MODELLING RELATIONSHIP USING ARCHIMEDEAN COPULA: An introduction to experimental study

Tomáš Bacigál

1. Preface

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. This article introduces the concept of a copula function as a 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 2, following [4]. After introducing the idea of copula theory, section 3 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 is geared to Archimedean class of copulas, pointing out the easiness with which they can be constructed, while the fifth section describes the estimation procedure in details. There we give an application to position dynamics observed by means of satellite based positioning technique (GPS). Finally, section 6 concludes.

All the theory preceding the experiment is presented in good belief that it will help specialists in applied sciences to quickly and easily adopt the advantage of relatively new procedures of modelling their data - by using copulas.

2. 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 pa-

rameter (correlation coefficient). However, there are many datasets, to that normal distribution does not provide an adequate approximation. For that reason, many non-normal distributions has 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 $U_1, U_2, \ldots, U_p$ whose joined distribution function C is defined as

$$C(u_1, u_2, \ldots, u_p) = \text{Prob}[U_1 \leq u_1, U_2 \leq u_2, \ldots, U_p \leq u_p], \tag{1}$$

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

$$C\Big(F_1(x_1), F_2(x_2), \ldots, F_p(x_p)\Big) = F(x_1, x_2, \ldots, x_p), \tag{2}$$

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 (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.

$$F(x_1, x_2, \ldots, x_p) = C\Big(F_1(x_1), F_2(x_2), \ldots, F_p(x_p)\Big), \tag{3}$$

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

Thus copula functions provide a unifying and flexible way to study joined 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 $M,\ \Pi$ and W defined on $[0,1]^2$ as follows:

$$\begin{aligned} M(u,v) &= \min(u,v), \\ \Pi(u,v) &= uv, \\ W(u,v) &= \max(u+v-1,0). \end{aligned} \tag{4}$$

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

$$W(u,v) \leq C(u,v) \leq M(u,v), \tag{5}$$

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 $M,\ \Pi$ 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 [9]). Because of the above mentioned appealing properties of normal distribution, the most commonly applied function is the normal copula

$$C_{_{normal}}(u_1,\ldots,u_n) = \Phi\Big(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_n)\Big), \tag{6}$$

where Φ denotes the joint distribution function of the n-variate standard normal distribution and Φ^{-1} the inverse of univariate normal standard distribution function (see [3]). 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.

3. 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 [9], [4], 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 is Pearson's product-moment correlation coefficient

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}\ , \tag{7}$$

however, it measures only a *linear* dependence between random variables. In context of joined 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 are 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 $t = (c-d)/(c+d) = (c-d)/\binom{n}{2}$, where c is the number of concordant, d the number of discordant pairs (x_i, y_i) and (x_j, y_j), n the number of observations and $\binom{n}{2}$ the number of all distinct pairs in the sample; 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

$$\tau = \tau_{X,Y} = \text{Prob}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \text{Prob}[(X_1 - X_2)(Y_1 - Y_2) < 0] \,, \quad (8)$$

where we assume (X_1, Y_1) and (X_2, Y_2) 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 (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

$$Q = Q(C_1, C_2) = 4 \iint_{[0,1]^2} C_2(u, v) dC_1(u, v) - 1 \quad (9)$$

shows, that this function depends on the distributions of the two vectors only through their copulas C_1 and C_2. According to (9) the population version of Kendall's tau in terms of copulas is given by

$$\tau_{X,Y} = \tau_C = Q(C, C) = 4 \iint_{[0,1]^2} C(u, v) dC(u, v) - 1 \,, \quad (10)$$

where C is the copula of X and Y. Integral, which appears in (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:

$$\rho = \rho_{X,Y} = 3\Big(\text{Prob}[(X_1 - X_2)(Y_1 - Y_3) > 0] - \text{Prob}[(X_1 - X_2)(Y_1 - Y_3) < 0]\Big) \,, \quad (11)$$

(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

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

The coefficient "3" that appears in (11) and (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" 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) 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 volume between the graph of the copula C and the product copula Π over the unit square $[0, 1]^2$.

4. Archimedean copula

In this chapter we focus on an important class of copulas (introduced above) 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. 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.

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

$$C_\phi(u, v) = \phi^{-1}\Big(\phi(u) + \phi(v)\Big) \quad \text{for } u, v \in (0, 1] \tag{13}$$

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. $C(C(u, v), w) = C(u, C(v, w))$ 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 [9]. If the generator is twice differentiable and the copula is absolutely continuous, the copula density (probability density function of random variables U and V) is given by

$$c_\phi(u, v) = \frac{\partial^2 C_\phi(u, v)}{\partial u \partial v} = \frac{-\phi''(C_\phi(u, v))\phi'(u)\phi'(v)}{[\phi'(C_\phi(u, v))]^3} \tag{14}$$

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 and the range of dependence parameters. Tab.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.

Tab.1 Archimedean copulas with their generators.

Family of copulas	Generator $\phi(t)$	Parameter θ	Bivariate copula $C_\phi(u, v)$	Special cases
Independence	$-\ln t$		uv	C=Π
Gumbel	$(-\ln t)^\theta$	$\theta \geq 1$	$e^{-[(-\ln u)^\theta + (-\ln v)^\theta]^{-1/\theta}}$	C_1=Π, C_∞=M
Clayton	$t^{-\theta} - 1$	$\theta > 0$	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	C_0=Π, C_∞=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=Π $C_{-\infty}$=W, C_∞=M

Now that we're talking about dependence, recall the population version of Kendall's tau whose evaluation requires the evaluation of the double integral in (10). For an Archimedean copula, the situation is simpler, in that τ can be evaluated directly from the generator of the copula

$$\tau_C = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)} dt \tag{15}$$

[5]. 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). Tab.2 shows particular closed forms of (15).

Tab.2 Measures of association related to Archimedean copulas

Family	Independence	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.				

5. 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 & Rivest [6], that is also referred to as *nonparametric* estimation of copula parameter. Then we use *semi-parametric* estimation method developed in [7] and finally the experiment with bivariate geodetic data is given to illustrate the proposed theory. The procedures are also discussed in [4], [8] and [1].

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

5.1. Nonparametric estimation

As [4] formulate, measures of association summarize information in the copula concerning the dependence, or association, between random variables. Thus, following [6] 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 $K(z) = \text{Prob}[Z_i \leq z]$. This distribution function is related to the generator of an Archimedean copula through the expression

$$K(z) = K_\phi(z) = z - \frac{\phi(z)}{\phi'(z)}\,. \tag{16}$$

To identify ϕ, we:

1. Find Kendall's tau using the usual (nonparametric or distribution-free) estimate
$$\tau_n = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \text{Sign}[(X_i - X_j)(Y_i - Y_j)]\,.$$
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 $i = 1, \ldots, n$:
$$Z_i = (n-1)^{-1} \sum_{j=1}^{n} \text{If}\,[X_j < X_i \;\&\&\; Y_j < Y_i, 1, 0]\,,$$

b) second, construct the estimate of K as proportion of $Z_i's \leq z$, that is

$$K_n(z) = n^{-1} \sum_{i=1}^{n} \mathrm{If}\left[Z_i \leq z, 1, 0\right],$$

where function If [*condition*, 1, 0] gives 1 if *condition* holds, and 0 otherwise. "&&" stands for logic operator "and".

3. Now construct a parametric estimate K_ϕ using the relationship (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 Tab.1, and for linking τ to θ, Tab.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 $\int_0^1 [K_{\phi_n}(z) - K_n(z)]^2 dz$ or graphically by (a) plotting of $z - K(z)$ versus z or (b) corresponding quantile-quantile (Q-Q) plots (see [6], [4], [2]). 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.

5.2. Semi-parametric estimation

To estimate dependence parameter θ, two strategies can be envisaged. First, the 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 [7] 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 previously, we look for $\hat{\theta}$ that maximizes the pseudo log-likelihood

$$L(\theta) = \sum_{i=1}^{n} \log\Big(c_\theta(F_n(x), G_n(y))\Big), \tag{17}$$

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

$$F_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathrm{If}\left[X_i \leq x, 1, 0\right], \tag{18}$$

$G_n(y)$ 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 [7] 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 (14). To examine a goodness of our estimation, there is the Akaike information criterion available for comparison: $AIC = -2(log\text{-}likelihood) + 2k$, where k is the number of parameters in the model (in our case, $k = 1$). The lowest AIC value determines the best estimator.

5.3. Application to point co-ordinate time-series analysis

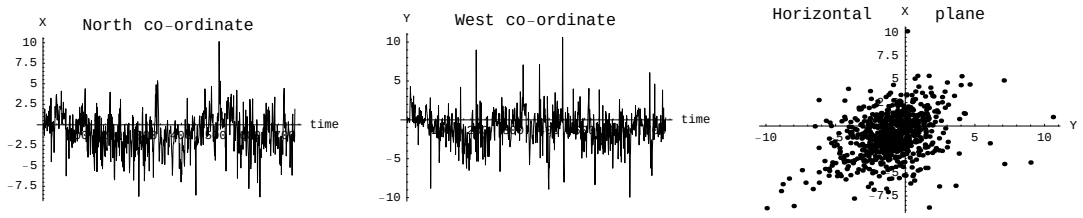


Fig.1 Two univariate time-series linked together to form bivariate random vector of a point location

Finally we have come to an experiment, that is to illustrate the above procedures. We employed bivariate time series - daily observations of plane co-ordinates of a point gathered 2 years, (which gives 728 realizations). Observations were made by means of NAVSTAR Global positioning system (GPS) on permanent station MOPI that takes part in European Reference Network. Establishment of such a network serves for various geodetic and geophysical purposes, e.g. for regular monitoring of recent kinematics of the Earth's crust (local, regional and global). The two random variables that make our bivariate observations thus share common physical phenomenon through the geometry and time reference. Indeed, as seen on Fig.1, we may expect some dependence. By default, position of a point is given in horizontal topocentric co-ordinate system, whose axes has north-east orientation (the third - vertical - we do not consider), however we swapped the east direction for the west, because the original configuration gives negative dependence and some copulas can model only positive dependencies.

The data was processed as follows. Firstly, we examined the two individual univariate time-series. Interestingly, both of them follow logistic distribution rather than normal. The logistic distribution with *mean* and *scale* parameter is frequently used in place of the normal distribution when a distribution with longer tails is desired. Nevertheless, further on we worked solely with the empirical marginal distribution function (18) to avoid any influence of a biased marginal model upon estimation of dependence structure. Next we computed scalar representatives of this structure, that is, measures of dependence

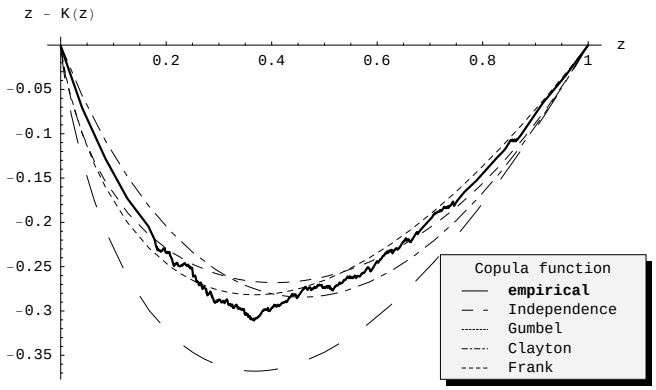
Correlation coefficient	Spearman's ρ	Kendall's τ
0.3670	0.3314	0.2343 .

Note that, if the data were nonstationary and required some variance stabilizing such as logarithmic transformation (which is strictly increasing), the pre-processing would have biased only the correlation coefficient, and none of the others.

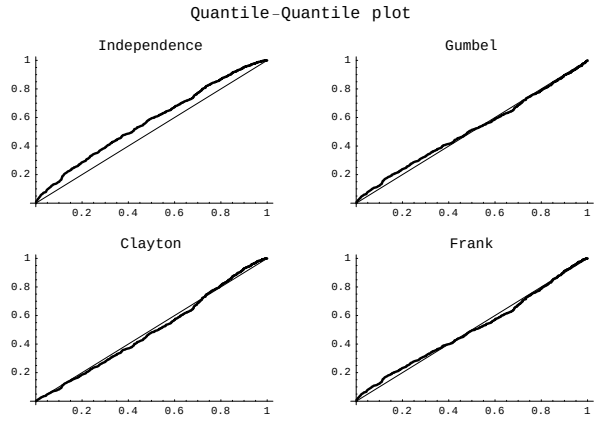
Following nonparametric procedure described in section 5.1, we estimated K_n, and using Kendall's τ also the three parametric estimates corresponding to

each one-parameter copula from Tab.1. Then, Fig.2 shows their "closeness" to K_n graphically, while Tab.3 numerically.

Within a semi-parametric procedure, (a) we firstly applied the procedure outlined in section 5.2, (b) then as an alternative (and as a backup too) we utilized nonlinear parametric least-square fit to empirical copula. For linking both (a) and (b) approaches, we computed L_2-norm distance between estimated and empirical copula. As seen from Tab.3, the differences are nonsignificant and in preferring Gumbel family to Frank and Clayton both methods agree with the nonparametric one. However, there seems to be a disharmony with AIC criterion of maximum likelihood estimate goodness, which surprisingly promotes the Clayton. On that account we performed some computations under different input conditions and figured out, that log-likelihood function of Clayton copula density (see Fig.3) is pretty sensitive to lower tail dependencies, namely to "perfect" extremes in data (notice the lower tail protruders in the very right-hand plot of Fig.1). Even just one (the most extremal) outlier chopped off from the lower tail of the data pushed the AIC of Clayton to between Frank and Gumbel. Dropping the other two degraded Clayton into "least appropriate" position among copulas under consideration. Upper tail extremes have no evident impact to Clayton likelihood estimate.



a)



b)

Fig.2 Graphical evaluation of nonparametric method:
a) Empirical function K_n fitted by K_ϕ of corresponding copula function
b) Quantile-quantile plots

This kinds of "revelations" appears to be quite important when choosing the best copula. Since nonlinear least-square fit demands a much more CPU time and memory[1], discussion of the nonparametric and semi-parametric (pseudo log-likelihood) is surely in order. As mentioned in [1], neither method is generally more convenient, but if there are outliers or if the marginal distributions are heavy tailed, it seems reasonable to choose the nonparametric approach. If we work with large data set, the likelihood estimator may be more precise.

Tab.3 Nonparametric and semi-parametric
estimates of copula dependence parameters θ

Family:	Gumbel	Clayton	Frank
Nonparametric procedure			
θ	1.3060	0.6120	2.2083
$d(K_\phi, K_n)$	0.445	0.542	0.492
Log-Likelihood procedure (semi-parametric)			
θ	1.3044	0.5638	2.3153
AIC	-106.2	-109.0	-90.7
$d(C_\theta, C_n)$	3.700	4.127	3.806
Nonlinear Fit procedure (semi-parametric)			
θ	1.3031	0.5595	2.103
$d(C_\theta, C_n)$	3.700	4.127	3.598

Influences on estimation is a subject to study and one has to be clear about what he prefer to understand, whether it is extreme situations, overall dependence structure or anything else. There are many other families of copula, that could be estimated by above procedures and, if necessary, should be considered as the alternatives to the three but mainly to most used Gaussian distribution, which - by its nature - cannot be satisfactory in numerous applications. In that of ours, the sum of squares of residuals unambiguously refused the appropriateness of bi-normal distribution.

6. Conclusion

From the very beginning of our paper we have outlined an approach of multivariate statistical analysis, that contemplates entirely the dependence structure, keeping individual variable properties isolated for optional concern. The approach is based on multivariate distribution function named copula, and we have provided a quick survey of definitions, properties, relation to dependence measures and a special class of copulas in order to interest any researcher in seeking new applications for this promising tool. As the copula functions are parametric families, an ordinary nonlinear least-squares fit can be applied for estimation. Moreover we described

[1]All the computation was made in Mathematica 5.0 on 1.2GHz CPU system with 256MB of RAM. With this configuration and for all the three copulas together the nonparametric procedure took few seconds, semi-parametric log-likelihood few minutes and least-square fit more than 2 hours. Usually, the size of fast physical memory RAM plays the crucial role when handling with a large data, however the most time consuming procedure here (summing up frequencies into the empirical distribution function) has the full CPU usage for about two hours without needing any access to a slow complementary (virtual) memory.

here other two methods, nonparametric and semi-parametric maximum likelihood. Also, an application to a position dynamics of GPS permanent station MOPI drew our attention to some pitfalls of a particular copula and method selection, more specifically the impact of tail dependencies in data.

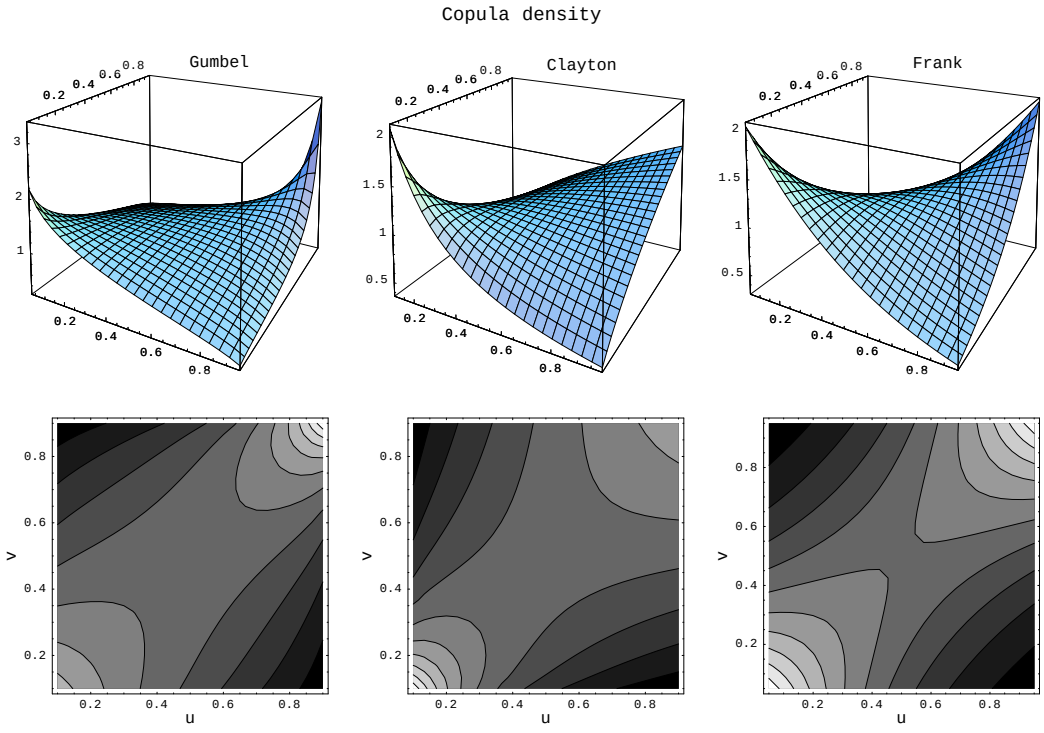


Fig.3 Copula density for three Archimedean families

Acknowledgements

I would like to document my immense gratitude to Professor Magda Komorníková and Professor Radko Mesiar for their encouragement and helpful comments.

Research supported by grants VEGA 1/1145/04 and APVT-20-003204

References

[1] Abid, F., Naifar, N.: The Impact of Stock Returns Volatility on Credit Default Swap Rates: A copula study, "http://www.defaultrisk.com/pdf__files/ The_Impact_o_Stock_Returns_Volatility_CDS_Rates.pdf" (2001).

[2] Durrleman, V., Nikeghbali, A., Roncali, T.: Which copula is the right one?, Working paper, Groupe de Recherche Operationnelle, Credit Lyonnais (2000).

[3] Embrechts, P., Lindskog, F., McNeil, A.: Modelling Dependence with Copulas and Applications to Risk Management, Handbook of Heavy Tailed Distributions in Finance, ed. S. Rachev, Elsevier (2001), pp. 329-384.

[4] Frees, E.W., Valdez, E.A.: Understanding Relationships Using Copulas, North American Actuarial Journal 2 (1998), pp. 1-25.

[**5**] Genest, C., MacKay, J.: The Joy of Copulas: Bivariate Distributions with Uniform Marginals, The American Statistician 40 (1986), pp. 280-283.

[**6**] Genest, C., Rivest, L.: Statistical Inference Procedures for Bivariate Archimedean Copulas, Journal of the American Statistical Association 88 (1993), pp. 1034-1043.

[**7**] Genest, C., Ghoudi, K., Rivest, L.: A Semi-parametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions, Biometrika 82 (1995), pp. 543-552.

[**8**] Melchiori, M.R.: Which Archimedean Copula is the right one?, YeldCurve.com e-Journal (2003).

[**9**] Nelsen, R.B.: An Introduction to Copulas, Lecture Notes in Statistics, Vol. 139, Springer (1998).

Author: Ing. Tomáš Bacigál, Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering SUT, Radlinského 11, 813 68 Bratislava,
e-mail: bacigal@math.sk