ and $y_i > y_j$.
- We say that $(x_i, yi)$ and $(x_j, y_j)$ are discordant if $x_i < x_j$ and $y_i > y_j$ or if $x_i > x_j$ and $y_i < y_j$

# Kendall's Tau

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ denote a random sample of $n$ observations from a vector $(X, Y)$ of continuous random variables. There are $\binom{n}{r}$ distinct pairs of $(x_i, y_i)$ and $(x_j, y_j)$ of observations in the sample, and each pair is either concordant or discordant.

- let $c$ denote the number of concordant pairs and $d$ the number of discordant pairs.
- Kendall's tau for the sample is defined as

$$t = \frac{c-d}{c+d} = \frac{c-d}{\binom{n}{r}}$$

- Population version of Kendall's tau is defined as

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

# Copula and Kendall's tau

Let $(X_1, Y_1)$ and $(X_2, Y_2)$ be independent vectors of continuous random variables with joint distributions $H_1$ and $H_2$, respectively, with common margins $F$ (of $X_1$ and $X_2$) and $G$ (of $Y_1$ and $Y_2$). Let $C_1$ and $C_2$ denotes the copulas of $(X_1, Y_1)$ and $(X_2, Y_2)$, respectively, so that $H_1(x, y) = C_1(F(x), G(y))$ and $H_2 = C_2(F(x), G(y))$. Let $Q$ denote the difference between the probabilities of concordance and discordance, i.e. let

$$Q = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

Then $Q = Q(C_1, C_2) = 4 \int \int_{I^2} C_1(u, v) dC_2(u, v) - 1$

The following theorem is used to calculate the Kendall's tau. A example is shown using the theorem in the next slide.

If $C_1$ and $C_2$ are two copula's. Then

$$\int \int_{I^2} C_1(u,v) dC_2(u,v) = \frac{1}{2} - \int \int_{I^2} \frac{\delta C(u,v)}{\delta u} \frac{\delta C(u,v)}{\delta v} dudv$$

Let $C_{\alpha,\beta}$ be a member of the Marshall-Olkin family of Copulas for $\alpha > 0$, $\beta < 1$ :

$$C_{\alpha,\beta} = \begin{cases} u^{1-\alpha}v & u^{\alpha} \geq v^{\beta} \\ uv^{1-\beta} & u^{\alpha} \leq v^{\beta} \end{cases}$$

The partials of $C_{\alpha,\beta}$ fails to exist only on the curve $u^\alpha = v^\beta$, so that

$$\frac{\delta C_{\alpha,\beta}(u,v)}{\delta u}\frac{\delta C_{\alpha,\beta}(u,v)}{\delta v} = \begin{cases}(1-\alpha)u^{1-2\alpha}v & u^\alpha > v^\beta \\ (1-\beta)uv^{1-2\beta} & u^\alpha < v^\beta\end{cases}$$

and hence

$$\int\int_{I^2} \frac{\delta}{\delta u}C_{\alpha,\beta}(u,v)\frac{\delta}{\delta v}C_{\alpha,\beta}(u,v) = \frac{1}{4}\left(1 - \frac{\alpha\beta}{\alpha - \alpha\beta + \beta}\right)$$

from which we obtain, Kendal Tau equals to

$$\tau_{\alpha,\beta} = \frac{\alpha\beta}{(\alpha - \alpha\beta + \beta)}$$

**Note : Given Kendal Tau we can set the parameters of the chosen copula.**

# Copula and Spearman's rank correlation co-efficient

Let F(X) and F(Y) be the cdf of X and Y. Therefore, Spearman's rank correlation co-efficient can be given by

$$\begin{aligned}\rho = Corr(F(X), F(Y)) &= 12 \int\int_{I^2} uv \; dC(u, v) - 3 \\ &= 12 \int\int_{I^2} C(u, v) du dv - 3\end{aligned}$$

# Outline

Let X and Y be continuous radom variable with distribution F and G, respectively. The upper tail dependence parameter $\lambda_U$ is the limit of conditional probability that $Y$ is greater than the 100t -th percentile of $F$ as t approaches 1, i.e.

$$\lambda_U = \lim_{t \to 1^-} P[Y > G^{-1}(t) | X > F^{(-1)}(t)]$$

Similarly lower tail dependence parameter $\lambda_L$ is the limit (if it exists) of the conditional probability that $Y$ is less than or equal to the 100t-th percentile of G given that X is less than or equal to the 100t-th percentile of $F$ as $t$ approaches to 0, i.e.

$$\lambda_L = \lim_{t \to 0^+} P[Y \leq G^{-1}(t) | X \leq F^{-1}(t)].$$

Let $X, Y, F, G, \lambda_U$ and $\lambda_L$ be as in the above slide and let C be the copula of X and Y, with diagonal section $\delta_C$. If the limits in the above slide exists, then

$$\lambda_U = 2 - \lim_{t \to 1^-} \frac{1 - C(t,t)}{1 - t} = 2 - \delta_C'(1^-)$$

and

$$\lambda_L = \lim_{t \to 0^+} \frac{C(t,t)}{t} = \delta_C'(0^+)$$

# Proof:

We see,

$$\begin{aligned}\lambda_U = \lim_{t\to 1^-} P[Y > G^{-1}(t)|X > F^{-1}(t)] \\ &= \lim_{t\to 1^-} P[G(Y) > t|F(X) > t] \\ &= \lim_{t\to 1^-} \frac{1-\bar{C}(t,t)}{1-t} \\ &= \lim_{t\to 1^-} \frac{1-2t+C(t,t)}{1-t} \\ &= 2-\delta'_C(1^-)\end{aligned}$$

**If $\lambda_U$ is in (0,1], we say C has upper tail dependence; if $\lambda_U = 0$, we say C has no upper tail dependence; similarly for $\lambda_L$**

# Outline

# Empirical Copulas

Let $\{x_k, y_k\}_{k=1}^{n}$ denote a sample of size n from a continuous bivariate distribution. The empirical copula is the function $C_n$ given by $C_n(\frac{i}{n}, \frac{j}{n}) = \frac{\text{number of pairs (x,y) in the sample with} x \leq x_i, y \leq y_j}{n}$ where $x_i$ and $y_j$, $1 \leq i, j \leq n$, denote the order statistics from the sample. The empirical copula frequency $c_n$ is given by

$$c_n(\frac{i}{n}, \frac{j}{n}) = \begin{cases} \frac{1}{n} & \text{if } \ (x_{(i)}, y_{(j)}) \ \text{ is an element of the sample} \\ , 0 & \text{otherwise} \end{cases}$$

# Spearman Rank correlation and Kendal's tau

Note that $C_n$ and $c_n$ are related via
$C_n(\frac{i}{n}, \frac{j}{n}) = \sum_{p=1}^{n} \sum_{q=1}^{n} c_n(\frac{p}{n}, \frac{q}{n})$
and $c_n(\frac{i}{n}, \frac{j}{n}) = C_n(\frac{i}{n}, \frac{j}{n}) - C_n(\frac{i-1}{n}, \frac{j}{n}) - C_n(\frac{i}{n}, \frac{j-1}{n}) + C_n(\frac{i-1}{n}, \frac{j-1}{n})$
Spearman rank correlation =

$$\frac{12}{n^2-1} \sum_{i=1}^{n} \sum_{j=1}^{n} [C_n(\frac{i}{n}, \frac{j}{n}) - \frac{i}{n}\frac{j}{n}]$$

Kendall's tau =

$$\frac{2n}{n-1} \sum_{i=2}^{n} \sum_{j=2}^{n} \sum_{p=1}^{i-1} \sum_{q=1}^{j-1} [c_n(\frac{i}{n}, \frac{j}{n}) c_n(\frac{p}{n}, \frac{q}{n}) - c_n(\frac{i}{n}, \frac{q}{n}) c_n(\frac{p}{n}, \frac{j}{n})]$$