Chapter 4

Copula

Geodesy and other technical disciplines have used in its history various mathematical models to describe observed as well as mediate variables of inspected phenomenons. Univariate behaviour first, then multivariate capturing mutual dependencies, the focus was always put to understanding and predicting the values of individual concern. A copula function is (recently very popular) tool for relating different dimensions of a data output.

Before we zoom to relevant theory, it may come handy to look "a little" back in section 4.1, following [19]. After introducing the idea of copula theory, section 4.2 gives an interesting look into dependence measuring, which is helpful in the discussion about association between random variables and the role that copulas play in it. Section 4.3 is geared to Archimedean class of copulas, pointing out the easiness with which they can be constructed, while the section 4.1 describes the estimation procedure in details.

It is necessary to remark in advance, that the notation in this chapter differs significantly from that in the other chapters. This is understandable since the concept of copula, which is fairly new to the time series analysis, looks at the modelled data from a different angle.

4.1 Introduction to copula

Understanding relationships among multivariate outcomes is a basic problem in statistical science. In the late nineteenth century, Sir Francis Galton made a fundamental contribution to understanding multivariate relationships with his introduction of regression analysis, by which he linked the distribution of heights of adult children to the distribution of their parents' heights. Galton showed not only that each distribution was approximately normal but also that the joint distribution could be described as a bivariate normal. Thus, the conditional distribution of adult children 's height, given the parents' height, could also be described by using normal distribution. Regression analysis has developed into the most widely applied statistical methodology and become an important component of multivariate analysis, because it allows researchers to focus on the effects of explanatory variables.

However, though widely applicable, regression analysis is limited by the basic setup that requires the analyst to identify one dimension of the outcome as the primary measure of interest (the dependent variable) and other dimensions as supporting or "explaining" this variable (the independent variables). This may generally be not of primary interest, thus our attention should be focused on the more basic problem of understanding the distribution of several outcomes, a multivariate distribution.

As normal distribution has the most practical use when describing one-dimensional data sets, it has long dominated the study of multivariate distributions as well. Multivariate normal distributions are appealing because the marginal distributions are normal too, and also because the association between any two random outcomes can be fully described knowing only the marginal distributions and additional parameter (correlation coefficient). However, there are many datasets, to that normal distribution does not provide an adequate approximation. For that reason, many non-normal distributions have been developed, mostly as immediate extensions of univariate distributions (Pareto, gamma, ...). Drawbacks of such a construction are that (a) a different family is needed for each marginal distribution, (b) extensions to more than just the bivariate case are not clear, (c) and measures of association often appear in the marginal distributions. A construction of multivariate distributions that does not suffer from these drawbacks is based on the *copula* function.

Copula is a function that links univariate marginals to their full multivariate distribution. To cast light on previous definition, consider p uniform (on the unit interval) random variables $U_1, U_2, \dots, U_p$ whose joint distribution function C is defined as

$$C(u_1, u_2, \dots, u_p) = \Pr[U_1 \leq u_1, U_2 \leq u_2, \dots, U_p \leq u_p], \tag{4.1}$$

where u denotes realizations. Those p variables are distribution functions (also referred to as probability integral transforms) of p outcomes $X_1, X_2, \dots,$ X_p (each of them being a continuous random variable) that we wish to understand. They are the marginal distribution functions $F_1, \dots, F_p$ of multivariate distribution function

$$C\Bigg(F_1(x_1), F_2(x_2), \dots, F_p(x_p)\Bigg) = F(x_1, x_2, \dots, x_p), \tag{4.2}$$

defined using a copula function, evaluated at realizations $x_1, x_2, \ldots, x_p$.

In 1959 Sklar formulated his famous theorem, where the converse of (4.2) was established, and that practically meant the foundation of whole copula theory. He proved that any joint distribution function F with univariate marginal distribution functions $F_1, \ldots, F_p$ can be seen as a copula function, i.e.

$$F(x_1, x_2, \ldots, x_p) = C\bigg(F_1(x_1), F_2(x_2), \ldots, F_p(x_p)\bigg). \tag{4.3}$$

He also showed that if the marginal distributions are continuous, then there is a unique copula representation (in general, C is unique on the $\mathrm{Ran}F_1 \times \mathrm{Ran}F_2 \times \cdots \times \mathrm{Ran}F_p$, where $\mathrm{Ran}F$ stands for the range of F).

Thus copula functions provide a unifying and flexible way to study joint distributions (with different marginals). Moreover, copula allows us to model the dependence structure independently from the marginal distributions.

As for the basic properties, following [9] and restricting ourselves to bivariate representation, copula is a function $C : [0,1]^2 \longrightarrow [0,1]$ which

- satisfies the boundary conditions $C(t,0) = C(0,t) = 0$ and $C(t,1) = C(1,t) = t$ for $t \in [0,1]$,
- satisfies the 2-increasing property:
  $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$ for all u_1, u_2, v_1, v_2 in $[0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

A copula is symmetric if $C(u,v) = C(v,u)$ for all (u,v) in $[0,1]^2$ and is asymmetric otherwise.

Now consider the functions M, Π and W defined on $[0,1]^2$ as follows:

$$\begin{aligned} M(u,v) &= \min(u,v), \\ \Pi(u,v) &= uv, \\ W(u,v) &= \max(u+v-1, 0). \end{aligned} \tag{4.4}$$

These functions are copulas, actually 2-copulas (i.e. copulas with two-dimensional domain), and M, W satisfy so-called Fréchet-Hoeffding bounds inequality

$$W(u,v) \leq C(u,v) \leq M(u,v), \tag{4.5}$$

where C is any 2-copula. W and M are called Fréchet-Hoeffding lower and upper bound, respectively. They represent perfect dependence, either negative or positive, whereas the product copula Π stands for perfect independence. If we extend the domain to $[0,1]^p$ for $p \geq 3$, (observe that M, Π and W are associative and thus their p-ary extension is trivial), still the bounds are M and W. However, the lower bound W is no more a p-copula (but still it is the best lower bound).

So far, numerous copulas have been developed and can be found listed in literature (for instance see [42]). Because of the above mentioned appealing properties of normal distribution, the most commonly applied function is the normal copula

$$C_{normal}(u_1, \ldots, u_n) = \Psi\bigg(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)\bigg), \tag{4.6}$$

where Ψ denotes the joint distribution function of the n-variate standard normal distribution and Φ^{-1} the inverse of univariate normal standard distribution function (see [13]). Multi-normal distribution belongs to the elliptical distributions, which captures only linear dependencies (the parameter set being correlation matrix) and therefore is inadequate in many multivariate analyses of data with probability density concentrated on tails (extreme values), for instance.

In this paper, our main concern is an interesting class of copulas, denoted Archimedean, that possess some outstanding useful properties. Archimedean copulas are going to be introduced after we say few words about measures of dependence.

4.2 Dependence and measures of association

In this section we recall some basic concepts of dependence or association between random variables and the role that copulas can play in this most widely studied subject in probability and statistics. Following [42], [19], there is a variety of ways to discuss and to measure dependence. Many of them are "scale-invariant", that is, they remain unchanged under strictly increasing transformations of the random variables. To understand the spirit of copula, consider two random variables X, Y and two functions f, g, strictly increasing (but otherwise arbitrary) over the range of X, Y. Then the transformed variables $f(X)$ and $g(Y)$ have the same copula as X and Y - in other words, the manner in which X and Y "move together" is captured by the copula, regardless of the scale in which each variable is measured.

The most famous and widely used measure of association is Pearson's product-moment correlation coefficient

$$corr(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}\ , \tag{4.7}$$

however, it measures only a *linear* dependence between random variables. In context of joint distributions, $corr(X, Y)$ depends not only on the copula but also on the marginal distributions, thus this measure is affected by (nonlinear) changes of scale. Since Pearson's coefficient has adopted the customary name, correlation coefficient, for scale-invariant measures we shall use more modern term "measure of association". The most widely known ones are the population versions of Kendall's tau (τ) and Spearman's rho (ρ), both of which measure a form of dependence known as *concordance*.

Informally, a pair of random variables is concordant if "large" values of one tend to be associated with "large" values of the other, and "small" values of one with "small" values of the other. More precisely, if (x_i, y_i) and (x_j, y_j) denote two observations of a vector (X, Y) of continuous random variables, we say that (x_i, y_i) and (x_j, y_j) are concordant if $(x_i - x_j)(y_i - y_j) > 0$, and discordant if $(x_i - x_j)(y_i - y_j) < 0$.

From the sample version of Kendall's tau defined as $t = (c - d)/(c + d)$, where c is the number of concordant and d the number of discordant pairs (x_i, y_i) and (x_j, y_j), we may work out easily that the population version of Kendall's tau will be defined as the probability of concordance minus the probability of discordance

$$\tau = \tau_{X,Y} = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0] \ , \quad (4.8)$$

where (X_1, Y_1) and (X_2, Y_2) are assumed to be independent and identically distributed random vectors. Before we link τ with copulas, define a "concordance function" Q in the same way as τ in (4.8), with that difference that the continuous random variables in the two vectors (X_1, Y_1) and (X_2, Y_2) have (possibly) different joint distributions H_1 and H_2, but common margins F and G. Then the equality

$$Q = Q(C_1, C_2) = 4 \iint_{[0,1]^2} C_2(u, v) dC_1(u, v) - 1 \quad (4.9)$$

shows, that this function depends on the distributions of the two vectors only through their copulas C_1 and C_2. According to (4.9) the population version of Kendall's tau in terms of copulas is given by

$$\tau_{X,Y} = \tau_C = Q(C, C) = 4 \iint_{[0,1]^2} C(u, v) dC(u, v) - 1 \ , \quad (4.10)$$

where C is the copula of X and Y. Integral, which appears in (4.10) can be interpreted as the expected value of the function $C(U, V)$ of random variables U and V uniform on $(0, 1)$ whose distribution function is C; then $\tau_C = 4E[C(U, V)] - 1$. Next section shows the taking advantage of linking τ to Archimedean copulas in their estimation.

Similarly, the population version of the measure of association known as Spearman's rho is based on concordance and discordance. Let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) be three independent random vectors with common joint distribution function H (whose margin are again F and G) and copula C. The population version of Spearman's rho is defined to be proportional to the probability of concordance minus the probability of discordance for the two vectors (X_1, Y_1) and (X_2, Y_3) – i.e., a pair of vectors with the same margins but one vector has distribution function H, while the components of the other are independent:

$$\rho = \rho_{X,Y} = 3\bigg(\Pr[(X_1 - X_2)(Y_1 - Y_3) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_3) < 0] \bigg) , \tag{4.11}$$

(the pair (X_3, Y_2) could be used equally as well). Note that while the joint distribution function of (X_1, Y_1) is $H(x, y)$, the joint distribution function of (X_2, Y_3) is $F(x)G(x)$ (since X_2 and Y_3 are independent) and their copula is Π. Then the population version of Spearman's rho is given by

$$\begin{aligned} \rho_{X,Y} = \rho_C = 3Q(C, \Pi) &= 12 \iint_{[0,1]^2} uv \, dC(u, v) - 3 \\ &= 12 \iint_{[0,1]^2} C(u, v) du dv - 3 \, . \end{aligned} \tag{4.12}$$

The coefficient "3" that appears in (4.11) and (4.12) is a "normalization" constant, since $Q(C, \Pi) \in [-1/3, 1/3]$, allowing ρ to satisfy the range property of measures of concordance.

Here we list some of the properties that a measure κ of association between two random variables X and Y should satisfy to be a measure of concordance:

- $-1 \leq \kappa_{X,Y} \leq 1, \ \kappa_{X,X} = 1, \ \kappa_{X,-X} = -1,$
- $\kappa_{X,Y} = \kappa_{Y,X},$
- if X and Y are independent, then $\kappa_{X,Y} = \kappa_{\Pi} = 0,$
- $\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}.$

Spearman's rho is also called a "grade"[1] correlation coefficient. For closer look, if x and y are observation from two random variables X and Y with distribution functions F and G, respectively, then the grades of x and y are given by $u = F(x)$ and $v = G(y)$. Note that the grades (u and v)

[1]Grade is the population analogue of rank

are observations from the uniform (0,1) random variables $U = F(X)$ and $V = G(Y)$ whose distribution function is copula C. Thus Spearman's rho for a pair of continuous random variables X and Y is identical to Pearson's product-moment correlation coefficient for the grades U and V:

$$\rho_{X,Y} = corr(F(X), G(Y)).$$

Another interpretation of Spearman's rho says that it is proportional to the average difference between the graph of the copula C and the product copula Π over the unit square $[0,1]^2$.

4.3 Archimedean copula

In this chapter we focus on an important class of copulas known as Archimedean copulas. They find a wide range of applications mainly because of (a) the ease with which they can be constructed, (b) the great variety of families of copulas which belong to this class, and (c) the many nice properties possessed by the members of this class. Archimedean copulas originally appeared not in statistics, but rather in the study of probabilistic metric spaces, where they were studied as a part of the development of a probabilistic version of the triangle inequality. Like a copula, a triangle norm, or t-norm maps $[0,1]^p$ to $[0,1]$ and joins distribution functions (here the resulting distribution function is univariate). Some t-norms (exactly those which are 1-Lipschitz) are copulas and vice versa, some copulas (exactly those which are associative) are t-norms. Moreover, Archimedean t-norms which are also copulas are called Archimedean copulas. Recall that (for $p = 2$) a mapping T: $[0,1]^2 \to [0,1]$ is called an Archimedean t-norm if it is continuous, associative, non-decreasing, 1 is its neutral element and $T(x,x) < x$ for all $x \in]0,1[$. Each Archimedean t-norm can be represented in the form $T(x,y) = t^{-1}\big(\min(t(x)+t(y), t(0))\big)$ where t: $[0,1] \to [0,\infty]$ is continuous, strictly decreasing and $t(1) = 0$, see [34].

The Archimedean representation allows us to reduce the study of a multivariate copula to a single univariate function. For simplicity, we consider bivariate copulas so that $p = 2$. Assume that ϕ is a convex, decreasing function with domain $(0,1]$ and range in $[0,\infty)$, that is ϕ: $(0,1] \to [0,\infty)$, such that $\phi(1) = 0$. Use ϕ^{-1} for the function which is inverse of ϕ on the range of ϕ and 0 otherwise. Then the function

$$C_\phi(u,v) = \phi^{-1}\Big(\phi(u) + \phi(v)\Big) \quad \text{for } u, v \in (0,1] \tag{4.13}$$

is said to be an Archimedean copula. ϕ is called a *generator* of the copula C_ϕ. Archimedean copula is symmetric, also associative, i.e. $C(C(u,v),w) =$

Table 4.1: Archimedean copulas with their generators.

Family of copulas	Generator $\phi(t)$	Param. θ	Bivariate copula $C_\phi(u,v)$	Special cases
product	$-\ln t$		uv	$C{=}\Pi$
Gumbel	$(-\ln t)^\theta$	$\theta \geq 1$	$e^{-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta}}$	$C_1{=}\Pi,\ C_\infty{=}M$
Clayton	$t^{-\theta} - 1$	$\theta > 0$	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$C_0{=}\Pi,\ C_\infty{=}M$
Frank	$-\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$	$\theta \in \Re$	$-\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right)$	$C_0{=}\Pi$ $C_{-\infty}{=}W,\ C_\infty{=}M$

$C(u, C(v, w))$ for all $u, v, w \in [0, 1]$, and for any constant $k > 0$ the $k\phi$ is also a generator of C_ϕ. Observe that Archimedean copulas (which are always 2-copulas) as p-ary operators need not be p-copulas. A necessary and sufficient condition for an Archimedean copula to be p-copula for each $p \geq 2$ is the total monotonicity of the function ϕ^{-1} [42]. If the generator is twice differentiable, the copula is absolutely continuous and the copula density (probability density function of random vector (U, V)) is given by

$$c_\phi(u,v) = \frac{\partial^2 C_\phi(u,v)}{\partial u \partial v} = \frac{-\phi''(C_\phi(u,v))\phi'(u)\phi'(v)}{[\phi'(C_\phi(u,v))]^3}. \tag{4.14}$$

As a generator uniquely determines an Archimedean copula, different choices of generator yield many families of copulas, that consequently, besides the form of generator, differ in the number of dependence parameters and their range. Table 4.1 summarizes the most important one-parameter families of Archimedean class. For convenience the copula notation C_ϕ is replaced by C_θ in the last column, where θ assumes its limiting values. Note, that Clayton and Gumbel copulas model only positive dependence, while Frank covers the whole range.

Now that we're talking about dependence, recall the population version of Kendall's tau whose evaluation requires the evaluation of the double integral

Table 4.2: Measures of association related to Archimedean copulas

Family	product	Gumbel	Clayton	Frank
Kendall's τ	0	$\frac{\theta-1}{\theta}$	$\frac{\theta}{\theta+2}$	$1 - \frac{4}{\theta}\{1 - D_1(\theta)\}$
Spearman's ρ	0	no closed form	complicated form	$1 - \frac{12}{\theta}\{D_1(\theta) - D_2(\theta)\}$
Note: $D_k(x) = \frac{k}{x^k}\int_0^x \frac{t_k}{e^t-1}dt$ is so called				"Debye" function.

in (4.10). For an Archimedean copula, the situation is simpler, in that τ can be evaluated directly from the generator of the copula

$$\tau_C = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt \tag{4.15}$$

[21]. Indeed, one of the reasons that Archimedean copulas are easy to work with is that often expressions with one-place function (the generator) can be employed rather than expressions with a two-place function (the copula). Table 4.2 shows particular closed forms of (4.15).

4.4 Fitting a copula to bivariate data

The Archimedean copula has simplified the construction of bivariate distributions and it has many families that are capable to present different structure of dependence and there are many different methods developed to estimate its parameters. We only need to find functions which will serve as generators, define the corresponding copulas and estimate their dependence parameters.

For identifying the copula, we focus on the procedure of Genest and Rivest [22], that is also referred to as *nonparametric* estimation of copula parameter. Then we use *semi-parametric* estimation method developed in [23] and finally the experiment with bivariate geodetic data is given to illustrate the proposed theory. The procedures are also discussed in [19], [40] and [1].

In the following, we consider the three most widely used Archimedean families of copula: Clayton, Gumbel and Frank.

Nonparametric estimation

As [19] formulate, measures of association summarize information in the copula concerning the dependence, or association, between random variables. Thus, following [22] we can also use those measures to specify a copula form in empirical applications.

Assume that we have a random sample of bivariate observations (X_i, Y_i) for $i = 1, \ldots, n$ available. Assume that the joint distribution function H has associated Archimedean copula C_ϕ; we wish to identify the form of ϕ. First to begin with, define an intermediate (unobserved) random variable $Z_i = H(X_i, Y_i)$ that has distribution function $K(z) = \Pr[Z_i \leq z]$. This distribution function is related to the generator of an Archimedean copula through the expression

$$K(z) = K_\phi(z) = z - \frac{\phi(z)}{\phi'(z)}\,. \tag{4.16}$$

To identify ϕ, we:

1. Find Kendall's tau using the usual (nonparametric or distribution-free) estimate
$$\tau_n = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1}\operatorname{sgn}[(X_i - X_j)(Y_i - Y_j)]}{\sum_{i=2}^{n}\sum_{j=1}^{i-1}|\operatorname{sgn}[(X_i - X_j)(Y_i - Y_j)]|}\ .$$

2. Construct a nonparametric estimate of K, as follows:

   a) first, define the pseudo-observations $Z_i = \{$ number of (X_j, Y_j) such that $X_j < X_i$ and $Y_j < Y_i\}/(n-1)$ for $i = 1, \ldots n$:
   $$Z_i = (n-1)^{-1}\sum_{j=1}^{n} \mathrm{I}[X_j < X_i \wedge Y_j < Y_i],$$

   b) second, construct the estimate of K as proportion of Z_is smaller than z, that is
   $$K_n(z) = n^{-1}\sum_{i=1}^{n} \mathrm{I}[Z_i \leq z],$$

   where indicator function $\mathrm{I}[A]$ gives 1 if A occurs and 0 otherwise.

3. Now construct a parametric estimate K_ϕ using the relationship (4.16). Illustratively, $\tau_n \longrightarrow \theta_n \longrightarrow \phi_n(t) \longrightarrow K_{\phi_n}(z)$, where subscript n denotes estimate. For various choices of generator, refer to Table 4.1, and for linking τ to θ, Table 4.2 is helpful.

The step 3 is to be repeated for every copula family we wish to compare. The best choice of generator then corresponds to the parametric estimate $K_{\phi_n}(z)$, that most closely resembles the nonparametric estimate $K_n(z)$. Measuring "closeness" can be done either by a (L_2-norm) distance such as $\int_0^1[K_{\phi_n}(z) - K_n(z)]^2dz$ or graphically by (a) plotting of $z - K(z)$ versus z or (b) corresponding quantile-quantile (Q-Q) plots (see [22], [19], [12]). Q-Q plots are used to determine whether two data sets come from populations with a common distribution. If the points of the plot, which are formed from the quantiles of the data, are roughly on a line with a slope of 1, then the distributions are the same.

Semi-parametric estimation

To estimate dependence parameter θ, two strategies can be envisaged. The first and straightforward one writes down a likelihood function, where the valid parametric models of marginal distributions are involved. The resulting estimate $\hat{\theta}$ would then be margin-dependent, just as the estimates of the parameters involved in the marginal distributions would be indirectly affected by the copula. As the multivariate analysis focus on the dependence structure, it requires the dependence parameter to be margin-free. That's why [23] proposed a semi-parametric procedure for the second strategy, when we don't want to specify any parametric model to describe the marginal distribution. This procedure consist of (a) transforming the marginal observations into uniformly distributed vectors using the it empirical distribution function, and (b) estimating the copula parameters by maximizing a *pseudo log-likelihood* function.

So, given a random sample as before, we look for $\hat{\theta}$ that maximizes the pseudo log-likelihood

$$L(\theta) = \sum_{i=1}^{n} \log\bigg(c_\theta(F_n(x), G_n(y)) \bigg) , \tag{4.17}$$

in which F_n, G_n stands for re-scaled empirical marginal distributions functions, i.e.,

$$F_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathrm{I}[X_i \leq x] , \tag{4.18}$$

$G_n(y)$ arise analogically. This re-scaling avoids difficulties from potential unboundedness of $\log(c_\theta(u,v))$ as u or v tend to one. Genest et al. in [23] examined the statistical properties of the proposed estimator and proved it to be consistent, asymptotically normal and fully efficient at independence case.

The copula density c_θ for each Archimedean copula can be acquired from (4.14). To examine a goodness of our estimation, there is the Akaike information criterion available for comparison: $AIC = -2(\textit{log-likelihood}) + 2k$, where k is the number of parameters in the model (in our case, $k = 1$). The lowest AIC value determines the best estimator.