

Cópuas: Utilização no R e aplicações com dados de sobrevivência

Rafael Barros Leal

Trabalho realizado sob orientação do
Prof. Doutor Wilton Bernardino da Silva

28 de fevereiro de 2015

Resumo Cópulas são funções que fazem uma ligação entre funções de distribuição multivariada às suas respectivas distribuições marginais unidimensionais. As utilidades das funções de cópulas são várias. Elas são muito eficientes para estudar medidas de dependência livres de escalas. São também um ponto de partida para construir famílias de distribuição bivariadas. Neste texto iremos explorar algumas aplicações das cópulas na análise de sobrevivência e as funções relacionadas às cópulas disponíveis no software R.

Palavras chaves: Teoria de Cópulas, Cópulas de sobrevivência, Pacote Copula.

Abstract Copulas are functions that provide a link between multivariate distribution functions and their univariate margins. Copulas can be used by statisticians in several ways. They are very efficient in the study of scale-free measures of dependence. Also, it can be used as a starting point for constructing families of bivariate distributions. In this paper, we will explore some of its applications regarding survival analysis and computational tools related to copulas available in the software R.

Palavras chaves: Copulas Theory, Survival Copulas, Copula Package.

Sumário

1	Introdução à Cópulas	4
1.1	Revisão da literatura	4
1.2	Definições e propriedades básicas	6
1.2.1	A cópula do produto.	7
1.2.2	Limites superior e inferior de Fréchet-Hoeffding	7
1.2.3	Teorema de Sklar	8
1.2.4	Famílias de Cópulas	10
1.2.5	Medidas de associação	13
2	Cópulas de Sobrevivência	17
2.1	Algumas cópulas importantes	18
2.1.1	Exponencial Bivariada de Gumbel	18
2.1.2	Farlie-Gumbel-Morgenstern	19
2.1.3	Weibull Bivariada	20
3	Aplicação das cópulas no R	21
3.1	Pacotes no R e algumas funções	21
3.2	Seleção de cópulas	25
4	Métodos para geração de dados	26
4.1	Abordagem pela distribuição condicional	26
4.2	Algoritmo para cópulas Arquimedianas em geral.	27
4.3	Abordagem de Marshall e Olkin	28
5	Estudo de simulação	30
6	Conclusões	35
A	Um exemplo utilizado na simulação.	37

1 Introdução à Cópulas

O estudo das cópulas e suas aplicações na estatística é um acontecimento relativamente moderno. Ainda existem muitos problemas em aberto e trabalho a ser feito adentro do tema das cópulas.

A palavra *cópula* vem do latim e significa “um elo, ligação, vínculo”. A palavra foi utilizada pela primeira vez num sentido ligado a estatística por Abe Sklar em seu teorema, que será melhor explorado mais a frente. Neste teorema, Sklar descrevia funções que “uniam” funções de distribuições unidimensionais para formar funções de distribuições multivariadas. Porém, as funções precedem o termo que as caracterizam, sendo citadas em trabalhos de Hoeffding(1941) e Fréchet(1951). Desde então, cópulas tiveram um crescente interesse, principalmente em áreas como bioestatística, atuária e finanças.

No presente trabalho, iremos explorar os principais resultados relativos às cópulas, a definição das cópulas, suas principais famílias, propriedades, pacotes de programação do software R e aplicações da teoria.

1.1 Revisão da literatura

Por mais que o tema da teoria de cópulas já possua mais de 50 anos de vida, a sua aplicação e estudo cresceram com mais força apenas na última década, devido aos avanços dos softwares estatísticos. Diversos artigos foram publicados recentemente com aplicações em seguradoras no mercado americano, possuidor do maior faturamento do mundo, mas poucos existem no Brasil. É interessante analisar esta teoria também em nosso país, visto que este possui o mercado segurador com maior faturamento da América latina e um dos maiores crescimentos percentuais anuais do mundo.

Em 1959, Sklar [14] mostrou que qualquer distribuição multivariada pode ser escrita em uma representação de cópulas, mesmo este processo não sendo sempre conveniente. Neste artigo ele mostrou que, caso as distribuições marginais sejam contínuas, existe uma representação única com base em cópulas.

Um dos maiores atuários com atuação no estudo das cópulas é Edward W. Frees. Em [3], Frees estuda um dos modelos de precificação de seguros mais utilizados nos estados unidos, atrelado a teoria de cópulas. Ele mostra que existe uma relação direta entre o método de precificação baseado em credibilidade e as distribuições de perda. Neste texto, ele avalia a qualidade de suas previsões em

relação às seguradoras de automóveis de Massachusetts e em comparação a outras previsões existentes na literatura. Em [5] ele também sugere um modelo de cópulas para precificação de rendas vitalícias em seguradoras.

Além destes trabalhos, em [4] Frees explora algumas aplicações práticas das cópulas, como em modelos conjuntos de mortalidade. Também é mostrado como simular resultados multivariados a partir de distribuições expressadas nas suas formas de cópulas. Este procedimento é realizado com auxílio de algoritmos utilizando cópulas Arquimedianas ou compostas. O método que utiliza cópulas compostas foi primeiramente sugerido por Marshall e Olkin [19].

Em [15], Staudt introduz várias considerações a serem tomadas quando da modelagem utilizando cópulas, mais especificamente em relação ao comportamento de distribuições com caudas pesadas, tanto em suas distribuições marginais quanto em suas distribuições conjuntas. Em seu artigo, Staudt sugere ao leitor considerar mais medidas de associação, visto que a correlação mede apenas a dependência linear. O Autor também aconselha a cautela quando da escolha das cópulas, visto que o formato mais natural da curva deveria ser um produto de considerações teóricas conhecidas e analisadas a priori.

Em [10] encontramos o provável melhor livro para iniciar o estudo da teoria de cópulas. Neste texto, Nelsen apresenta os principais pontos do estudo e as principais definições, de forma clara e didática. O texto possui certa complexidade, assim como tudo que tange o assunto das cópulas, mas é apropriado o suficiente para o estudo em nível de graduação. O intuito do livro é de ser utilizado por qualquer pessoa com boa base matemática e probabilística.

Alguns fatos e limitações do uso das cópulas para modelar distribuições conjuntas com marginais discretas são expostos em [20]. Parte destas limitações são consequência direta da ausência da propriedade de unicidade das cópulas para distribuições discretas. Neste texto também é explicado que a dependência entre as distribuições marginais não é função apenas da cópula.

Dadas distribuições marginais discretas, cópulas podem ser utilizadas para estudos de robustez incluindo análises de simulação. Porém, no texto é mostrado que este campo de estudo ainda precisa de avanços científicos para ser tão funcional como é o caso com marginais contínuas.

Diversos textos na literatura mostram a eficácia do uso da teoria de cópulas na modelagem de dados relacionados a seguradoras, principalmente no que se re-

fere a análise de sobrevivência. O uso das cópulas aplicado às ciências atuariais já é executado de forma relativamente ampla fora do Brasil, porém poucos trabalhos foram realizados no Brasil.

Associado ao software R, Jun Yan escreveu em [17] um dos textos de mais fácil interpretação e uso computacional. O texto apresenta o uso das funções mais importantes e básicas do pacote `copula`. Neste texto, o autor sugere em diversas situações que uma boa abordagem para a decisão de qual cópula utilizar deve ser baseada na estimação por máxima verossimilhança.

O pacote `copula` é provavelmente o mais avançado dos pacotes relativos às cópulas implementado no R, porém ainda está longe de ser perfeito e apenas possui implementadas algumas famílias de cópulas.

1.2 Definições e propriedades básicas

Para iniciar este capítulo, vamos descrever uma definição informal para as cópulas. Considere duas variáveis aleatórias X e Y com funções de distribuição $F(x) = P[X \leq x]$ e $G(y) = P[Y \leq y]$, respectivamente, e distribuição conjunta $H(x, y) = P[X \leq x, Y \leq y]$, então a cópula é uma função $C : [0, 1]^2 \rightarrow [0, 1]$ tal que,

$$H(x, y) = C\{F(x), G(y)\}, \quad (x, y) \in \mathbf{R}^2$$

Formalmente, temos que: Uma cópula bidimensional é uma função $C : [0, 1]^2 \rightarrow [0, 1]$ tal que,

1. $C(0, y) = C(x, 0) = 0$, $C(1, y) = y$ e $C(x, 1) = x \quad \forall \quad x, y \in I$

2. C é **2 – crescente**, ou seja, para $a, b, c, d \in I$ com $a \leq b$ e $c \leq d$,

$$V_c([a, b] \times [c, d]) = C(b, d) - C(a, d) - C(b, c) + C(a, c) \geq 0,$$

Em que a função V_c é chamada de C-Volume do retângulo $[a, b] \times [c, d]$. Isto quer dizer que, caso $H(x, y)$ possua segunda derivada, então a derivada de $F(x)$ e de $G(y)$ é maior ou igual a zero, como é provado em [10]

Note que, por definição, uma cópula não necessita que seus argumentos sejam funções de distribuição. Porém, quando estes argumentos são funções de distribuições univariadas, a cópula é uma função de distribuição bivariada com marginais iguais aos seus argumentos.

Para facilitar um pouco o entendimento, vamos citar algumas cópulas importantes e mostrar que elas seguem todas as propriedades acima.

1.2.1 A cópula do produto.

Vamos mostrar que a função bivariada $H(u, v) = uv$ é uma cópula de $u, v \in [0, 1]$

- Domínio de $H = S_u \times S_v = I^2 = [0, 1]^2$
- $H(0, v) = H(u, 0) = 0$
- $H(1, v) = v$ e $H(u, 1) = u$
- Seja $B = [u_1, u_2] \times [v_1, v_2] \rightarrow V_H(B) \geq 0 \quad \rightarrow H$ é 2-crescente

É fácil provar que isso vale para todo $u_1, u_2, v_1, v_2 \in [0, 1]$ fazendo a substituição $u_2 = u_1 + \Delta_u$ e $v_2 = v_1 + \Delta_v$ com $\Delta_u, \Delta_v \geq 0$.

Após algumas manipulações algébricas chegamos no seguinte resultado:

$$V_H(B) = \Delta_u \Delta_v \geq 0$$

A cópula do produto é comumente escrita como $\prod(u, v) = uv$

1.2.2 Limites superior e inferior de Fréchet-Hoeffding

Antes de propriamente citar os limites superior e inferior de Fréchet-Hoeffding, vamos enunciar algumas outras definições e um teorema.

Uma subcópula é uma função mais ampla que as cópulas (e o conjunto das cópulas está contido no conjunto das subcópulas). Vamos citar uma definição antes de partir para as subcópulas.

Definição. Sejam S_1 e S_2 subconjuntos de $\bar{\mathbf{R}}$, onde $\bar{\mathbf{R}}$ representa a linha dos números reais estendida $[-\infty, +\infty]$. Sejam s_1 e s_2 os menores elementos de S_1 e S_2 , respectivamente. Seja H uma função definida de $S_1 \times S_2 \rightarrow \mathbf{R}$. Dizemos que H é uma função *grounded* se $H(x, s_2) = 0 = H(s_1, y)$ para todo $x, y \in S_1 \times S_2$

A diferença entre subcópulas e cópulas é que subcópulas são funções *grounded* que não precisam necessariamente estar definidas em $\mathbf{I} = [0, 1]$ mas em um subconjunto de \mathbf{I} que contenha 0 e 1. Subcópulas estão melhor definidas, por exemplo, em [1].

Sendo assim, toda propriedade de uma subcópula também é propriedade da cópula.

O teorema a seguir discute os limites superior e inferior para uma subcópula definida em um domínio contido em $\bar{\mathbf{R}}^2$, o plano real estendido definido como $\bar{\mathbf{R}}^2 = \bar{\mathbf{R}} \times \bar{\mathbf{R}}$.

Teorema Seja C' uma subcópula. Então para todo $(u, v) \in \text{Dom}C'$,

$$\max(u + v - 1, 0) \leq C'(u, v) \leq \min(u, v)$$

Prova Seja (u, v) um ponto qualquer que pertença a $\text{Dom}C'$. Então, $C'(u, v) \leq C'(u, 1) = u$ e $C'(u, v) \leq C'(v, 1) = v$, logo, $C'(u, v) \leq \min(u, v)$. Além disso, temos que

$$V_{C'}([u, 1] \times [v, 1]) \geq 0 \rightarrow 1 - u - v + C'(u, v) \geq 0 \rightarrow C'(u, v) \geq u + v - 1$$

Combinando isto com o fato de $C'(u, v) \geq 0$, temos que $C'(u, v) \geq \max(u + v - 1, 0)$.

Como toda cópula é uma subcópula, a inequação acima vale para cópulas. É fácil provar que os dois limites são também cópulas, da mesma forma que foi feito acima com a cópula do produto $\prod(u, v)$.

Estes tais limites são denotados por $M(u, v) = \min(u, v)$ e $W(u, v) = \max(u + v - 1, 0)$ e são conhecidos, respectivamente, como limites superior e inferior de *Fréchet-Hoeffding*.

1.2.3 Teorema de Sklar

Agora que já foi descrito acima algumas definições e propriedades importantes das cópulas, iremos finalmente mostrar como existe a aplicação da teoria das cópulas no âmbito da Estatística. O teorema de Sklar é extremamente importante e mostra a relação que existe entre funções de distribuições multivariadas e suas respectivas marginais univariadas.

Definição. Uma função de distribuição conjunta é uma função H com domínio em $\bar{\mathbf{R}}^2$ tal que:

- H é 2-,

- $H(x, -\infty) = H(-\infty, y) = 0$ e $H(\infty, \infty) = 1$.

Logo H é grounded e, como $DomH = \bar{\mathbf{R}}^2$, H possui como marginais as funções de distribuições F e G dadas por $F(x) = H(x, \infty)$ e $G(y) = H(\infty, y)$.

Lemma 1. Seja H função de distribuição conjunta com marginais F e G , então existe uma única subcópula tal que:

- $DomC' = \text{Img}F \times \text{Img}G$
- Para todo $x, y \in \bar{\mathbf{R}}$, $H(x, y) = C'(F(x), G(y))$

Lemma 2. Seja C' uma subcópula. Então existe uma cópula C tal que

$$C(u, v) = C'(u, v) \quad \forall (u, v) \in DomC'.$$

Ou seja, toda subcópula pode ser estendida para uma cópula.

As provas para os dois Lemmas citados acima são encontradas em [14] ou em [10].

Teorema de Sklar. Seja H função de distribuição conjunta com marginais F e G , então existe uma cópula C tal que, para todo $x, y \in \bar{\mathbf{R}}$,

$$H(x, y) = C(F(x), G(y)). \quad (1)$$

Se F e G são contínuas, então C é única. Caso contrário, C é unicamente definida em $\text{Img}F \times \text{Img}G$. Reciprocamente, se C é uma cópula, F e G são funções de distribuição, então a função H definida em (1) é uma função de distribuição conjunta com marginais F e G .

A equação (1) dá uma expressão que escreve uma função de distribuição conjunta em termos da cópula e das funções de distribuições marginais univariadas.

Note que todas essas definições foram dadas ao contexto de funções de distribuição, mas não foi indicado que, por exemplo, $F(x)$ e $G(y)$ eram funções de distribuição de uma respectiva variável aleatória. A partir de agora iremos começar a associar as funções de distribuição a variáveis aleatórias e verificar a aplicação das teorias das cópulas junto a funções de distribuições acumuladas.

Sendo assim, podemos reescrever o teorema de Sklar em termos de variáveis aleatórias e suas distribuições.

Teorema de Sklar. Sejam X e Y variáveis aleatórias com funções de distribuição F e G respectivamente e distribuição conjunta H . Então existe uma cópula C tal como foi definido em (1). Se F e G são contínuas, C é única. Caso contrário, C é unicamente definida em $\text{Img}F \times \text{Img}G$. A cópula definida acima é chamada de cópula de X e Y e pode ser denotada como C_{XY} . A extensão da cópula bivariada para a cópula multivariada é simples.

1.2.4 Famílias de Cópulas

Existem várias famílias de cópulas importantes e com várias propriedades diferentes. Utilizando a distribuição normal multivariada é possível construir a **família de Cópulas Normais(ou Gaussianas)**. A matriz de correlações é quem parametriza esta cópula. Caso a matriz de correlação seja a matriz diagonal, a cópula Gaussiana é reduzida a cópula do produto.

Uma das famílias de cópulas mais importantes é a família de cópulas Arquimedianas, que possui implementação computacional relativamente fácil devido às suas propriedades matemáticas.

Seja $\varphi(x) : [0, 1] \rightarrow [0, +\infty]$ uma função contínua e estritamente decrescente, tal que $\varphi(0) = +\infty$ e $\varphi(1) = 0$ e $\frac{d^2\varphi}{dx^2} > 0$, então a cópula Arquimediana pode ser escrita como

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)),$$

onde $\varphi(x)$ é chamada de função geradora da cópula e $\varphi^{-1}(x)$ representa a inversa da geradora.

Existem várias funções $\varphi(x)$ que fornecem cópulas Arquimedianas importantes. A Tabela 1 descreve algumas dessas cópulas para diversas funções geradoras φ .

Tabela 1.

Nome	Função geradora $\varphi(x)$	Inversa da geradora $\varphi^{-1}(x)$	Parâmetro $\theta \in$
Ali-Mikhail-Haq	$\log\left(\frac{1-\theta+\theta x}{x}\right)$	$\frac{1-\theta}{\exp(x)-\theta}$	$[-1, 1)$
Clayton	$\frac{x^{-\theta}-1}{\theta}$	$(1+\theta x)^{-1/\theta}$	$[-1, +\infty) \setminus 0$
Frank	$-\log\left(\frac{\exp(-\theta x)-1}{\exp(-\theta)-1}\right)$	$-\frac{\log(1+\exp(-x)(\exp(-\theta)-1))}{\theta}$	$(-\infty, +\infty) \setminus 0$
Gumbel-Hougaard	$(-\log(x))^\theta$	$\exp(-x^{1/\theta})$	$[1, +\infty)$
Joe	$1 - (1 - \exp(-x))^{1/\theta}$	$-\log(1 - (1 - x)^\theta)$	$[1, +\infty)$

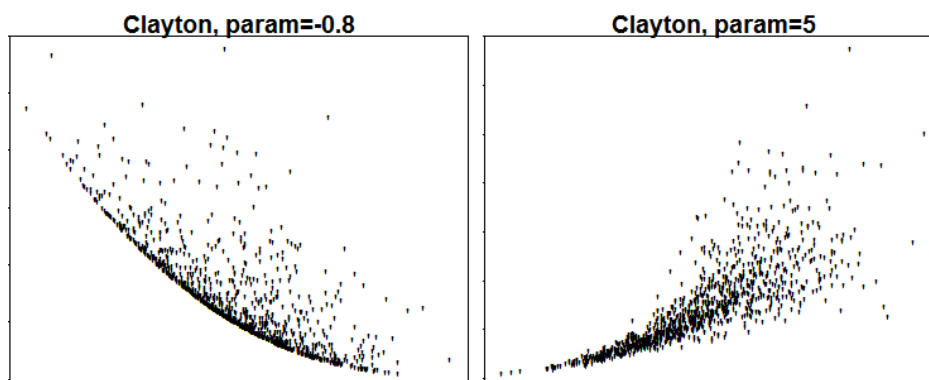
Nome	Cópula Bivariada $C_\theta(u, v)$
Ali-Mikhail-Haq(1978)	$\frac{uv}{1-\theta(1-u)(1-v)}$
Clayton(1978)	$(\max\{u^{-\theta} + v^{-\theta} - 1; 0\})^{-1/\theta}$
Frank(1979)	$-\frac{1}{\theta} \log\left(1 + \frac{\exp(-\theta u) - 1}{\exp(-\theta) - 1} \frac{\exp(-\theta v) - 1}{\exp(-\theta) - 1}\right)$
Gumbel-Hougaard(1960)	$\exp(-((-\log(u))^\theta + (-\log(v))^\theta)^{1/\theta})$
Joe(1997)	$1 - ((1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta)^{1/\theta}$

Alguns autores chamam a família de cópulas de Gumbel-Hougaard apenas de família de Gumbel, visto que esta foi abordada pela primeira vez por Gumbel em [6]. Porém, não faremos o mesmo, pois existe uma outra família de cópulas Arquimedianas conhecida como família de Gumbel. Esta outra família será descrita com mais detalhes abaixo.

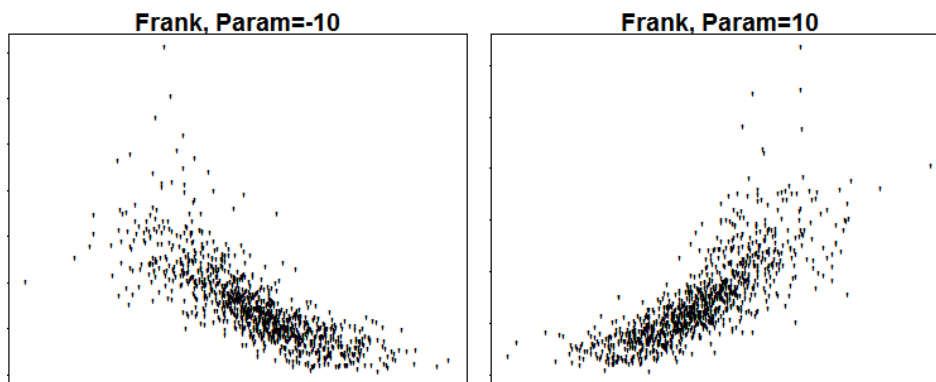
A cópula de Frank possui uma propriedade interessante. Ela é a única família de cópulas Arquimedianas em que $\hat{C}(u, v) = C(u, v)$, ou seja, sua cópula de sobrevivência tem forma igual a sua função de cópula. A função \hat{C} definida como cópula de sobrevivência será melhor detalhada no capítulo 2.

Além disso, se $\varphi(x) = \exp(-x)$, obtemos uma cópula do produto.

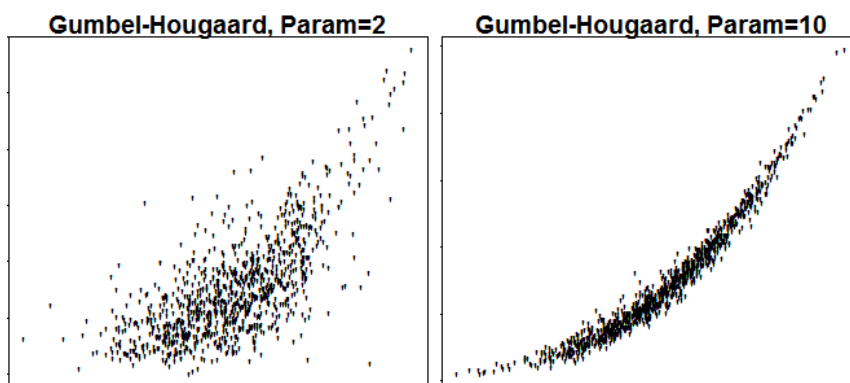
Seguem, a título de exemplificação, alguns diagramas de dispersão e suas respectivas cópulas.



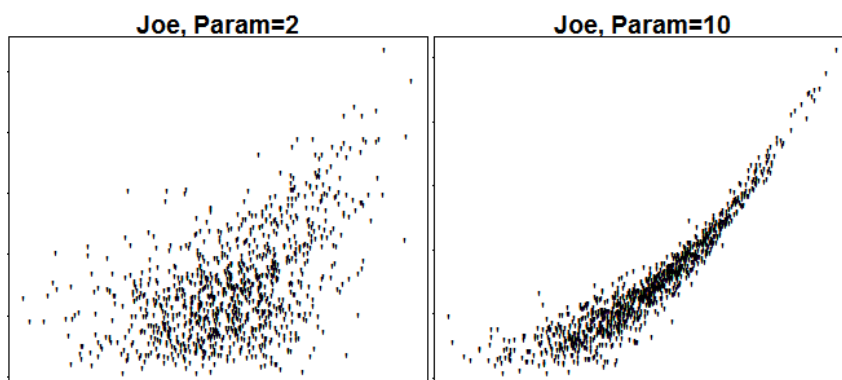
Diagramas de dispersão da cópula de Clayton e seus respectivos parâmetros θ



Diagramas de dispersão da cópula de Frank e seus respectivos parâmetros θ



Diagramas de dispersão da cópula de Gumbel-Hougaard e seus respectivos parâmetros θ



Diagramas de dispersão da cópula de Joe e seus respectivos parâmetros θ

Além do que foi definido neste trabalho, existem funções com propriedades

semelhantes às das cópulas, porém não cumprem todos os pré-requisitos para serem caracterizadas como cópulas. Estas funções são chamadas de quasi-cópulas. As propriedades e aplicações das quasi-cópulas não estão indicadas neste texto, porém podem ser encontrada, por exemplo, em [10].

1.2.5 Medidas de associação

Uma das utilidades das cópulas está presente no estudo de medidas de dependência ou associação entre variáveis aleatórias. Para este fim, as medidas de associação mais famosas são o rho de Spearman e o tau de Kendall. Ambas medem a probabilidade de **concordância** entre variáveis aleatórias com uma respectiva cópula.

A fim de explanar melhor essas medidas, vamos primeiro esclarecer a definição de “concordante”.

Definição. Sejam (x_i, y_i) e (x_j, y_j) duas observações num vetor (X, Y) de variáveis aleatórias. Dizemos que (x_i, y_i) e (x_j, y_j) são concordantes se, caso $x_i < x_j$ e $y_i < y_j$ ou se $x_i > x_j$ e $y_i > y_j$. Caso contrário, dizemos que (x_i, y_i) e (x_j, y_j) são *discordantes*. Ou seja, (x_i, y_i) e (x_j, y_j) são concordantes se $(x_i - x_j)(y_i - y_j) > 0$ e são discordantes se $(x_i - x_j)(y_i - y_j) < 0$.

Informalmente, dizemos que um par de variáveis aleatórias são concordantes caso valores altos de uma tendem a estar associados a valores altos da outra e valores baixos de uma tendem a estar associados a valores baixos da outra.

Tau de Kendall

O tau de Kendall é um teste popular por sua simplicidade, seja na sua interpretação quanto na sua estrutura algébrica.

Seja $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ uma amostra aleatória de n observações do vetor aleatório (X, Y) . Sendo assim, temos $\binom{n}{2}$ pares distintos (x_i, y_i) e (x_j, y_j) de observações aleatórias na amostra, e cada um desses pares é concordante ou discordante. Seja c o número de pares concordantes e d o número de pares discordantes, então o tau de Kendall para a amostra é definido como:

$$t = \frac{c - d}{c + d} = (c - d) / \binom{n}{2}$$

Se separarmos esta última razão em duas frações, t representa a probabilidade

de concordância menos a probabilidade de discordância para um par (x_i, y_i) e (x_j, y_j) retirado aleatoriamente da amostra. É de forma semelhante que é definido o tau de Kendall para a população.

A versão populacional do tau de Kendall é dada por:

$$\tau_{X,Y} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

Em que (X_1, Y_1) e (X_2, Y_2) são vetores aleatórios independentes e identicamente distribuídos, cada um com função de distribuição conjunta H .

A utilidade das cópulas nas aplicações relacionadas a medidas de associação, como por exemplo o tau de Kendall é demonstrada pelo seguinte:

Sejam (X_1, Y_1) e (X_2, Y_2) variáveis aleatórias contínuas, com funções de distribuição conjuntas H_1 e H_2 diferentes, mas marginais comuns F e G . Podemos calcular $Q = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$, onde Q é definida como a função de concordância. Esta função depende da distribuição de (X_1, Y_1) e (X_2, Y_2) somente através de suas cópulas.

Teorema Sejam (X_1, Y_1) e (X_2, Y_2) vetores independentes de variáveis aleatórias contínuas, com funções de distribuição conjuntas H_1 e H_2 , respectivamente, com marginais comuns F para X_1 e X_2 e G para Y_1 e Y_2 . Sejam C_1 e C_2 as cópulas de (X_1, Y_1) e (X_2, Y_2) , respectivamente, tal que $H_1(x, y) = C_1(F(x), G(y))$ e $H_2(x, y) = C_2(F(x), G(y))$. Seja Q a função de concordância definida acima, dada por:

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

Então,

$$Q = Q(C_1, C_2) = 4 \int \int_{I^2} C_2(u, v) dC_1(u, v) - 1.$$

A prova deste teorema pode ser vista em [10].

Algumas das propriedades importantes da função Q , dados C_1 e C_2 , são as seguintes.

1. Q é simétrica em relação aos seus argumentos: $Q(C_1, C_2) = Q(C_2, C_1)$
2. Q é não decrescente em cada um dos seus argumentos.
3. As cópulas podem ser substituídas pelas cópulas de sobrevivência: $Q(C_1, C_2) = Q(\hat{C}_1, \hat{C}_2)$. Tais cópulas serão definidas no capítulo 2

Teorema. Sejam X e Y variáveis aleatórias contínuas com cópula C . Então o tau de Kendall populacional é dado por:

$$\tau_{X,Y} = \tau_C = Q(C, C) = 4 \int \int_{I^2} C(u, v) dC(u, v) - 1$$

Note que a integral dupla que aparece acima pode ser interpretada como o valor esperado da função $C(U, V)$ onde U e V são variáveis aleatórias com distribuição uniforme $(0, 1)$ e C é a sua função de distribuição conjunta. Ou seja:

$$\tau_C = 4E(C(U, V)) - 1$$

Rho de Spearman

Assim como o tau de Kendall, a versão populacional do rho de Spearman é baseado nos conceitos de concordância e discordância. Sejam:

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3),$$

três vetores independentes de variáveis aleatórias com função de distribuição conjunta H comum e com marginais F e G . O rho de Spearman populacional $\rho_{X,Y}$ é proporcional a probabilidade de concordância menos a probabilidade de discordância entre os dois vetores (X_1, Y_1) e (X_2, Y_3) , isto é, ambos os vetores possuem as mesmas distribuições marginais, mas um deles tem função de distribuição conjunta H e o outro tem componentes independentes.

$$\rho_{X,Y} = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0])$$

Poderíamos também ter utilizado o par (X_3, Y_2) . Note que a função de distribuição conjunta de (X_1, Y_1) é $H(x, y)$, enquanto a função de distribuição conjunta de (X_2, Y_3) é dada por $F(x)G(y)$, pois X_2 e Y_3 são independentes. Dessa forma, a cópula de X_2 e Y_3 é a cópula do produto \prod .

Teorema. Sejam X e Y variáveis aleatórias contínuas com cópula C , então o rho de Spearman populacional de X e Y é dado por:

$$\rho_{X,Y} = \rho_C = 3Q(C, \prod) = 12 \int \int_{I^2} uv dC(u, v) - 3 = 12 \int \int_{I^2} C(u, v) dudv - 3 \quad (2)$$

Como pode ser visto em [10], o rho de Spearman para um par de variáveis aleatórias X e Y é idêntico ao coeficiente de correlação de Pearson para as variáveis aleatórias $U = F(X)$ e $V = G(Y)$, com função de distribuição conjunta C . Uma interpretação dada para a integral em (2) é de que ela representa o volume abaixo do gráfico da cópula C e acima de $I^2 = [0, 1] \times [0, 1]$. Dessa forma, o rho de

Spearman é então o volume abaixo do gráfico da cópula C , mas reescalado para sempre resultar em um valor que esteja no intervalo $[-1, 1]$.

Ambas as medidas (rho de Spearman e tau de Kendall) expressam a probabilidade de concordância entre variáveis aleatórias com uma respectiva cópula, porém, geralmente os valores de ρ e τ são diferentes. Porém o limite para o quão diferentes essas medidas podem ser é definido pelas seguintes inequações:

$$\frac{3\tau - 1}{2} \leq \rho \leq \frac{1 + 2\tau - \tau^2}{2}, \tau \geq 0$$

$$\frac{\tau^2 + 2\tau - 1}{2} \leq \rho \leq \frac{1 + 3\tau}{2}, \tau \leq 0$$

Abaixo seguem algumas das cópulas mais importantes e o tau de Kendall e rho de Spearman associado a cada uma delas.

Tabela 2.

Nome	Tau de Kendall
Ali-Mikhail-Haq(1978)	$\frac{3\theta-2}{3\theta} - \frac{2(1-\theta)^2}{3\theta^2} \ln(1-\theta)$
Clayton(1978)	$\frac{\theta}{\theta+2}$
Frank(1979)	$1 - \frac{4}{\theta} [1 - (\frac{1}{\theta} \int_0^\theta \frac{t}{e^t-1} dt)]$
Gumbel-Hougaard(1960)	$\frac{\theta-1}{\theta}$
FGM(1960)	$\frac{2\theta}{9}$

Nome	Rho de Spearman
Ali-Mikhail-Haq(1978)	$\frac{12(1+\theta)}{\theta^2} \int_1^{(1-\theta)} \frac{\ln(t)}{1-t} dt - \frac{24(1-\theta)}{\theta^2} \ln(1-\theta) - \frac{3(\theta+12)}{\theta}$
Clayton(1978) ¹	$12 \int \int_{I_2} (\max(u^{-\theta} + v^{-\theta} - 1; 0))^{-\frac{1}{\theta}} dudv - 3$
Frank(1979)	$1 - \frac{12}{\theta} [(\frac{1}{\theta} \int_0^\theta \frac{t}{e^t-1} dt) - (\frac{2}{\theta^2} \int_0^\theta \frac{t^2}{e^t-1} dt)]$
Gumbel-Hougaard(1960) ¹	$12 \int \int_{I_2} (\exp(-((- \log(u))^\theta + (- \log(v))^\theta)^{1/\theta})) dudv - 3$
FGM(1960)	$\frac{\theta}{3}$

¹Esta é uma conta complicada e difícil de resolver. Em alguns casos não existe forma explícita

2 Cópulas de Sobrevivência

Agora que algumas propriedades e definições relacionadas as cópulas já foram dadas, podemos tomar o foco principal do nosso estudo: a aplicação das cópulas na análise de sobrevivência.

O estudo das cópulas na modelagem de dados de sobrevivência multivariada foi abordado por vários estudiosos como Gumbel, Clayton, Oakes e outros.

Em várias aplicações, as variáveis aleatórias de interesse medem o tempo até um respectivo evento, por exemplo, a probabilidade de um animal sobreviver além de um tempo x . Essa probabilidade é dada pela função de sobrevivência $\bar{F}(x) = P[X > x] = 1 - F(x)$, onde F representa a função de distribuição de X .

Dado um par de variáveis aleatórias (X, Y) com função de distribuição conjunta H , a função de sobrevivência conjunta é dada por $\bar{H}(x, y) = P[X > x, Y > y]$. As marginais de \bar{H} são as funções $\bar{H}(x, -\infty)$ e $\bar{H}(-\infty, y)$, que são as funções de sobrevivência univariadas \bar{F} e \bar{G} , respectivamente.

Assim como existe uma cópula entre as funções de distribuição conjuntas e suas respectivas funções de distribuição univariadas, existe uma relação entre as funções de sobrevivência conjuntas \bar{H} e as respectivas funções de sobrevivência

univariadas \bar{F} e \bar{G} , como é possível ver em [10]. Temos que:

$$\begin{aligned}\bar{H}(x, y) &= 1 - F(x) - G(y) + H(x, y) \\ &= \bar{F}(x) + \bar{G}(y) - 1 + C(F(x), G(y)) \\ &= \bar{F}(x) + \bar{G}(y) - 1 + C(1 - \bar{F}(x), 1 - \bar{G}(y)),\end{aligned}$$

Dessa forma, podemos definir a função \hat{C} de \mathbf{I}^2 em \mathbf{I} como,

$$\hat{C}(u, v) = u + v - 1 + C(1 - u, 1 - v),$$

e assim, temos

$$\bar{H}(x, y) = \hat{C}(\bar{F}(x), \bar{G}(y))$$

e nos referimos a \hat{C} como a cópula de sobrevivência entre X e Y . Podemos provar que \hat{C} é uma cópula da mesma forma que fizemos acima com a cópula do produto.

- $\hat{C}(x, 1) = x + C(1 - x, 0) = x$ e $\hat{C}(1, y) = y$
- \hat{C} é uma função *grounded*:

$$\hat{C}(k, 0) = \hat{C}(0, k) = k - 1 + C(1, 1 - k) = k - 1 + 1 - k = 0$$

- \hat{C} é 2-crescente, ou seja, $V_{\hat{C}}([u_1, v_1] \times [u_2, v_2]) =$

$$C(1 - v_1, 1 - v_2) - C(1 - v_1, 1 - u_2) - C(1 - u_1, 1 - v_2) + C(1 - u_1, 1 - u_2)$$

Assim, se a cópula entre variáveis aleatórias é conhecida, então a sua cópula de sobrevivência pode ser obtida a partir desta. Em muitos casos é mais simples trabalhar com funções de sobrevivência do que com funções de distribuição, da mesma forma que em alguns casos a cópula de sobrevivência entre variáveis aleatórias é mais simples que a função de cópula entre as variáveis aleatórias.

2.1 Algumas cópulas importantes

2.1.1 Exponencial Bivariada de Gumbel

Consideremos a distribuição exponencial Bivariada de Gumbel, com função de distribuição acumulada dada por:

$$F(x, y) = 1 - e^{-x} - e^{-y} + e^{-(x+y+\theta xy)}, \quad x, y \geq 0, \quad 0 \leq \theta \leq 1$$

Dessa forma, é possível verificar que a cópula é dada por

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

e aplicando $\hat{C}(u, v) = u + v - 1 + C(1 - u, 1 - v)$, facilmente encontramos a cópula de sobrevivência dada por

$$\hat{C}_\theta(u, v) = uv e^{-\theta \ln(u) \ln(v)},$$

que é uma função mais simples que a cópula C_θ . Porém, existem também desvantagens na utilização da exponencial bivariada, visto que não é possível obter correlações fora do intervalo $[-.25, .25]$ entre processos de sobrevivência, como é demonstrado em [12].

2.1.2 Farlie-Gumbel-Morgenstern

Esta cópula foi discutida por Morgenstern(1956), Gumbel(1958) e Farlie(1960). A sua forma é dada por:

$$C_\theta(u, v) = uv(1 + \theta(1 - u)(1 - v)), -1 \leq \theta \leq 1$$

Essa cópula é uma variação da cópula do produto \prod . Note que se $\theta = 0$, obtemos exatamente a cópula do produto, porém, as cópulas da família FGM são não-Arquimedianas para todos os outros valores de θ .

Esta é uma cópula muito utilizada na modelagem de dados e em testes de associação devido a simplicidade da sua forma analítica. Se fizermos $\hat{C}_\theta(u, v) = u + v - 1 + C(1 - u, 1 - v)$, temos que:

$$\begin{aligned} \hat{C}_\theta(u, v) &= u + v - 1 + (1 - u)(1 - v)(1 + \theta uv) \\ &= u + v - 1 + (1 - u)(1 - v) + \theta uv(1 - u)(1 - v) \\ &= u + v - 1 + 1 - u - v + uv + \theta uv(1 - u)(1 - v) \\ &= uv + \theta uv(1 - u)(1 - v) = C_\theta(u, v) \end{aligned}$$

E concluímos que as cópulas da família Farlie-Gumbel-Morgenstern possuem simetria radial, ou seja, $C_\theta = \hat{C}_\theta$. Estas cópulas também possuem a propriedade de serem absolutamente contínuas. As distribuições do tipo FGM são vastamente utilizadas em modelagem, testes de associação e no estudo da eficiência de procedimentos não paramétricos, devido a sua forma analítica simples, como podemos ver em [10]. Por outro lado, as cópulas da família Farlie-Gumbel-Morgenstern não são apropriadas para modelar dados onde há uma forte relação de dependência, vide [12].

Além da família de cópulas acima, alguns estudiosos como Kotz e Johnson(1977) e Lin(1987) criaram novas versões da família de cópulas FGM utilizando a função de sobrevivência conjunta,

$$\bar{C}_\theta = 1 - u - v + C_\theta(u, v) = (1 - u)(1 - v)(1 + \theta uv)$$

e perceberam que \bar{C}_θ e \hat{C}_θ possuem em sua formulação a função de distribuição conjunta uv e a função de sobrevivência conjunta $(1 - u)(1 - v)$.

Kotz e Johnson substituíram os termos $(1 - u)(1 - v)$ por \bar{C}_θ , mas com um novo parâmetro ψ de forma a obter a *FGM iterada de Kotz e Johnson*:

$$C_{\theta,\psi}(u, v) = uv + \theta uv(1 - u)(1 - v)[1 + \psi uv]$$

De forma semelhante, Lin substituiu, em \bar{C}_θ , o termo uv por C_θ , mas com um novo parâmetro ψ de forma a obter a *FGM iterada de Lin*:

$$C_{\theta,\psi}(u, v) = uv + \theta uv(1 - u)(1 - v)[1 + \psi(1 - u)(1 - v)]$$

Ao aplicar $\hat{C}(u, v) = u + v - 1 + C(1 - u, 1 - v)$ na *FGM iterada de Kotz e Johnson*, temos que:

$$\begin{aligned} \hat{C}(u, v) &= u + v - 1 + (1 - u)(1 - v) + \theta uv(1 - u)(1 - v)[1 + \psi(1 - u)(1 - v)] \\ &= u + v - 1 + 1 - u - v + uv + \theta uv(1 - u)(1 - v)[1 + \psi(1 - u)(1 - v)] \\ &= uv + \theta uv(1 - u)(1 - v)[1 + \psi(1 - u)(1 - v)] \end{aligned}$$

Ou seja, a *FGM iterada de Lin* é a cópula de sobrevivência da *FGM iterada de Kotz e Johnson*.

2.1.3 Weibull Bivariada

Uma vantagem em relação a exponencial bivariada de Gumbel, é que a Weibull Bivariada não possui um problema significativo em relação ao parâmetro de correlação.

Sejam as funções de distribuição de Weibull univariadas dadas por $F(x) = 1 - e^{-(\frac{x}{\lambda_x})^{p_x}}$ e $G(y) = 1 - e^{-(\frac{y}{\lambda_y})^{p_y}}$

Sejam agora duas funções de distribuição Weibull bivariadas, dadas por:

$$W(x, y | \lambda_x, \lambda_y, p_x, p_y, \theta) = 1 - e^{-(\frac{x}{\lambda_x})^{p_x}} - e^{-(\frac{y}{\lambda_y})^{p_y}} + e^{-(\frac{x}{\lambda_x})^{p_x} - (\frac{y}{\lambda_y})^{p_y} - \theta(\frac{x}{\lambda_x})^{p_x}(\frac{y}{\lambda_y})^{p_y}}$$

$$Z(x, y | \lambda_x, \lambda_y, p_x, p_y, \theta) = [1 - e^{-(\frac{x}{\lambda_x})^{p_x}}][1 - e^{-(\frac{y}{\lambda_y})^{p_y}}][1 + \theta e^{-(\frac{x}{\lambda_x})^{p_x} - (\frac{y}{\lambda_y})^{p_y}}]$$

De acordo com $C' = H[F^{-1}, G^{-1}]$ onde C' é uma cópula e $H(x, y)$ é uma função de distribuição bivariada, é possível mostrar que as cópulas de X e Y entre as funções acima são dadas respectivamente por:

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

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

Note também que a cópula de sobrevivência $\hat{C}(u, v)$ para a primeira função é idêntica a cópula de sobrevivência da exponencial bivariada de Gumbel, ou seja, $\hat{C}(u, v) = uv e^{-\theta \ln(u)\ln(v)}$. E isto não é por acaso. Note que se $p_x = p_y = 1$, então tanto $W(x, y)$ quanto $Z(x, y)$ são de distribuição exponencial bivariada de Gumbel.

O parâmetro θ , citado algumas vezes acima tem importante função na abordagem das cópulas, sendo a covariância e algumas medidas de associação função deste θ .

Este parâmetro é essencial na estimação de distribuições bivariadas e geralmente define um limite para o intervalo onde pode estar a correlação linear entre as marginais univariadas. Dessa forma, a Weibull bivariada se mostra muito superior a exponencial bivariada de Gumbel na análise de sobrevivência, visto que o parâmetro de correlação na exponencial bivariada é fortemente limitado.

3 Aplicação das cópulas no R

Existem diversos pacotes que fornecem aplicações das cópulas na modelagem multivariada. Nestes pacotes estão implementadas, inclusive, diversas classes contendo várias famílias de cópulas citadas anteriormente.

3.1 Pacotes no R e algumas funções

O pacote “copula” possui diversas funções que auxiliam na modelagem e simulação. O artigo [17] proporciona, de forma simples, um entendimento das principais funções do pacote “copula”, assim como alguns exemplos de suas aplicações. O pacote possui uma boa implementação da classe das cópulas Arquimedianas.

As linhas abaixo mostram como fazer isso no R:

```
>library(copula)
>frankCop<-archmCopula(family="frank",dim=2,param=1)
```

Onde a família é definida pelo argumento `family`. `dim` representa a dimensão da cópula (no nosso caso é uma cópula bivariada), e `param` representa o nosso θ . É possível também construir distribuições multivariadas a partir destas cópulas.

```
>biv<-mvdc(copula=frankCop,margins=c("norm","gamma")
+,paramMargins=list(list(mean=0,sd=1),list(shape=2,scale=1)))
```

Os comandos a seguir servem para gerar valores aleatórios da distribuição bivariada (ou multivariada) a partir da cópula e em seguida utilizamos esses valores na função de distribuição acumulada conjunta:

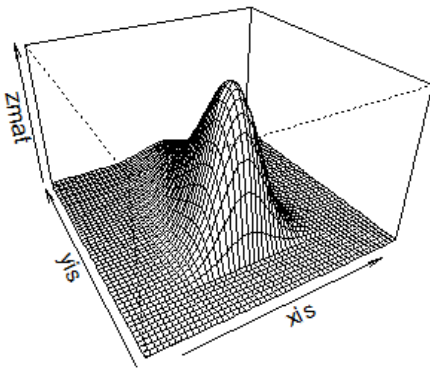
```
>x<-rMvdc(5,biv)
>x
      [,1]      [,2]
[1,] -0.2084895  1.5398263
[2,] -0.1552468  0.6294017
[3,] -0.5732844  3.1235366
[4,] -0.6690505  0.8293065
[5,]  0.7985776  2.4248667
>pMvdc(x,biv)
[1] 0.22001065 0.07177836 0.24601974 0.06649306 0.56707343
```

Existem também vários comandos que servem para realizar gráficos a partir das cópulas e das distribuições multivariadas construídas pelas cópulas. Caso tivéssemos uma distribuição trivariada gerada por uma cópula, poderíamos utilizar o pacote `scatterplot3d` para gerar, de forma simples, diagramas de dispersão.

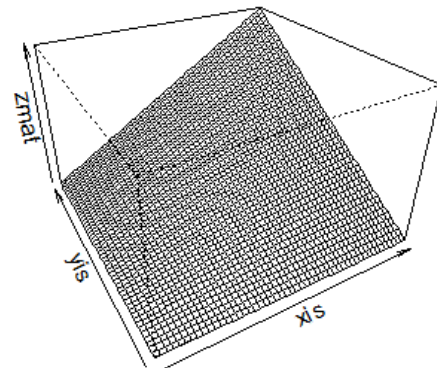
Para as cópulas e para as distribuições multivariadas podemos gerar gráficos de contorno e de perspectiva com os comandos `contour` e `persp` respectivamente:

```
>par(mfrow=c(1,2))
>persp(biv,dMvdc,xlim=c(-5,5),ylim=c(-1,5)
+,main="densidade da dist. bivariada")
>persp(franksCop,pCopula,xlim=c(0,1),ylim=c(0,1)
+,main="função de distribuição da cópula")
```

densidade da dist. bivariada

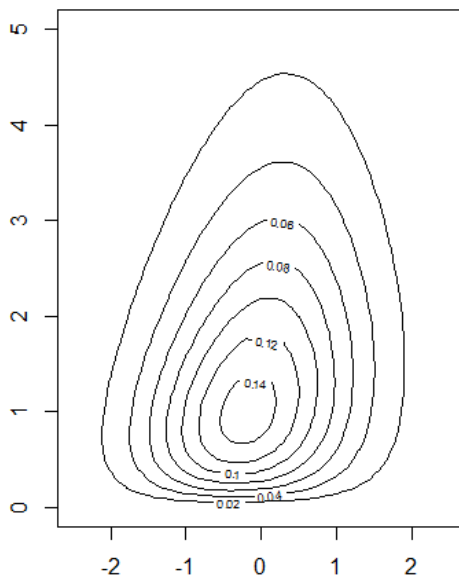


função de distribuição da cópula

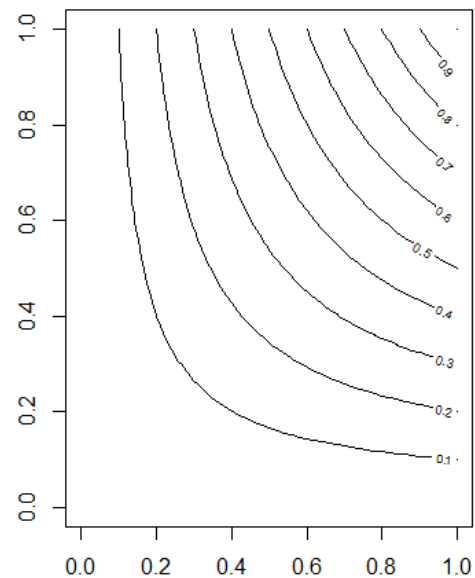


```
>par(mfrow=c(1,2))
>contour(biv,dMvdc,xlim=c(-2.5,2.5),ylim=c(0,5),
+main="Densidade da dist. bivariada")
>contour(francop,pCopula,xlim=c(0,1),ylim=c(0,1),
+main="Função de distribuição da cópula")
```

Densidade da dist. bivariada



Função de distribuição da cópula



Também existe a possibilidade de obter um ajuste para uma base de dados. A função `fitCopula`, a partir de um conjunto de observações (ou pseudo-observações) e um objeto da classe `mvdc`, nos retorna estimativas dos parâmetros, suas respectivas variâncias e a log-verossimilhança calculadas a partir dos valores estimados.

Note que o objeto da classe `mvdc` já possui em sua forma uma cópula com um respectivo parâmetro θ , porém este valor só serve de valor inicial para obter o $\hat{\theta}$.

O comando `fitCopula` pode utilizar vários argumentos para estimar os seus parâmetros. O argumento `method="ml"` utiliza o método da máxima verossimilhança, que tem pressuposto de que os dados vem de observações de uma cópula desconhecida. O método padrão, `method="mpl"` pressupõe que os dados são compostos de pseudo-observações construídas, por exemplo, a partir das distribuições marginais.

Caso escolhamos utilizar o método da máxima verossimilhança, são necessários alguns pressupostos em relação as distribuições marginais. Se as marginais forem especificadas corretamente, o estimador de máxima verossimilhança possui as propriedades ótimas usuais dos estimadores de máxima verossimilhança

O comando `fitMvdc` funciona de forma semelhante. Ele utiliza um conjunto de observações e um objeto da classe `mvdc` e nos retorna, além de outras informações, estimativas dos parâmetros das distribuições marginais e do parâmetro θ .

Os dois comandos aceitam vários métodos para realizar a otimização e o método padrão é o *Nelder-Mead*.

As funções `dCopula` e `pCopula` calculam, respectivamente, a função densidade e a função de distribuição acumulada de uma cópula em um ponto específico.

Além dessas e outras funções, o pacote `copula` também possui todas as funções que haviam sido anteriormente implementadas no pacote `nacopula`. Estas funções se referem a cópulas arquimedianas aninhadas, que são cópulas que generalizam as cópulas arquimedianas, porém conseguem permitir um pouco de assimetria.

As cópulas arquimedianas aninhadas são construídas a partir de uma combinação de outras cópulas arquimedianas. Para isso, deve-se substituir um dos argumentos de uma cópula arquimediana por uma outra cópula arquimediana. Por exemplo:

$$C(u_1, u_2, u_3) = C(u_1, C(u_2, u_3))$$

As funções mais importantes do pacote `nacopula` são as seguintes:

- `rnacopula(n,copula)`: gera n vetores aleatórios a partir da cópula em questão.
- `prob(copula,l,u)`: calcula a probabilidade de um vetor aleatório, com distribuição obtida a partir da cópula em questão, estar contido no cubo com

ponto inferior l e superior u .

- `copula@iTau(KendTau)` a partir de uma copula de classe `acopula` ,como por exemplo `copClayton`, obtemos o θ necessário para que o nosso tau de kendall seja igual a `KendTau`.
- `rho(copula)` retorna o rho de Spearman da respectiva cópula. O comando `tau(copula)` funciona de modo semelhante.

O artigo [8] também proporciona um bom entendimento do pacote `nacopula` .

A função `gofCopula` serve para verificar a qualidade do ajuste entre os dados e uma dada família de cópulas. O artigo [18] possui diretrizes em relação a este tema.

Além destes pacotes, existem vários outros pacotes no R que merecem atenção e devem ser estudados, como por exemplo:

- `CDvine`: Análise exploratória de dados bivariados. Possui métodos de amostragem e de geração de gráficos.
- `copBasic`: Possui implementadas as cópulas do produto, limite inferior e superior de *Fréchet-Hoeffding*, cópula de Plackett. Possui várias funções semelhantes as encontradas no pacote `copula` .
- `copulaedas`, `fgac`, `HAC`, `pencopula`, etc.

3.2 Seleção de cópulas

Existem vários meios para selecionar qual família de cópulas se ajusta melhor aos dados. Além disso, é virtualmente possível selecionar infinitas combinações de cópulas e distribuições marginais para escolher qual se ajusta melhor aos dados. Isto torna a escolha muito mais difícil.

Teoricamente, o primeiro passo é estimar as distribuições. Este processo pode ser realizado de diversas formas e não será tema de discussão deste estudo, visto que o nosso foco é o estudo das cópulas.

Em nosso estudo de simulação, utilizaremos o método da máxima verossimilhança para a escolha das cópulas a partir de um conjunto de famílias de um parâmetro de cópulas arquimedianas. O método da máxima verossimilhança consegue, ao mesmo tempo, estimar os parâmetros da família de cópulas selecionadas e os parâmetros associados às distribuições marginais.

Para além disso, a função `mlcbbse1` do pacote `mlCopulaSelection` no R, permite-nos calcular a máxima log-verossimilhança para um conjunto de 40 famílias de cópulas Arquimedianas.

Além do método da máxima verossimilhança, é possível utilizar outros métodos para a seleção das cópulas. O método da máxima verossimilhança as vezes requer muito poder computacional e, em algumas situações não se mostra muito eficiente.

Para estes casos, podemos utilizar outros métodos de escolha de cópulas. Tais métodos geralmente supõem que as distribuições marginais já foram estimadas. O Critério de Informação de Akaike também pode ser utilizado, em alguns casos, para selecionar a cópula. Em [9] é possível verificar uma forma de aplicar o teste de Kolmogorov-Smirnov na seleção das cópulas.

Em [29] é possível entender o funcionamento e eficiência de alguns testes paramétricos e não-paramétricos. Em [22], Joe(1997) também sugere a aplicação de diferentes métodos de seleção de cópulas e suas vantagens computacionais.

4 Métodos para geração de dados

O objetivo deste capítulo é apresentar algoritmos para simular variáveis aleatórias bivariadas a partir de cópulas.

Suponha novamente que $H(x, y) = C(F(x), G(y))$, onde F representa a distribuição marginal de X , G representa a distribuição marginal de Y . Caso C seja uma cópula Arquimediana, então existe um gerador φ tal que:

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$$

Onde φ^{-1} representa a função inversa da geradora.

4.1 Abordagem pela distribuição condicional

Um dos algoritmos mais populares para simulação de variáveis aleatórias é baseado na abordagem pela distribuição condicional. Este método pode ser utilizado para várias cópulas Arquimedianas bivariadas e seu algoritmo é baseado no seguinte:

Considere duas variáveis aleatórias uniformes U_1 e U_2 com cópula C conhecida. Assumindo condições de regularidade suficientes, que podem ser encontradas em

[10], obtemos a função de distribuição contínua :

$$C_{2|1}(u_2|u_1) = P(U_2 \leq u_2|U_1 = u_1) = \frac{\partial C(u_1, u_2)}{\partial u_1}$$

O procedimento consequente é:

1. Simule duas variáveis aleatórias $U(0, 1)$ independentes, (v_1, v_2) .
2. Faça $u_1 = v_1$.
3. Faça $u_2 = C_{2|1}^{-1}(v_2|v_1)$, onde $C_{2|1}^{-1}$ representa a pseudo-inversa de $C_{2|1}$.
Logo, o par (u_1, u_2) é constituído de variáveis aleatórias uniformes com cópula C .
4. Os valores simulados que queremos gerar são $x = F^{-1}(u_1)$ e $y = G^{-1}(u_2)$

Dessa forma, para a cópula de Clayton,

$$u_2 = [u_1^{-\theta}(v_2^{\frac{-\theta}{(1+\theta)}} - 1) + 1]^{\frac{-1}{\theta}}$$

Para a cópula de Frank,

$$u_2 = -\frac{1}{\theta} \ln \left(1 + \frac{v_2(1 - e^{-\theta})}{v_2(e^{-\theta u_1} - 1) - e^{-\theta u_1}} \right)$$

Para a cópula de Ali-Mikhail-Haq,

1. Primeiramente, faça $a = 1 - v_1$; $b = 1 - \theta(1 + 2av_2) + 2\theta^2 a^2 v_2$
e $c = 1 + \theta(2 - 4a + 4av_2) + \theta^2(1 - 4av_2 + 4a^2 v_2)$
2. Faça $u_1 = v_1$
3. E então, faça $u_2 = 2t(a\theta - 1)^2 / (b + \sqrt{c})$

No pacote `copula`, a função `rCopula` diz utilizar o método da abordagem pela distribuição condicional para gerar dados aleatórios a partir de cópulas com `dim=2`, ou seja, cópulas bivariadas.

4.2 Algoritmo para cópulas Arquimedianas em geral.

Para algumas cópulas, a distribuição condicional não é diretamente invertível. Esse é o caso das cópulas de Gumbel-Hougaard e da cópula de Joe.

Para essas cópulas podemos utilizar um outro algoritmo popular para simular variáveis aleatórias a partir de uma cópula Arquimediana. Este algoritmo é baseado no seguinte teorema:

Teorema. Sejam U_1 e U_2 variáveis aleatórias uniformes $(0,1)$ com distribuição bivariada definida por uma cópula Arquimediana com função geradora φ . Então a função de distribuição de $C(U_1, U_2)$ é dada por $K_c(t) = t - \frac{\varphi(t)}{\varphi'(t)}$. A prova deste teorema pode ser encontrada em [10].

A partir disso, o algoritmo para geração de valores aleatórios é o seguinte:

1. Simule duas variáveis aleatórias $U(0, 1)$ independentes, (v_1, v_2)
2. Faça $t = K_c^{-1}(v_2)$, onde $K_c(t) = t - \frac{\varphi(t)}{\varphi'(t)}$
3. Defina $u_1 = \varphi^{-1}(v_1\varphi(t))$ e $u_2 = \varphi^{-1}((1 - v_1)\varphi(t))$
4. Os valores simulados que queremos gerar são $x = F^{-1}(u_1)$ e $y = G^{-1}(u_2)$.

Dessa forma, para a cópula de Gumbel-Hougaard,

$$u_1 = \exp[v_1^{\frac{1}{\theta}} \ln(w)]; u_2 = \exp[(1 - v_1^{\frac{1}{\theta}}) \ln(w)],$$

w é definido de forma a solucionar $K_c(w) = w \left(1 - \frac{\ln(w)}{\theta}\right) = v_2, 0 < w < 1$.

Para a cópula de Joe,

$$u_1 = 1 - [1 - [1 - (1 - w)^{\theta}]^{v_1}]^{\frac{1}{\theta}}; u_2 = 1 - [1 - [1 - (1 - w)^{\theta}]^{1-v_1}]^{\frac{1}{\theta}},$$

onde w é definido de forma a solucionar:

$$K_c(w) = w - \frac{[\ln(1 - (1 - w)^{\theta})] [(1 - (1 - w)^{\theta})]}{\theta(1 - w)^{\theta-1}} = v_2, 0 < w < 1.$$

4.3 Abordagem de Marshall e Olkin

Algumas cópulas caem em uma classe conhecida como *LT-Archimedean copulas* e possuem uma representação definida a partir de transformadas de Laplace. Este é o caso das cópulas de Joe, Frank, Clayton e Gumbel-Hougaard.

Para as *LT-Archimedean copulas*, a inversa da função geradora φ possui uma representação a partir de uma transformada de Laplace de uma função G , isto é,

$$\varphi = \mathcal{L}(G).$$

Infelizmente, não existe um método conhecido para encontrar esta função G explicitamente.

O algoritmo baseado na representação a partir de transformadas de Laplace é dado por:

1. Crie uma variável $V \sim G = \mathcal{L}^{-1}(\varphi)$, com a transformada de Laplace de G dada por $\psi(t) = \int_0^\infty e^{tx} dG(x), t \geq 0$
2. Simule duas variáveis aleatórias $U(0, 1)$ independentes, (v_1, v_2)
3. Faça $u_i = \psi(-\frac{\ln(v_i)}{V}), i = 1, 2$
4. Então o vetor (u_1, u_2) possui distribuição associada a cópula Arquimediana com função geradora $\varphi = \psi^{-1}$

Para a cópula de Clayton,

- $V \sim \text{Gamma}(\frac{1}{\theta}, 1)$.
- $\psi(t) = (1 + t)^{-1/\theta}$.

Para a cópula de Gumbel-Hougaard,

- V segue distribuição estável dada por $St(\frac{1}{\theta}, 1, (\cos(\frac{\pi}{2\theta}))^\theta, 0; 1)$.
- $\psi(t) = \exp(-t^{1/\theta})$.

Para a cópula de Frank,

- V é discreta com $P(V = k) = (1 - e^{-\theta})^k / (k\theta)$.
- $\psi(t) = -\frac{1}{\theta} \ln[1 + e^{-t}(e^{-\theta} - 1)]$.

Para a cópula de Ali-Mikhail-Haq,

- V é discreta com $P(V = k) = (1 - \theta)(\theta)^{k-1}$.
- $\psi(t) = \frac{1-\theta}{e^t - \theta}$

Para a cópula de Joe,

- V é discreta com $P(V = k) = (-1)^{k+1} \binom{1}{k}$.
- $\psi(t) = 1 - (1 - e^{-t})^{1/\theta}$

Este método funciona também para cópulas de dimensão superior a 2 e é o método utilizado pelo comando `rCopula` do pacote `copula`, de acordo com os escritores do pacote.

5 Estudo de simulação

No presente capítulo iremos realizar algumas simulações para testar a eficiência das estimativas dos parâmetros e da seleção de modelos baseados em cópulas.

Faremos várias simulações com dados gerados a partir de uma cópula e marginais previamente definidas.

A partir daí, iremos ajustar vários modelos baseados em cópulas a partir da família de cópulas de Clayton, Joe, Frank e Gumbel-Hougaard e selecionaremos o melhor modelo a partir do método da máxima verossimilhança. Note que os métodos implementados para estimação de modelos baseados em cópulas no pacote `copula` utilizam métodos iterativos para estimar os parâmetros, e nem sempre esses métodos são convergentes. Isto faz com que, para alguns conjuntos de dados, tenhamos falhas na estimação dos modelos para algumas cópulas. Estas falhas podem ser reduzidas com alguns procedimentos.

Alguns dos possíveis meios de reduzir estas falhas são os seguintes:

- Modificar o método de otimização em `fitMvdc(...,method=c(),...)` .
- Em `fitMvdc(...,optim.control=list(),...)` , Modificar os argumentos de controle do processo iterativo.
- Em `fitMvdc(...,start=c(),...)` fazer alguns cálculos para estimar os valores iniciais do processo iterativo.

Por exemplo, se uma das marginais segue uma distribuição Normal de parâmetros desconhecidos, calcular a média e a variância da coluna de observações respectiva desta marginal.

Estes e outros erros que podem ser encontrados no pacote `copula` são tratados e parcialmente esclarecidos no artigo [17].

Para as simulações abaixo, poderíamos ter utilizado uma maior quantidade de famílias de cópulas e de marginais, mas por questão de praticidade, utilizaremos apenas as cópulas arquimedianas de Clayton, Joe, Frank e Gumbel.

Nas tabelas abaixo, o item “n° de seleções” representa a quantidade de vezes que a cópula obteve a maior verossimilhança dentro das 4 possibilidades.

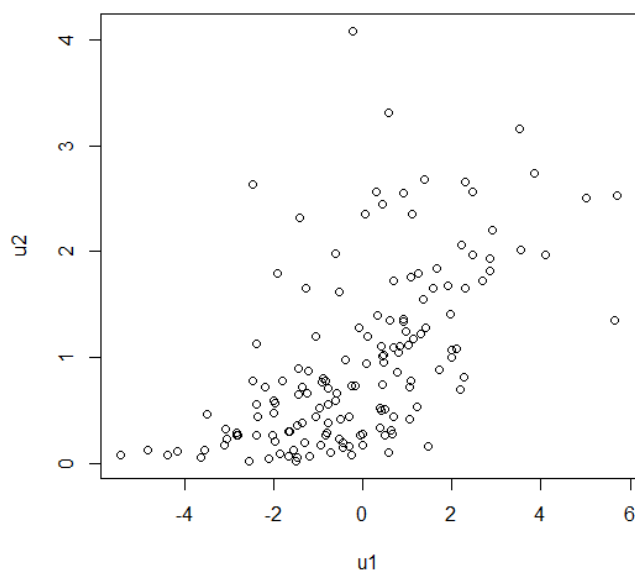
O item “ $\theta \in \text{int. de conf de } \hat{\theta}$ ” representa a quantidade de vezes em que o nosso parâmetro θ real estava dentro do intervalo de confiança estimado.

Cento e cinquenta observações geradas por uma cópula de Frank de $\theta = 5.736283$, tau de Kendal=0.5 e marginais $Normal(0, 2)$ e $Gama(1, 1)$ em 500 simulações. Intervalo de confiança gerado com $\alpha = 0.05$.

	Clayton	Joe	Frank	Gumbel
n° de seleções	0	0	489	11
Percentagem de seleções	0	0	97,8%	2,2%
$\theta \in$ int. de conf de $\hat{\theta}$			500	
Percentagem de $\theta \in$ int. de conf de $\hat{\theta}$			100%	

Uma das 500 simulações está representada no gráfico abaixo.

150 observações geradas por uma cópula de Frank

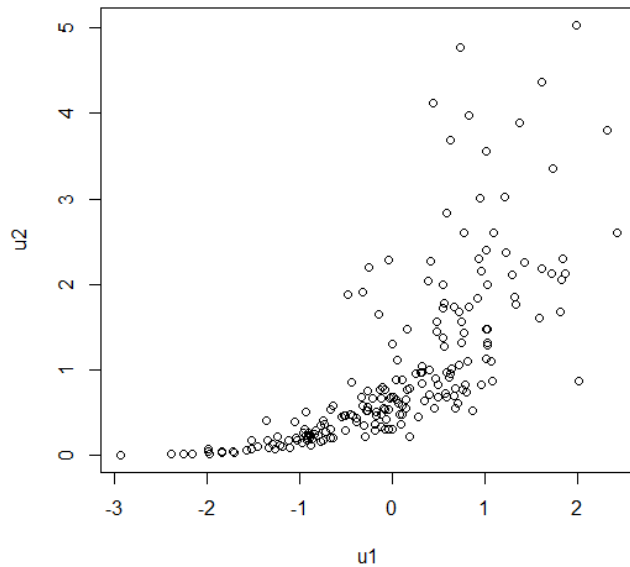


Duzentas observações geradas por uma cópula de Clayton de $\theta = 4.667$, tau de Kendall=0.7 e marginais $Normal(0, 1)$ e $Exponencial(1)$ em 1000 simulações. Intervalo de confiança gerado com $\alpha = 0.05$

	Clayton	Joe	Frank	Gumbel
n° de seleções	1000	0	0	0
Percentagem de seleções	100%	0	0	0
$\theta \in$ int. de conf de $\hat{\theta}$	1000			
Percentagem de $\theta \in$ int. de conf de $\hat{\theta}$	100%			

Uma das 1000 simulações está representada no gráfico abaixo.

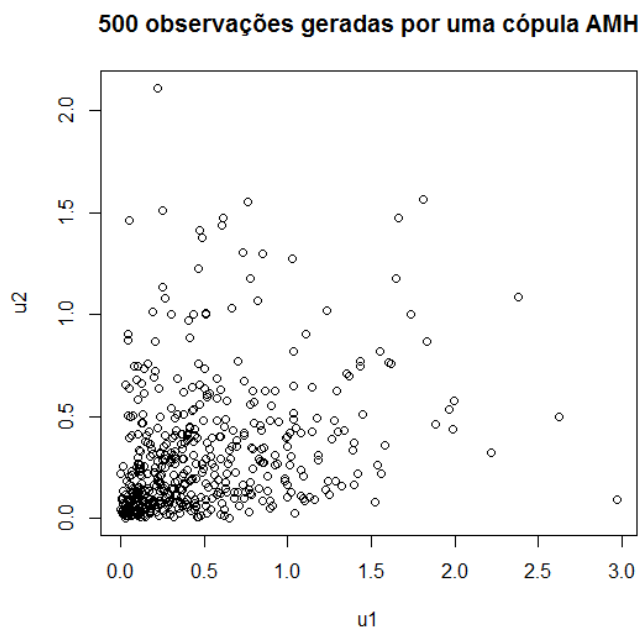
200 observações geradas a partir da cópula de Clayton



Quinhentas observações gerados por uma cópula da família Ali-Mikhail-Haq de $\theta = 0.8384794$, tau de Kendall=0.25 e marginais *Exponencial*(2) e *Exponencial*(3) em 500 simulações. Intervalo de confiança gerado com $\alpha = 0.05$

	Clayton	Joe	Frank	Gumbel	AMH
nº de seleções	34	0	25	0	441
Percentagem de seleções	6,8%	0	5%	0	88,2%
$\theta \in \text{int. de conf de } \hat{\theta}$					500
Percentagem de $\theta \in \text{int. de conf de } \hat{\theta}$					100%

Uma das 500 simulações está representada no gráfico abaixo.



Além destes tipos de simulação, é possível fazer o mesmo tipo de ajuste, mas sem tomar em consideração as distribuições marginais associadas a cópula. Pode ser omitida a distribuição marginal na geração de dados ou no ajuste, ou até em ambos.

Neste estudo, omitiremos as distribuições marginais tanto na geração de dados quanto no ajuste. O estudo será semelhante ao realizado anteriormente.

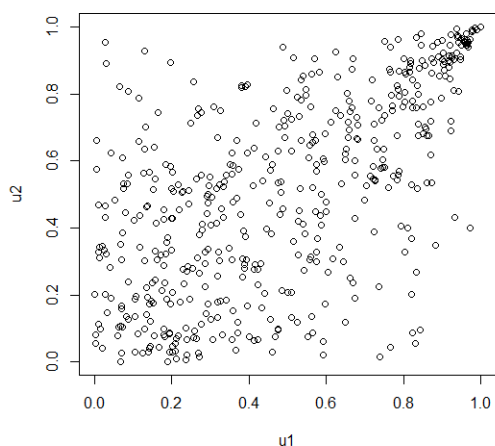
Assim como foi citado anteriormente quando foi falado do pacote `copula`, é possível realizar um ajuste de uma cópula a partir de observações ou de pseudo-observações. Apenas 1 dos 4 métodos existentes na função `fitCopula` utiliza observações verdadeiras, e não pseudo-observações, o método da máxima verossimilhança.

Cinco mil observações gerados por uma cópula da família Joe de $\theta = 2.219066$, tau de Kendall=0.4 em 5000 simulações. Intervalo de confiança gerado com $\alpha = 0.05$. O método utilizado aqui foi o da máxima verossimilhança

	Clayton	Joe	Frank	Gumbel
n° de seleções	0	4952	0	58
Percentagem de seleções	0	98,84%	0	1,16%
$\theta \in \text{int. de conf de } \hat{\theta}$		5000		
Percentagem de $\theta \in \text{int. de conf de } \hat{\theta}$		100%		

Uma das 5000 simulações está representada no gráfico abaixo.

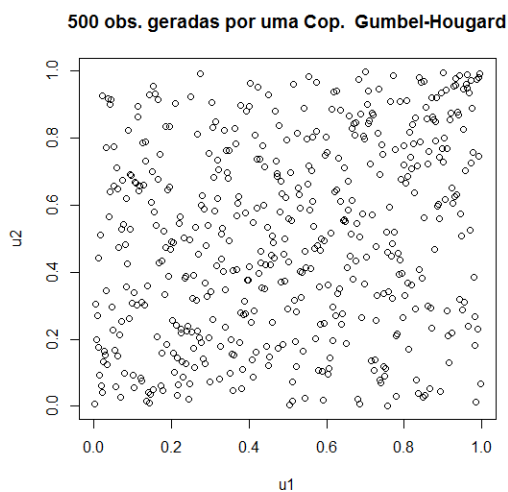
500 observações geradas por uma cópula de Joe



Quinhentas observações gerados por uma cópula da família Gumbel-Hougaard de $\theta = 1.25$, tau de Kendall=0.2 em 5000 simulações. Intervalo de confiança gerado com $\alpha = 0.05$. O método utilizado aqui foi o da máxima pseudo-verossimilhança.

	Clayton	AMH	Frank	Gumbel
n° de seleções	1	13	141	4845
Percentagem de seleções	0,02%	0,26%	2,82%	96,9%
$\theta \in \text{int. de conf de } \hat{\theta}$				5000
Percentagem de $\theta \in \text{int. de conf de } \hat{\theta}$				100%

Uma das 5000 simulações está representada no gráfico abaixo.



Note que em todas as simulações realizadas, as estimativas do R tiveram "sucesso" em indicar a qual família de cópulas os dados melhor se adaptavam em mais de 85% dos casos. Além disso, e mais importantemente, em todas as simulações nos obtivemos 100% dos parâmetros reais dentro do intervalo de confiança estimado para cada uma das simulações

6 Conclusões

As cópulas são funções que conectam as distribuições marginais e as distribuições conjuntas. Elas são ferramentas boas para construção de distribuições multivariadas e para estudar as relações de dependência das distribuições.

A utilização das funções de cópulas está sendo cada vez maior em tópicos como Finanças, Ciências Atuariais e Econometria. Neste texto, disponibilizamos informações básicas, porém suficientes para realizar estudos semelhantes.

Nosso estudo de simulação atingiu bons resultados a partir da estimação baseada na verossimilhança. Houve excelência na estimativa dos parâmetros. Como verificado, 100% dos parâmetros reais estavam dentro do intervalo de confiança estimado.

Todo esse estudo de simulação também pode ser prolongado para o caso multivariado, porém, quanto maior a dimensão da cópula, maior será a dificuldade de estimar os parâmetros associados a esta.

As funcionalidades das cópulas em softwares open-source está cada vez mais

desenvolvida, com um vasto acervo de pacotes livres para serem utilizados.

Em trabalhos futuros queremos introduzir a censura nos dados de análise de sobrevivência e tratar da aplicação da teoria de cópulas em dados relativos a seguradoras e às ciências atuariais. Além disso, existem diversos problemas abertos relacionados às cópulas que devem ser estudados e solucionados.

A Um exemplo utilizado na simulação.

No R:

```
library(copula)
###Cópula de Clayton de marginais Normal e exponencial.

theta<-copClayton@iTau(0.7)
ClaytonCop<-archmCopula(family="clayton",dim=2,param=theta)
contagem<-c(0,0,0,0)
contagem2<-c(0)

bivClayton<-mvdc(copula=ClaytonCop,margins=c("norm","exp")
,paramMargins=list(list(mean=0,sd=1),list(rate=1)))
bivJoe<-mvdc(copula=archmCopula(family="joe",dim=2,param=1),
margins=c("norm","exp")
,paramMargins=list(list(mean=0,sd=1),list(rate=1)))
bivFrank<-mvdc(copula=archmCopula(family="frank",dim=2,param=1),
margins=c("norm","exp")
,paramMargins=list(list(mean=0,sd=1),list(rate=1)))
bivGumbelHougard<-mvdc(copula=archmCopula(family="Gumbel",dim=2,
param=1),margins=c("norm","exp")
,paramMargins=list(list(mean=0,sd=1),list(rate=1)))
bivAMH<-mvdc(copula=archmCopula(family="amh",dim=2,param=0.5),
margins=c("norm","exp")
,paramMargins=list(list(mean=0,sd=1),list(rate=1)))

for(i in 1:500){
dados<-rMvdc(200,bivClayton)
fitClayton<-fitMvdc(dados,bivClayton,c(1,1,1,1))
fitJoe<-fitMvdc(dados,bivJoe,c(1,1,1,2))
fitFrank<-fitMvdc(dados,bivFrank,c(1,1,1,1))
fitGumbelHougard<-fitMvdc(dados,bivGumbelHougard,c(1,1,1,2))
#fitAMH<-fitMvdc(dados,bivAMH,c(1,1,1,0.1))

verossimi<-c(0,0,0,0)
verossimi[1]<-fitClayton@loglik
verossimi[2]<-fitJoe@loglik
verossimi[3]<-fitFrank@loglik
verossimi[4]<-fitGumbelHougard@loglik
```

```
#verossimi[5]<-fitAMH@loglik
if (max(verossimi)==verossimi[1]) {contagem[1]<-contagem[1]+1}
if (max(verossimi)==verossimi[2]) {contagem[2]<-contagem[2]+1}
if (max(verossimi)==verossimi[3]) {contagem[3]<-contagem[3]+1}
if (max(verossimi)==verossimi[4]) {contagem[4]<-contagem[4]+1}
#if (max(verossimi)==verossimi[5]) {contagem[5]<-contagem[5]+1}
if ((fitClayton@estimate[4]<confint(fitClayton,"param")[2])&
(fitClayton@estimate[4]>confint(fitClayton,"param")[1]))
{contagem2<-contagem2+1}
}
```

Referências

- [1] Assis, R. and Laurini, M. (2008), Funções de Cópula na Precificação de Opções, Insper Working Papers, Insper Working Paper, Insper Instituto de Ensino e Pesquisa.
- [2] Balakrishnan, N. and Lai, C. Continuous bivariate distributions. Springer Science & Business Media, 2009.
- [3] Chen, Y. and Tu, A., (2013), Estimating hedged portfolio value-at-risk using the conditional copula: An illustration of model risk, *International Review of Economics & Finance*, 27, issue C, p. 514-528.
- [4] Cherubini, U., Luciano E. and Vecchiato, W.(2004). *Copula Methods in Finance*.
- [5] Emura, Takeshi, Lin, Chien-Wei and Wang(2010), A goodness-of-fit test for Archimedean copula models in the presence of right censoring, *Computational Statistics & Data Analysis*, 54, issue 12, p. 3033-3043.
- [6] Fréchet, M.(1951), Sur les tableaux de corrélation dont les marges sont données. *Ann Univ Lyon, Sect A* 9:53-77
- [7] Frees, E.W. and Wang, P. (2005).Credibility Using Copulas. *North American Actuarial Journal*, 9(2), 31-48.
- [8] Frees, E.W. and Valdez, E.A. Understanding Relationships Using Copulas.
- [9] Frees, E.W., Carriere, J. and Valdez, E. Annuity valuation with dependent mortality, *J. Risk Insurance*, 63 (1996), pp. 229-261.
- [10] Genest, C. and Neslehova, J. 2007.A primer on copulas for count data *Astin Bulletin*, Vol. 37.
- [11] Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292), 698-707.
- [12] Hoeffding, W.(1994) Scale-Invariant Correlation Theory. In: *The collected works of Wassily Hoeffding*. Springer New York, p. 57-107.
- [13] Hofert, M. and Maechler, M.(2011), Nested Archimedean Copulas Meet R: The nacopula Package, *Journal of Statistical Software*, 39, issue i09.
- [14] Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

- [15] Kakouris, I. and Rustem, B.(2014) Robust portfolio optimization with copulas. *European Journal of Operational Research*, v. 235, n. 1, p. 28-37.
- [16] Kole, E., Koedijk, K. and Verbeek, M. Selecting copulas for risk management. *Journal of Banking & Finance*, v. 31, n. 8, p. 2405-2423, 2007.
- [17] Mariana, F.(2011). *Aplicação de cópulas na modelação do número de sinistros de grupos de risco homogéneos no seguro automóvel* [Lisboa]: [s.n.].
- [18] Marshall, A.and Olkin, I. 1967. "A Multivariate Exponential Distribution," *Journal of the American Statistical Association* 62:30-44.
- [19] Münnix, Michael C. and Schäfer, R. (2011), A copula approach on the dynamics of statistical dependencies in the US stock market, *Physica A: Statistical Mechanics and its Applications*, 390, issue 23, p. 4251-4259.
- [20] Nelsen, R (2006), *An Introduction to Copulas* .
- [21] Nelsen, R. *Properties and applications of copulas: A Brief Survey*.
- [22] Palaro, A., Helder P. and Hotta, L.(2006) Using conditional copula to estimate value at risk. *Journal of Data Science*, v. 4, p. 93-115.
- [23] Pierre, G.(2001). *Multivariate survival modelling: A unified approach with copulas*.
- [24] Righi, M. and Ceretta, P.(2013), Analyzing the dependence structure of various sectors in the Brazilian market: A Pair Copula Construction approach, *Economic Modelling*, 35, issue C, p. 199-206.
- [25] Romeo, J., Tanaka, N. and Pedroso, A.(2006) Bivariate survival modeling: a Bayesian approach based on Copulas. *Lifetime Data Analysis*, 12, 205-222.
- [26] Shemyakiny, E. and Youn, H. (2006), Copula models of joint last survivor analysis. *Appl. Stochastic Models Bus. Ind.*, 22: 211-224.
- [27] Shemyakiny, E. and Youn H., *Copula Models of Joint Survival Analysis*.
- [28] Sklar, A.(1959)"Fonctions de répartition à n dimensions et leurs marges", *Publ. Inst. Statist. Univ. Paris* (in French) 8: 229-231.
- [29] Staudt, A. "Tail risk, systemic risk and copulas." *Casualty Actuarial Society E-Forum*, Fall 2010-Volume 2. 2010.

- [30] Swanepoel, J. and Allison, J. (2013), Some new results on the empirical copula estimator with applications, *Statistics & Probability Letters*, 83, issue 7, p. 1731-1739
- [31] Trivedi P. and Zimmer D.(2005), *Copula Modeling:An Introduction for Practitioners*.
- [32] Xubiao, E. and Gong, P.(2009) Measuring the coupled risks: A copula-based CVaR model. *Journal of Computational and Applied Mathematics*, v. 223, n. 2, p. 1066-1080.
- [33] Yan, J. (2007). Enjoy the Joy of Copulas: With a Package copula. *Journal of Statistical Software*, 21(4), 1-21
- [34] Yan, J. and Kojadinovic, I.(2010), Modeling Multivariate Distributions with Continuous Margins Using the copula R Package, *Journal of Statistical Software*, 34, issue 9.
- [35] Zhao, X. and Zhou, X.(2010).”Applying copula models to individual claim loss reserving methods,”*Insurance: Mathematics and Economics*, Elsevier, vol. 46(2), pages 290-299, April.
- [36] Zongwu, C.(2008) Selection of copulas with applications in finance. Unpublished manuscript, University of North Carolina at Charlotte, Charlotte, NC.