Copulas: An Introduction
I - Fundamentals

Johan Segers

Université catholique de Louvain (BE)
Institut de statistique, biostatistique et sciences actuarielles

Columbia University, New York City
9–11 Oct 2013

The starting point: Margins versus dependence

Decomposition of a multivariate cdf F into

- univariate margins $F_1, \ldots, F_d$
- copula C

Idea: the copula C captures the dependence among the d variables, irrespective of their marginal distributions.

Course aim

Introduction to the basic concepts and main principles

I Fundamentals

II Models

III Inference

Caveats:

- Personal selection of topics in a wide and fast-growing field
- Speaker's bias towards (practically useful) theory
- References are a random selection from an ocean of literature

Some references to start with

Jaworski, P., F. Durante, W. Härdle, and T. Rychlik (2010). *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*. Lecture Notes in Statistics. Berlin: Springer.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Kojadinovic, I. and J. Yan (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software 34*(9), 1–20.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton: Princeton University Press. Chapter 5, "Copulas and Dependence".

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer.

Trivedi, P. K. and D. M. Zimmer (2005). Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics 1*(1), 1–111.

+ books on the use of copulas in specific domains, notably finance

Copulas: An Introduction
I - Fundamentals

Copulas: An Introduction
I - Fundamentals

Sklar's theorem

Densities and conditional distributions

Copulas for discrete variables

Measures of association

Generalized inverse functions

The left-continuous generalized inverse function of a univariate cdf F is defined as

$$F^{\leftarrow}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}, \qquad 0 < u < 1$$

Ex. Make a picture of $F^{\leftarrow}(u) = x$ in case

1. F is continuous and increasing in x
2. F is continuous but flat in x
3. F has an atom at x

Ex. Work out $F^{\leftarrow}$ if F is the cdf of a rv X with $\mathrm{P}(X = 1) = p = 1 - \mathrm{P}(X = 0)$.

Properties of generalized inverse functions

Let F be a univariate cdf, not necessarily continuous.

- $F(F^{\leftarrow}(u)) \geq u$
- $F(x) \geq u$ iff $x \geq F^{\leftarrow}(u)$
- If U is uniform $(0, 1)$, then $X = F^{\leftarrow}(U)$ has cdf F.

Ex. Prove these properties.
[Hint: F is right continuous.]

Ex. How would the second result help you to generate random numbers from F?

Probability integral transform: Reduction to uniformity

If X is a random variable with continuous cdf F, then the distribution of $U = F(X)$ is $\mathrm{Uniform}(0,1)$, i.e.

$$\mathrm{P}[F(X) \le u] = u, \qquad u \in [0,1]$$

Ex. What goes wrong if F is not continuous? Take for instance X Bernoulli(p).

Ex. Prove the above property.
[Hint: Justify the equalities in
$\mathrm{P}[F(X) \ge u] = \mathrm{P}[X \ge F^{\leftarrow}(u)] = 1 - F(F^{\leftarrow}(u)) = 1 - u$.]

Ex. Generate a pseudo-random sample $X_1, \ldots, X_n$ from your favourite continuous distribution F. Compute $F(X_1), \ldots, F(X_n)$ and assess its 'uniformity' (e.g. histogram, kernel density estimate, QQ-plot, ...).

So what's a copula?

A d-variate copula $C : [0,1]^d \to [0,1]$ is the cdf of a random vector $(U_1, \ldots, U_d)$ with Uniform$(0,1)$ margins:

$$C(\boldsymbol{u}) = \mathrm{P}[U_1 \le u_1, \ldots, U_d \le u_d]$$

where

$$\mathrm{P}[U_j \le u_j] = u_j$$

for $j \in \{1, \ldots, d\}$ and $0 \le u_j \le 1$.

Remark: Alternative definition possible, in terms of properties of C as a function.

The representation of a copula as a cdf implies a number of properties

$$C(\boldsymbol{u}) = \mathrm{P}[U_1 \le u_1, \ldots, U_d \le u_d], \qquad U_j \sim \text{Uniform}(0,1)$$

1. If some component u_j is 0, then $C(\boldsymbol{u}) = 0$.
2. $C(1, \ldots, 1, u_j, 1, \ldots, 1) = u_j$ if $0 \le u_j \le 1$.
3. C is d-increasing, e.g. if $d = 2$ and $a_j \le b_j$,

$$0 \le C(b_1, b_2) - C(a_1, b_2) - C(b_1, a_2) + C(a_1, a_2)$$

4. C is nondecreasing in each of the d variables.
5. C is Lipschitz and hence continuous:

$$|C(\boldsymbol{u}) - C(\boldsymbol{v})| \le |u_1 - v_1| + \cdots + |u_d - v_d|$$

Ex. Prove these properties.

Sklar's theorem I: How to construct a multivariate cdf

Let C be a d-variate copula and let $F_1, \ldots, F_d$ be univariate cdf's. Then the function

$$F(\boldsymbol{x}) = C\big(F_1(x_1), \ldots, F_d(x_d)\big) \tag{Skl}$$

is a d-variate cdf with margins $F_1, \ldots, F_d$.

Proof.

Let $(U_1, \ldots, U_d) \sim C$ and put

$$X_j = F_j^{\leftarrow}(U_j) \sim F_j.$$

Then $\boldsymbol{X} \sim F$. □

Sklar's theorem II: Any multivariate cdf has a copula

If F is a d-variate cdf with univariate cdf's $F_1, \ldots, F_d$, then there exists a copula C such that (Skl) holds.

If the margins are continuous, then C is unique and is equal to

$$C(\boldsymbol{u}) = F\big(F_1^{\leftarrow}(u_1), \ldots, F_d^{\leftarrow}(u_d)\big)$$

Proof.

Assume the margins are continuous. Let $\boldsymbol{X} \sim F$ and put

$$U_j = F_j(X_j) \sim \text{Uniform}(0, 1).$$

Then $\boldsymbol{U} \sim C$ with C as given in the display, and (Skl) holds. $\square$

Elementary examples

Let (X, Y) be a random vector with continuous margins and copula C.

- X and Y are independent if and only if their copula is

$$C(u, v) = uv$$

- If $Y = g(X)$ with g increasing, then

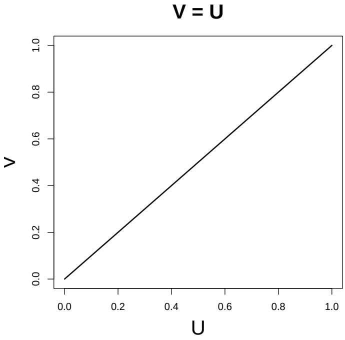
$$C(u, v) = \min(u, v) =: M(u, v)$$

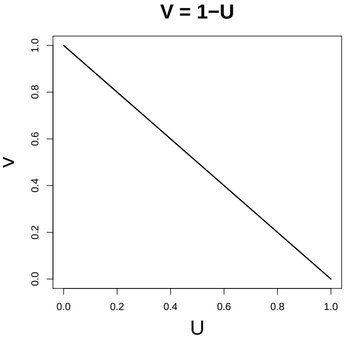
- If $Y = g(X)$ with g decreasing, then

$$C(u, v) = \max(u + v - 1, 0) =: W(u, v)$$

Ex.
1. Show the above relations.
2. Show that M is the cdf of (U, U). What is its support?
3. Show that W is the cdf of $(U, 1 - U)$. What is its support?

Fréchet–Hoeffding upper and lower bounds: Supported on the (anti)diagonal





$$M(u,v) = \min(u,v) \qquad W(u,v) = \max(u+v-1,0)$$

Fréchet–Hoeffding bounds

Any bivariate copula C verifies

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

Ex. Show these inequalities.
Hint: use the Bonferroni inequalities

$$\mathrm{P}(A)+\mathrm{P}(B)-1 \le \mathrm{P}(A\cap B) \le \min\{\mathrm{P}(A),\mathrm{P}(B)\}$$

Ex. Extend the bounds to d-variate copulas.

- The upper bound is the copula of the random vector $(U,\ldots,U)$.
- The lower bound is not a copula if $d \ge 3$.

Invariance under monotone transformations

If

- C is a copula of $\boldsymbol{X} \sim F$
- $T_1, \ldots, T_d$ are increasing functions

then

- C is also a copula of $(T_1(X_1), \ldots, T_d(X_d))$

Ex. Show the above property.
[Hint: the cdf of $T_j(X_j)$ is $F_j(T_j^{-1})$. Calculate the joint cdf of $(T_1(X_1), \ldots, T_d(X_d))$, using Sklar's representation of F.]

Survival copulas: Linking joint and marginal survival functions

Assume continuous margins. If $\boldsymbol{X} = (X_1, \ldots, X_d)$ and $U_j = F_j(X_j)$, then $1 - U_j$ is uniform on $(0, 1)$ too.

The cdf $\bar{C}$ of $(1 - U_1, \ldots, 1 - U_d)$ is the survival copula of $\boldsymbol{X}$, and

$$\mathrm{P}[X_1 > x_1, \ldots, X_d > x_d] = \bar{C}\big(\bar{F}_1(x_1), \ldots, \bar{F}_d(x_d)\big)$$

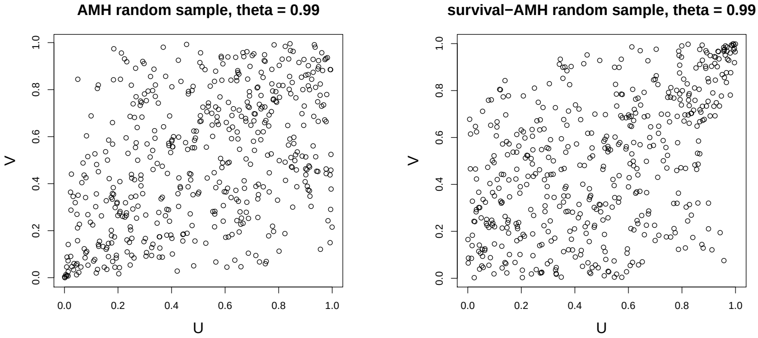
linking the joint survival function with the marginal ones,

$$\bar{F}_j(x_j) = 1 - F_j(x_j) = \mathrm{P}[X_j > x_j]$$

This way of modelling dependence is popular in survival analysis.

Example: the Ali–Mikhail–Haq (survival) copula

$$C_\theta(u, v) = \frac{uv}{1 - \theta\,(1-u)\,(1-v)}, \qquad \theta \in [-1, 1)$$



Survival copulas are copulas too

Ex. In dimension $d = 2$, show that

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

Ex. Show that if C is the copula of $(X_1, \ldots, X_d)$,
then $\bar{C}$ is the copula of $(-X_1, \ldots, -X_d)$,
or more generally of $(T_1(X_1), \ldots, T_d(X_d))$ for *decreasing* functions T_j.

Ex. If $(U, V) \sim C$, calculate the cdf's (copulas) of $(1 - U, V)$ and $(U, 1 - V)$.
More generally, to a d-variate copula C, one can associate 2^d copulas by considering transformations $(T_1, \ldots, T_d)$ with T_j in/de-creasing.

Symmetries

Let $\boldsymbol{U} \sim C$.

The copula C is called symmetric or exchangeable
if, for any permutation, σ, of $\{1, \ldots, d\}$,

$$(U_{\sigma(1)}, \ldots, U_{\sigma(d)}) \stackrel{d}{=} (U_1, \ldots, U_d)$$

The copula C is called radially symmetric if $\bar{C} = C$:

$$(1 - U_1, \ldots, 1 - U_d) \stackrel{d}{=} (U_1, \ldots, U_d)$$

Presence or absence of certain symmetries can be a guide towards model selection.

Example: the Plackett copula is (radially) symmetric

The *Plackett* copula arises in the study of 2×2 contingency tables.

	$U \leq u$	$U > u$
$V \leq v$	$C(u,v)$	$v - C(u,v)$
$V > v$	$u - C(u,v)$	$1 - u - v + C(u,v)$

$C_\theta(u,v)$ is defined as the smaller one of the two roots of the equation

$$\text{odds ratio } \theta = \frac{C_\theta(u,v)\,\{1 - u - v + C_\theta(u,v)\}}{\{u - C_\theta(u,v)\}\,\{v - C_\theta(u,v)\}} \in (0, \infty)$$

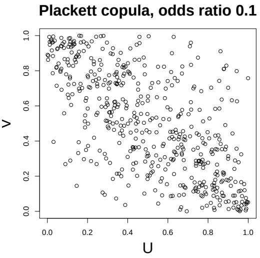
Ex. Show that the Plackett copula is both

- exchangeable
- radially symmetric

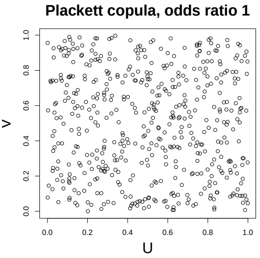
[Hint: either solve for $C_\theta(u,v)$ and verify the two symmetries by computation, or prove the two properties from inspecting the equation.]

Random samples from the Plackett copula

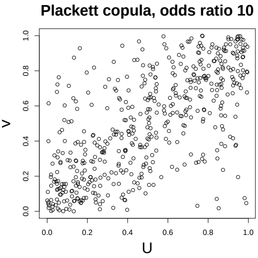
Random sample of size 500 from C_θ



$\theta = 0.1$



$\theta = 1$



$\theta = 10$

Sklar's theorem and weak convergence

Let $F_n(\boldsymbol{x}) = C_n(F_{n,1}(x_1), \ldots, F_{n,d}(x_d))$ and similarly for F. Assume continuous margins. Then

$$\begin{aligned} & F_n(\boldsymbol{x}) \to F(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \\ \iff & \begin{cases} C_n(\boldsymbol{u}) & \to C(\boldsymbol{u}) \qquad \forall \boldsymbol{u} \\ F_{n,j}(x_j) & \to F_j(x_j) \qquad \forall j, \forall x_j \end{cases} \end{aligned}$$

Proof.

- $\Rightarrow$ Continuous mapping theorem, uniform convergence to continuous limits.
- $\Leftarrow$ Uniform convergence to continuous limits.

□

Example: the sample maximum and minimum

Let $X_1, X_2, \ldots$ be iid with continuous distribution F. The copula of

$$\big(\max(X_1, \ldots, X_n), -\min(X_1, \ldots, X_n)\big)$$

is given by the *Clayton* copula with parameter $\theta = -1/n$

$$C_n(u, v) = \max(u^{1/n} + v^{1/n} - 1, 0)^n \qquad \text{(MaxMin)}$$

Ex. Show (MaxMin).
[Hint: $-\min(x_1, \ldots, x_n) = \max(-x_1, \ldots, -x_n)$.]

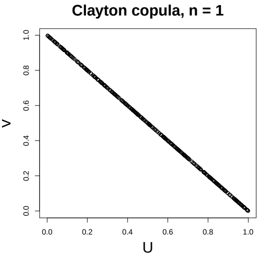
Ex. Show that

$$\lim_{n \to \infty} C_n(u, v) = uv$$

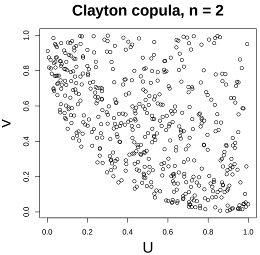
The sample maximum and minimum are 'asymptotically independent'.
[Hint: $n(u^{1/n} - 1) \to \log(u)$ and $(1 + x/n)^n \to e^x$.]

Random samples from the Clayton copula

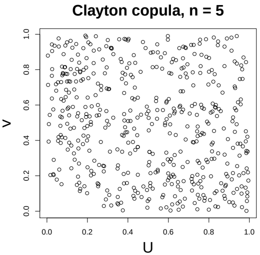
Random sample of size 500 from C_n



$n = 1$



$n = 2$



$n = 5$

Sklar's theorem: Some literature

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer. Chapter 2.

Ruschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference 139*, 3921–3927.

Sklar, A. (1959). Fonctions de répartition à *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–331.

Copulas: An Introduction
I - Fundamentals

Copula density

A copula C being a multivariate cdf, its density c, if it exists, is just

$$c(\boldsymbol{u}) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C(\boldsymbol{u})$$

Ex. Recall the Clayton copula C_n in (MaxMin).

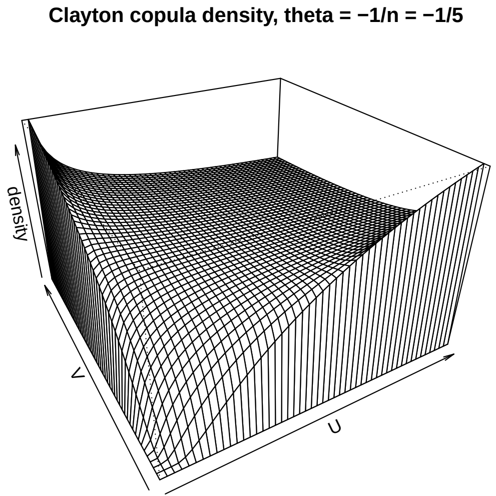
- Compute its density c_n.
- Show analytically or graphically that $c_n(u, v) \to 1$ as $n \to \infty$.

Ex. Compute the density of the *Gumbel–Hougaard* copula:

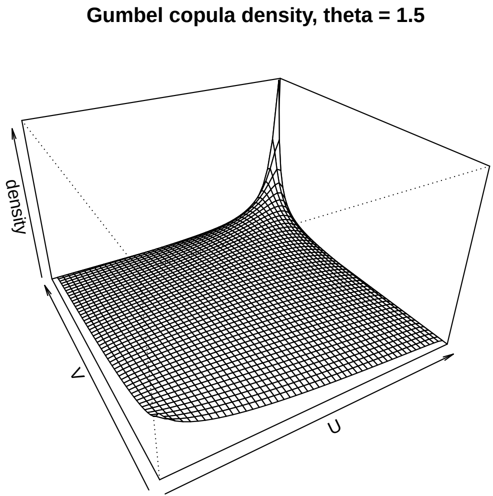
$$C(\boldsymbol{u}) = \exp\left[-\left\{(-\log u_1)^\theta + \cdots + (-\log u_d)^\theta\right\}^{1/\theta}\right], \qquad \theta \geq 1$$

Up to which d do you get?

Density of the Clayton copula



Density of the Gumbel-Hougaard copula



The joint density of a multivariate cdf factors into the marginal densities and the copula density

If the margins of F admit densities $f_1, \ldots, f_d$ and if the copula C admits a density c, then F admits a joint density

$$f(\boldsymbol{x}) = c\big(F_1(x_1), \ldots, F_d(x_d)\big) f_1(x_1) \cdots f_d(x_d)$$

Inversely, the copula density can be found from

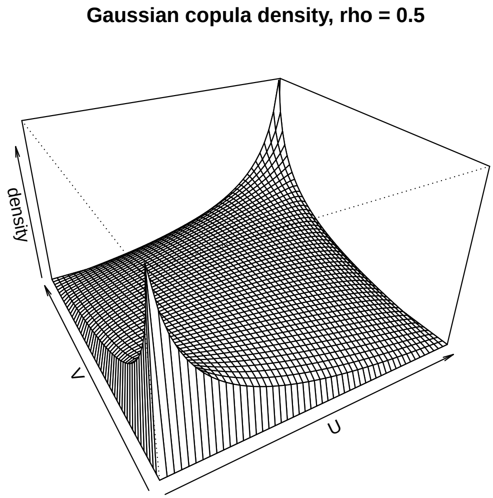
$$c(\boldsymbol{u}) = \frac{f(\boldsymbol{x})}{f_1(x_1) \cdots f_d(x_d)}, \qquad x_j = F_j^{-1}(u_j)$$

Ex. Prove these formulas.

Ex. Find the density of the *Gaussian* copula, i.e. the copula of the multivariate Gaussian distribution with invertible correlation matrix R. Hint: the density of such a Gaussian distribution is

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{d/2} \det(R)^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}' R^{-1} \boldsymbol{z}\right), \qquad \boldsymbol{z} \in \mathbb{R}^d$$

Density of the Gaussian copula



Conditional copula densities given a single variable are equal to the joint density

The density of a uniform variable being 1 on $[0,1]$, the conditional density of $\boldsymbol{U}_{-j}$ given $U_j = u_j$ is just c itself:

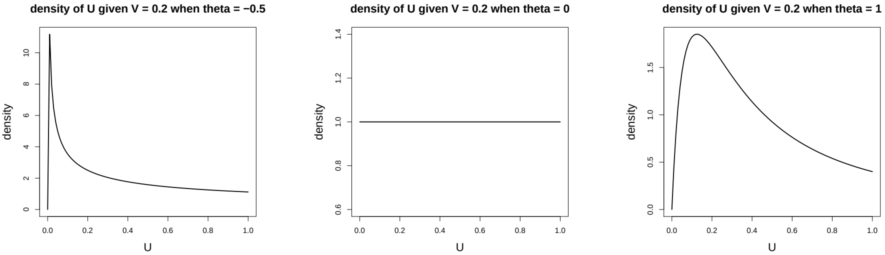
$$c_{\boldsymbol{U}_{-j}|U_j}(\boldsymbol{u}_{-j} \mid u_j) = c(\boldsymbol{u})$$

Ex. For the copula C_n in (MaxMin), check that the function $u \mapsto c_n(u, v)$, for fixed v, indeed defines a univariate density with 'parameter' v. Plot these densities and study the impact of n and v. What happens as $n \to \infty$?

Ex. For fixed j and u_j, is the function $\boldsymbol{u}_{-j} \mapsto c(\boldsymbol{u})$ again a copula density? Why (not)?

Conditional densities of the Clayton copula

Conditional pdf of $U \mid V = 0.2$ for the Clayton copula



$\theta = -0.5$	$\theta = 0$	$\theta = 1$
$n = 2$	$n \to \infty$	

Conditional distribution functions

The cdf of the conditional distribution of $\boldsymbol{U}_{-j}$ given $U_j = u_j$ is

$$\partial C(\boldsymbol{u}) / \partial u_j$$

Ex. Is the function $\boldsymbol{u}_{-j} \mapsto \partial C(\boldsymbol{u}) / \partial u_j$ a copula? Why (not)?

Ex. Compute $\partial C(u, v) / \partial v$ for

- $C(u, v) = uv$
- $C(u, v) = M(u, v) = \min(u, v)$
- $C(u, v) = W(u, v) = \max(u + v - 1, 0)$

What are the corresponding distributions for $U \mid V = v$?

Ex. Compute $\partial C_n(u, v) / \partial v$ with C_n as in (MaxMin).

The Gaussian copula density generates a two-parameter family of densities on the unit interval

Density of the bivariate Gaussian copula with parameter $\rho \in (-1, 1)$:

$$c_\rho(u, v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2} \frac{\rho^2 x^2 - 2\rho x y + \rho^2 y^2}{1-\rho^2} \right),$$

$$x = \Phi^{-1}(u),\ y = \Phi^{-1}(v)$$

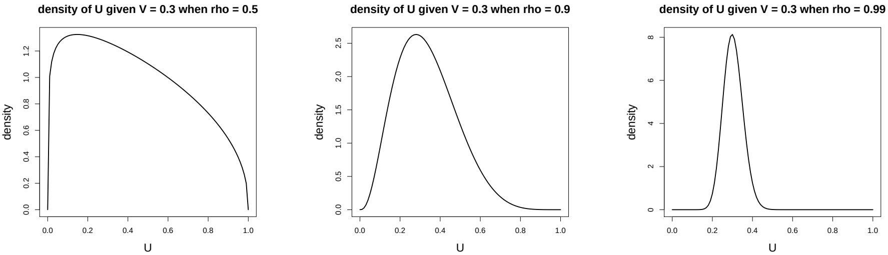
View this as a two-parameter family of densities on $(0, 1)$ via

$$u \mapsto c_\rho(u, v), \qquad \text{parameter } (\rho, v) \in (-1, 1) \times (0, 1)$$

This is the pdf of $U \mid V = v$ if $(U, V) \sim c_\rho$.

Conditional densities of the bivariate Gaussian copula

Conditional pdf of $U \mid V = 0.3$ if $(U, V) \sim C_\rho$



$\rho = 0.5$ $\rho = 0.9$ $\rho = 0.99$

Conditional copula densities and kernel smoothing on a compact interval

Ex. Show that if $(U, V) \sim C_\rho$ (Gaussian copula), then

$$(U \mid V = v) \stackrel{d}{=} \Phi\bigl(\rho\, \Phi^{-1}(v) + (1-\rho^2)^{1/2} Z\bigr), \qquad Z \sim N(0,1)$$

What happens if $\rho \to 1$?

Ex. Suppose one wants to estimate a density f on $(0,1)$ based on a sample $X_1, \ldots, X_n$. Heuristically motivate the following kernel density estimator:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n c_\rho(x, X_i), \qquad x \in (0,1)$$

the 'bandwidth' being $h = (1-\rho^2)^{1/2}$.

A variant of the probability integral transform: the Rosenblatt transform

Random pair $(X, Y) \sim F$. Conditional cdf

$$F(y|x) = \mathrm{P}[Y \le y \mid X = x]$$

Suppose that $y \mapsto F(y|x)$ is continuous for all x.

Rosenblatt transform

$$W = F(Y|X)$$

- $W \sim \mathrm{Uniform}(0, 1)$
- X and W are independent

Extends to higher dimensions: $X_1, F_{2|1}(X_2|X_1), F_{3|21}(X_3|X_1, X_2), \ldots$

Turning the inverse Rosenblatt transform into a simulation algorithm

If $(U, V) \sim C$, then

$$P[V \leq v \mid U = u] = \frac{\partial C(u, v)}{\partial u} =: \dot{C}_1(u, v)$$

Defining $W = \dot{C}_1(U, V)$, it follows that

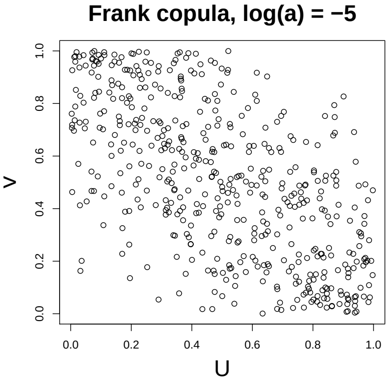
- U and W are independent Uniform$(0, 1)$ rv's
- $(U, q(W, U)) \sim C$ with q defined by $q(w, u) = v \iff \dot{C}_1(u, v) = w$

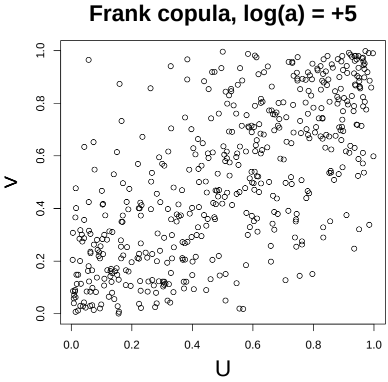
$\Rightarrow$ Generic way to generate random variates from a copula C.

Ex. Write and implement a simulation algorithm for the *Frank* copula

$$C(u, v) = \frac{1}{\log(a)} \log\left(1 + \frac{(a^u - 1)(a^v - 1)}{a - 1}\right), \qquad a \in (0, \infty) \setminus \{1\}$$

Random samples from a Frank copula





$\log a = -5$

$\log a = +5$

In a triple, apply the Rosenblatt transform to pairs

Uniform triple $(U_1, U_2, U_3) \sim C$.

Rosenblatt transforms for (U_1, U_2) and (U_3, U_2) conditionally on U_2:

$$U_{1|2} = \left. \frac{\partial C_{12}(u_1, u_2)}{\partial u_2} \right|_{(u_1,u_2)=(U_1,U_2)} =: C_{1|2}(U_1|U_2)$$

$$U_{3|2} = \left. \frac{\partial C_{32}(u_3, u_2)}{\partial u_2} \right|_{(u_3,u_2)=(U_3,U_2)} =: C_{3|2}(U_3|U_2)$$

Then

- $U_{1|2}$ and $U_{3|2}$ are again Uniform$(0, 1)$;
- $U_{1|2}$ and $U_{3|2}$ are both independent of U_2.

Still,

- the pair $(U_{1|2}, U_{3|2})$ is in general *not* independent of U_2.

Dependence or independence?
A brain teaser

Ex. For the *Farlie–Gumbel–Morgenstern* copula

$$C(u_1, u_2, u_3) = u_1 u_2 u_3 \left(1 + \theta\,(1 - u_1)(1 - u_2)(1 - u_3)\right), \qquad \theta \in [-1, 1],$$

check that

- the variables U_1, U_2, U_3 are *pairwise* independent
- and thus $U_{1|2} = U_1$ and $U_{3|2} = U_3$

although

- $(U_{1|2}, U_{3|2}) = (U_1, U_3)$ is *not* independent of U_2

Let's simplify: After conditioning, independence

Simplifying assumption

The copula of the conditional distribution of $(U_1, U_3) \mid U_2 = u_2$ *does not depend on the value of* u_2.

Equivalently:

$(U_{1|2}, U_{3|2})$ *and* U_2 *are independent.*

In this case, the conditional copula of $(U_1, U_3) \mid U_2 = u_2$, whatever u_2, is equal to the *unconditional* copula (cdf) of $U_{1|2}, U_{3|2}$:

$$C_{13|2}(u_1, u_3) = \mathrm{P}[U_{1|2} \leq u_1, U_{3|2} \leq u_3]$$

Ex. Does the simplifying assumption hold for the trivariate FGM copula?

The simplifying assumption allows a reduction to pair copulas

Under the simplifying assumption, the trivariate copula C is determined by the three pair copulas $C_{12}, C_{23}, C_{13|2}$:

$$\begin{aligned} C_{12} &\dashrightarrow \text{conditional distribution of } U_1 \text{ given } U_2, \\ C_{32} &\dashrightarrow \text{conditional distribution of } U_3 \text{ given } U_2, \\ C_{13|2} &\dashrightarrow \text{copula of the conditional distribution of } (U_1, U_3) \text{ given } U_2 \end{aligned}$$

In terms of densities:

$$c(u_1, u_2, u_3) = c_{13|2}\big(C_{1|2}(u_1|u_2),\ C_{3|2}(u_3|u_2)\big)\ c_{12}(u_1, u_2)\ c_{32}(u_3, u_2)$$

Higher-dimensional extensions lead to vine copulas or pair copula constructions.

For the Gaussian copula, the simplifying assumption holds

The copula of the multivariate normal distribution:

$$C_R(\boldsymbol{u}) = \Phi_R\big(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\big)$$

- R is a $d \times d$ correlation matrix
- Φ_R is the cdf of $N_d(\mathbf{0}, R)$
- Φ^{-1} is the $N(0,1)$ quantile function

Ex. What if we also allow for non-zero means or non-unit variances?

Ex. For the Gaussian copula, the *simplifying assumption* holds. Which are the pair copulas? Hint: if $(Z_1, Z_2, Z_3) \sim N_3(\mathbf{0}, R)$, then $(Z_1, Z_3) | Z_2 = z_2$ is bivariate Gaussian with correlation equal to the *partial correlation*

$$\rho_{13|2} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{(1-\rho_{12}^2)^{1/2}\,(1-\rho_{23}^2)^{1/2}}$$

Densities and conditional distributions: Some literature

Hofert, M., M. Mächler, and A. J. McNeil (2012). Likelihood inference for archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis 110*, 133–150.

Hofert, M. and D. Pham (2013). Densities of nested archimedean copulas. *Journal of Multivariate Analysis 118*, 37–52.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall. Section 4.5.

Jones, M. and D. Henderson (2007). Miscellanea kernel-type density estimation on the unit interval. *Biometrika 94*(4), 977–984.

Copulas: An Introduction
I - Fundamentals

Sklar's theorem

Densities and conditional distributions

Copulas for discrete variables

Measures of association

Multivariate discrete distributions:

Which multivariate discrete distributions do you know?

- Multinomial
- Negative multinomial
- Multivariate Poisson
- ...

Limited number of parametric families, with specific margins and dependence structures

Sklar's theorem revisited

Margins $F_1, \ldots, F_d$ and copula C, then

$$F(\boldsymbol{x}) = C\big(F_1(x_1), \ldots, F_d(x_d)\big)$$

is a d-variate cdf with margins $F_1, \ldots, F_d$,
even if (some of) $F_1, \ldots, F_d$ are discrete.

Proof.

If $F_j^{\leftarrow}(u) = \inf\{x \in \mathbb{R} : F_j(x) \geq u\}$ denotes the left-continuous inverse of F_j, then the rhs above is the cdf of

$$\big(F_1^{\leftarrow}(U_1), \ldots, F_d^{\leftarrow}(U_d)\big)$$

with $(U_1, \ldots, U_d) \sim C$. $\square$

Probability mass function

The pmf follows from the inclusion-exclusion formula:
For a pair of count variables $(X_1, X_2) \sim F$ and for $(x_1, x_2) \in \mathbb{N}$,

$$\begin{aligned} p(x_1, x_2) &= \mathrm{P}[X_1 = x_2, X_2 = x_2] \\ &= C\big(F_1(x_1), F_2(x_2)\big) - C\big(F_1(x_1 - 1), F_2(x_2)\big) \\ &\qquad - C\big(F_1(x_1), F_2(x_2 - 1)\big) + C\big(F_1(x_1 - 1), F_2(x_2 - 1)\big) \end{aligned}$$

From the pmf, one retrieves the conditional distributions.

Ex. Let (X_1, X_2) be a pair of Bernoulli variables with success probabilities p_1 and p_2, linked via a copula C.

1. Calculate the pmf of (X_1, X_2).
2. Show that C_1 and C_2 induce the same distribution on (X_1, X_2) as soon as

$$C_1(1 - p_1, 1 - p_2) = C_2(1 - p_1, 1 - p_2)$$

Non-uniqueness and (lack of) identifiability: The issue

The copula C is determined only on $F_1(\mathbb{R}) \times \cdots \times F_d(\mathbb{R})$. Hence, the copula C of F is not unique if $F_j(\mathbb{R}) \neq (0, 1)$, i.e. if F_j is not continuous. The copula is *non-identifiable*.

If $C_1(\boldsymbol{u}) = C_2(\boldsymbol{u})$ for all $\boldsymbol{u} \in F_1(\mathbb{R}) \times \cdots \times F_d(\mathbb{R})$, then

$$C_1\big(F_1(x_1), \ldots, F_d(x_d)\big) = C_2\big(F_1(x_1), \ldots, F_d(x_d)\big)$$

and both C_1 and C_2 are copulas of F, even if $C_1 \neq C_2$.

Non-uniqueness and (lack of) identifiability: A solution

For *parametric* models $\{C_\theta : \theta \in \Theta\}$, the parameter θ usually is identifiable by the values of C_θ on $F_1(\mathbb{R}) \times \cdots \times F_d(\mathbb{R})$.

Ex. For a pair of Bernoulli variables (X_1, X_2) with

$$P(X_j = 1) = p_j = 1 - P(X_j = 0), \qquad j \in \{1, 2\},$$

linked by the *Farlie–Gumbel–Morgenstern* copula

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

show that the parameter θ is identifiable.
[Hint: Calculate $P[X_1 = 0, X_2 = 0]$.]

Model construction

Sklar's theorem yields endless possibilities to construct multivariate distributions with discrete margins.

Ex.

- Invent a new parametric family of distributions for bivariate count data by combining margins and a copula of your choice. (Modestly name it after yourself.)
- Write software to compute its pmf and implement the maximum likelihood estimator for the parameter vector.
- Apply to it a fashionable data set.
- Publish the results in a prestigious journal.

Finding a copula for a multivariate discrete distribution: The issue

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a random vector with values in $\mathbb{N}^d$. The function

$$\boldsymbol{u} \mapsto F\big(F_1^{\leftarrow}(u_1), \ldots, F_d^{\leftarrow}(u_d)\big)$$

is *not* a copula (its margins are not uniform, since $F_j(F_j^{\leftarrow}(u_j)) \neq u_j$).

How to find a copula C for F?

Finding a copula for a multivariate discrete distribution: A solution

Let $V_1, \ldots, V_d$ be independent uniform $(0, 1)$ random variables, independent of $\boldsymbol{X}$. Consider

$$Y_j = X_j + V_j - 1, \qquad \boldsymbol{Y} = (Y_1, \ldots, Y_d)$$

Then Y_j is continuous and

$$\{Y_j \le x_j\} = \{X_j \le x_j\}, \qquad x_j \in \mathbb{N}.$$

The (unique) copula C of $\boldsymbol{Y}$ is also a copula of $\boldsymbol{X}$.

Ex. Given the cdf of X_j, draw the one of Y_j.

Ex. Apply this construction to find a copula for the Bernoulli pair X_1, X_2 above $\mathrm{P}[X_1 = 1, X_2 = 1] = p_{12}$. Explain the name ‘checker-board copula’.

Copulas for discrete variables: Some literature

Denuit, M. and P. Lambert (2005). Constraints on concordance measures in bivariate discrete data. *J. Multivariate Anal. 93*(1), 40–57.

Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *Astin Bull. 37*(2), 475–515.

Genest, C., J. Nešlehová, and B. Rémillard (2013). On the empirical multilinear copula process for count data. *Bernoulli*, in press.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall. Chapters 7 and 11.

Copulas: An Introduction
I - Fundamentals

Reducing a copula to a number

- Copulas are a fairly complex way to describe dependence.
- Simplify to numerical summary measures of the dependence structure.
- Different summary measures focus on different aspects.
- Distinct copulas may share the same value of a summary measure.
  - Zero correlation does not imply independence.
    E.g. $X \sim N(0, 1)$ and $Y = X^2$
- For parametric copula families, the value of a numerical summary measure may sometimes identify the parameter.

To avoid problems with ties, restrict to *continuous* distributions.

Association versus dependence

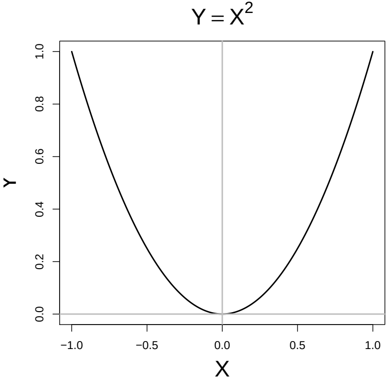
Association: The extent up to which large (small) values of X go together with large (small) values of Y.

Dependence: The extent up to which the outcome of Y is predictable from the outcome of X.

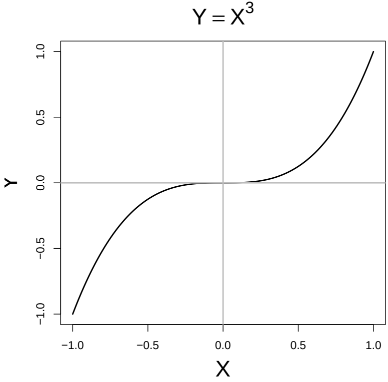
- Example: If $X \sim N(0, 1)$ and $Y = X^2$, then X and Y are perfectly dependent but not associated.

In this section, we will consider measures of association.

Association, dependence, and linear correlation



perfectly dependent
but not at all associated



perfectly associated
but not perfectly correlated

Criticisms on Pearson's linear correlation

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\ \operatorname{var}(Y)}} \in [-1, 1]$$

- Does not even exist if $\mathrm{E}[X^2] = \infty$ or $\mathrm{E}[Y^2] = \infty$
- Even for increasing f and g, in general $\operatorname{cor}(f(X), g(Y)) \neq \operatorname{cor}(X, Y)$
- Even if X and Y are perfectly associated, $\operatorname{cor}(X, Y)$ need not be 1

Ex. Calculate $\operatorname{cor}(X, X^3)$ for $X \sim N(0, 1)$.
[Hint: $\mathrm{E}[X^{2p}] = (2p - 1) \times (2p - 3) \times \cdots \times 1$ for integer $p \geq 1$.]

Kendall's tau: concordance versus discordance

Measure assocation by probabilities of con/dis-cordance:
if (X_1, Y_1) and (X_2, Y_2) are iid F, then

$$\begin{aligned}\tau(F) = \quad & \mathrm{P}[X_1 - X_2 \text{ and } Y_1 - Y_2 \text{ have the same sign}] \\ & - \mathrm{P}[X_1 - X_2 \text{ and } Y_1 - Y_2 \text{ have opposite signs}]\end{aligned}$$

Ex. Draw pairs of points (x_1, y_1) and (x_2, y_2) in the plane which are

- concordant
- discordant

Ex. Show that $\tau(W) = -1 \leq \tau(F) = \tau(C) \leq 1 = \tau(M)$ with M and W the Fréchet–Hoeffding upper and lower bounds.

Kendall's tau as a copula property

Since $\tau(F)$ is invariant if we apply increasing transformations f and g to X and Y, respectively, one can show that

$$\tau(F) = \tau(C) = 4 \int_{[0,1]^2} C(u,v) \, \mathrm{d}C(u,v) - 1$$

Ex. Show that $\tau(C_\theta) = 2\theta/9$ for C_θ the *FGM* copula

$$C_\theta(u,v) = uv \left(1 + \theta \, (1-u) \, (1-v)\right), \qquad -1 \le \theta \le 1.$$

How does this impair the applicability of the FGM copula?

Spearman's rho: Pearson's linear correlation revisited

Random pair (X, Y) with margins F and G.
Put $U = F(X)$ and $V = G(Y)$, so $(U, V) \sim C$.

$$\rho_S(C) = \operatorname{cor}(U, V) = 12 \int_{[0,1]^2} C(u, v)\, \mathrm{d}u\, \mathrm{d}v - 3$$

Ex. Prove the second equality.

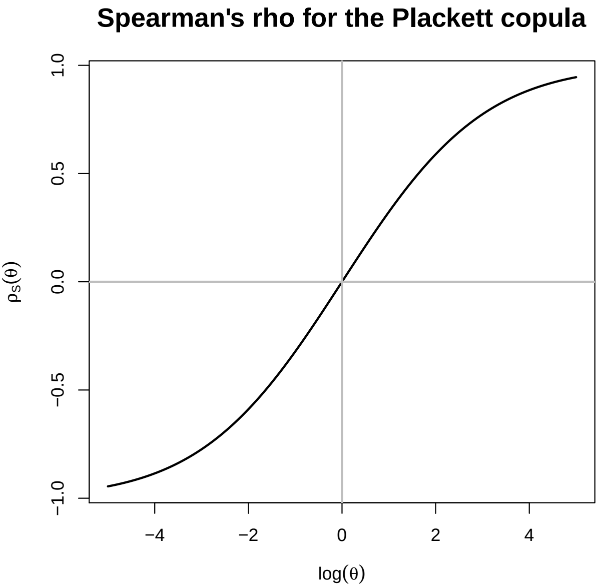
Ex. Show that $\rho(W) = -1 \leq \rho(F) = \rho(C) \leq 1 = \rho(M)$ with M and W the Fréchet–Hoeffding upper and lower bounds.

Ex. For the *Plackett* copula C_θ with odds ratio $\theta > 0$, show that

$$\rho_S(C_\theta) = \frac{\theta + 1}{\theta - 1} - \frac{2\theta}{(\theta - 1)^2} \log \theta$$

What happens if $\theta \to 0$, $\theta = 1$, or $\theta \to \infty$? First guess, then compute.

Spearman's rho of the Plackett copula



Coefficients of tail dependence: Joint exceedances below or above thresholds

If focus is on joint exceedances below (small) thresholds, consider

$$\operatorname{cor}(\mathbf{1}\{U \le w\}, \mathbf{1}\{V \le w\}) = \frac{C(w,w) - w^2}{w(1-w)}, \qquad 0 < w < 1$$

Coefficient of lower tail dependence:

$$\begin{aligned}\lambda_L(C) &= \lim_{w \downarrow 0} \operatorname{cor}(\mathbf{1}\{U \le w\}, \mathbf{1}\{V \le w\}) \\ &= \lim_{w \downarrow 0} \frac{C(w,w)}{w} \in [0,1]\end{aligned}$$

Coefficient of upper tail dependence:

$$\lambda_U(C) = \lambda_L(\bar{C}) = \lim_{w \downarrow 0} \frac{2w - 1 + C(1-w, 1-w)}{w}$$

Coefficients of tail dependence: An exceedance given an exceedance

Lower tails:

$$\frac{C(w,w)}{w} = \mathrm{P}(U \le w \mid V \le w) = \mathrm{P}(V \le w \mid U \le w)$$

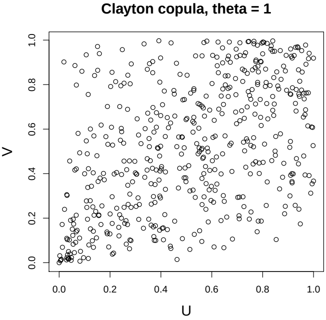
Upper tails:

$$\frac{2w - 1 + C(1-w, 1-w)}{w} = \mathrm{P}(U \ge 1-w \mid V \ge 1-w) = \mathrm{P}(V \ge 1-w \mid U \ge 1-w)$$

- Coefficients of tail dependence $\lambda_L(C)$ and $\lambda_U(C)$: limits as $w \downarrow 0$
- Asymptotic tail independence: if $\lambda_{L/U}(C) = 0$.

The Clayton copula: lower tail dependence

$$C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \qquad \theta > 0$$



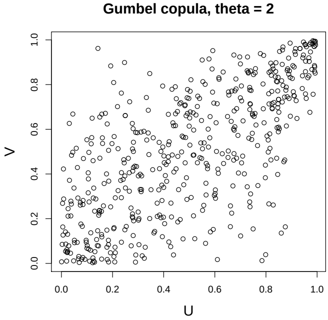
Ex. Show that

$$\lambda_L(C_\theta) = 2^{-1/\theta}$$
$$\lambda_U(C_\theta) = 0$$

What happens if $\theta \to 0$ or $\theta \to \infty$?

The Gumbel copula: upper tail dependence

$$C_\theta(u, v) = \exp[-\{(-\log u)^\theta + (-\log v)^\theta\}^{1/\theta}], \qquad \theta \geq 1$$



Ex. Show that

$$\lambda_L(C_\theta) = 0$$
$$\lambda_U(C_\theta) = 2 - 2^{1/\theta}$$

What happens if $\theta = 1$ or $\theta \to \infty$?

Many other measures of association

- Spearman's footrule
- Gini's gamma
- Blomqvist beta
- van der Waerden rank correlation
- Extensions to more than two variables:
  - *within* random vectors
  - *between* random vectors
- More refined tail dependence coefficients in case of asymptotic independence
- …

Remarks on association measures

- One-parameter copula families: often a one-to-one relation between the parameter and the value of an association measure
  $\Rightarrow$ reparametrization in terms of this association measure
- Different association measures intend to measure the same thing
  $\Rightarrow$ various relations (inequalities etc.) among such measures
- Which association measure to use? No clear rules. Depends on
  - Mathematical convenience
  - Personal preferences
  - …

Measures of association: Some literature

Christian Genest, C., Nešlehová, and N. Ben Ghorbal (2010). Spearman's footrule and Gini's gamma: a review with complements. *Journal of Nonparametric Statistics 22*(8), 937–954.

Coles, S., J. Heffernan, and J. Tawn (1999). Dependence measures for extreme value analyses. *Extremes 2*(4), 339–365.

Grothe, O., J. Schnieders, and J. Segers (2013). Measuring association and dependence between random vectors. *Journal of Multivariate Analysis (to appear)*, arXiv:1107.4381.

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer. Section 5.1.

Schmid, F., R. Schmidt, T. Blumentritt, S. Gaisser, and M. Ruppert (2010). Copula-based measures of multivariate association. In P. Jaworski, F. Durante, W. K. Härdle, and T. Rychlik (Eds.), *Copula Theory and Its Applications*, Lecture Notes in Statistics, pp. 209–236. Berlin: Springer.