Department of Mathematics
Faculty of Mechanical Engineering
Slovak University of Technology in Bratislava
Apli mat
predbežná hlavička
4th International Conference
APLIMAT 2005

Fitting Archimedean copulas to bivariate geodetic observations.

BACIGÁL Tomáš [1] [2]

Abstrakt. Kopule sú funkcie, ktoré prepájajú jednorozmerné marginálne rozdelenia náhodných premenných s ich zdruzenými rozdelením, teda modelujú výhradne ich vzájomný vzťah. V našej práci sa zameriavame na tri najznámejsie rodiny Archimedovských kopúl a popisujeme spôsob ich odhadu, v ktorom sa využíva fakt, že Archimedovské kopule sú generované jednorozmernou funkciou. Na príklade modelovania dvojrozmerných časových radov pozorovaní polohy bodu demonštrujeme neparametrický a semi-parametrický prístup k odhadu parametrov kopule. Zaujímavosťou je odhad lineárnej kombinácie dvoch kopúl, ktorá je tiež kopulou a dokáže výrazne lepšie aproximovat skutočné rozdelenie pravdepodobnosti.

Abstract. Copulas are functions, that link univariate marginals to their joint distribution function. Thus, applied to multivariate observations, copula captures entirely the relationships among individual variables. In our paper we focus on three Archimedean families and outline their estimation, which reflects the fact that Archimedean copula is built from a single univariate function. Nonparametric and semi-parametric procedures are considered and an application to modelling bivariate point position time-series is given. Moreover, we show that linear convex combinations of any two copulas (which is still a copula) can significantly improve their fit to empirical data.

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 skips the general introduction to copula theory, interested reader is referred to [9], [3] and others. To briefly line out the concept of a copula function as a tool for relating

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

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

different dimensions of an data output, define the bivariate joint distribution function H of random variables X and Y, in terms of copula C and marginals F and G

$$H(X,Y) = C\Big(F(X), G(Y)\Big). \tag{1}$$

Then copula $C(U,V)$, where $U = F(X)$ and $V = G(X)$ are random variables uniformly distributed on [0,1], captures the dependency structure of $H(X,Y)$. Every copula is bounded by Fréchet-Hoeffding lower $W(u,v) = \max(u+v-1, 0)$ and upper $M(u,v) = \min(u,v)$ bound, that represent perfect (negative and positive) dependence, while copula $\Pi(u,v) = uv$ means perfect independence.

2 Archimedean copulas

In this chapter we focus on an important class of copulas known as Archimedean. 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.

The Archimedean representation allows us to reduce the study of a multivariate copula to a single univariate function. For simplicity, we consider bivariate copulas. 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

$$C_\phi(u,v) = \phi^{-1}\Big(\phi(u) + \phi(v)\Big) \quad \text{for } u, v \in (0,1] \tag{2}$$

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. 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

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

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. Table 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.

The dependence parameters are tied with the measures of association, most used being Kendal's tau and Spearman's rho, that capture more than a linear dependence unlike the well known correlation coefficient. For the Archimedean copulas, Kendall's tau τ can be evaluated directly from the generator

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

Table 1: Archimedean copulas with their generators.

Family of copulas	Generator $\phi(t)$	Prameter 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

[5], instead of the more general evaluation from copula function through double integral. 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 2 shows particular closed forms of (4).

Table 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.				

3 Fitting a copula to bivariate data

For identifying the copula, we focus on the procedure of [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], [1]. In our application, we consider the three most widely used Archimedean families of copula: Clayton, Gumbel and Frank.

3.1 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

$$K(z) = K_\phi(z) = z - \frac{\phi(z)}{\phi'(z)}\,. \tag{5}$$

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
   $Z_i = (n-1)^{-1} \sum_{j=1}^{n} \text{If}\,[X_j < X_i \;\&\&\; Y_j < Y_i, 1, 0]$,for $i = 1, \ldots, n$

   b) second, construct the estimate of K
   $K_n(z) = n^{-1} \sum_{i=1}^{n} \text{If}\,[Z_i \leq z, 1, 0]$,

   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 (5). 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 1, and for linking τ to θ, Table 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.

3.2 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

$$L(\theta) = \sum_{i=1}^{n} \log\Big(c_\theta(F_n(x), G_n(y))\Big) \;, \tag{6}$$

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

$$F_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} \text{If}\,[X_i \leq x, 1, 0] \,, \tag{7}$$

$G_n(y)$ 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 the independence case.

The copula density c_θ for each Archimedean copula can be acquired from (3). 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.

3.3 Application to point co-ordinate time-series analysis

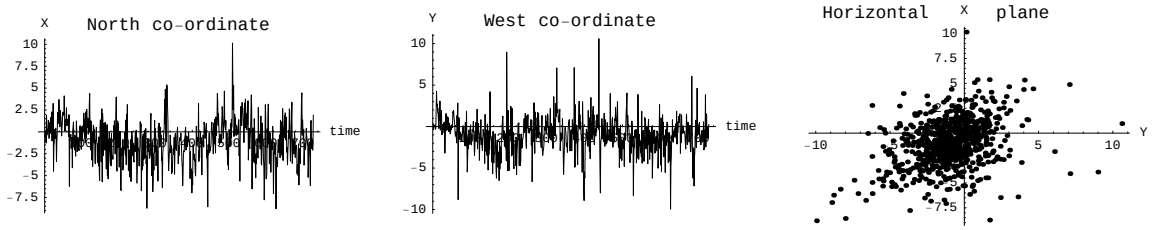
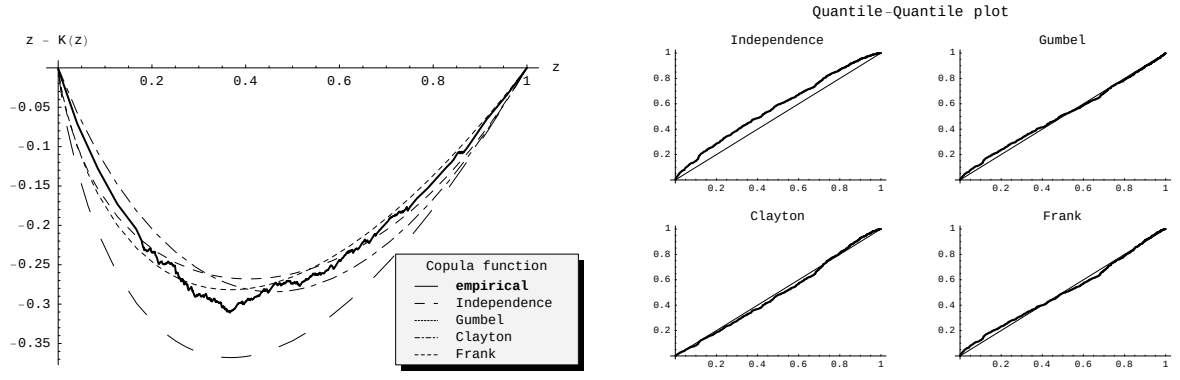


Figure 1: Two univariate time-series linked together to form bivariate random vector of a point location

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 Figure 1, we may expect some dependence.



a) b)

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

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 (7) 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 coef.	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 3.1, we estimated K_n, and using Kendall's τ also the three parametric estimates corresponding to each one-parameter copula from Table 1. Then, Figure 2 shows their "closeness" to K_n graphically, while Table 3 numerically.

Table 3: Nonparametric and semi-parametric estimates of copula dependence parameters

a)

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

b)

Linear convex combination:	Clayton-Gumbel	Clayton-Frank	Frank-Gumbel
Nonlinear Fit procedure (semi-parametric)			
α	0.4507	0.3714	0.5548
$d\Big(\alpha\, C_{\theta_1} + (\alpha-1)C_{\theta_2},\, C_n\Big)$	2.609	3.280	3.407

Within a semi-parametric procedure, (a) we firstly applied the procedure outlined in section 3.2, (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 Table 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 Figure 2). 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.

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, 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.

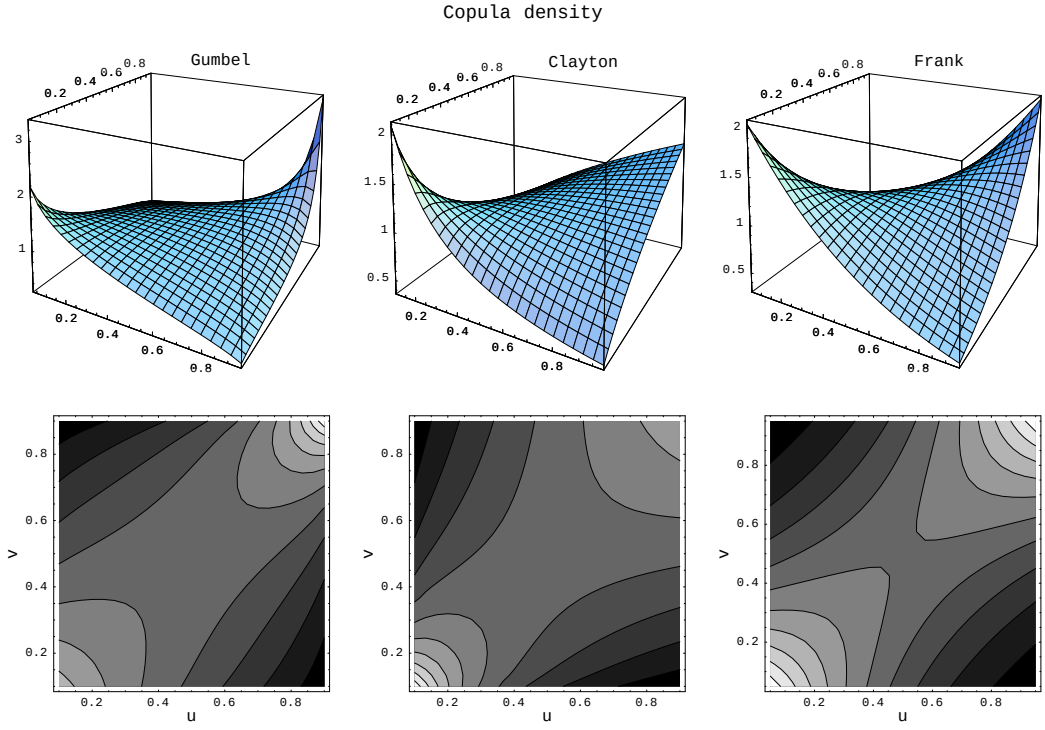


Figure 3: Copula density for three Archimedean families

There are many families of copula, that could be estimated by above procedures and, if necessary, should be considered as the alternatives to the three above 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.

Finally, as we have estimated the copula parameters by particular method and chose "the best" of them, we contemplated a possibility to improve the nonlinear fit of parametric copulas by simply fitting their linear convex combinations to empirical copula and compare the L_2 distances. It can be shown, that the linear convex combination $\alpha C_1 + (\alpha - 1)C_2$ of any two copulas C_1 and C_2 is also a copula with parameter $\alpha \in [0, 1]$. Such a copula may posses benefits of both parents when fitting empirical copula. And

indeed, Table 3 supports this assumption. The best combination is given by Clayton and Gumbel, with a slight dominance of Gumbel.

4 Conclusion

From the very beginning of our paper we have outline 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, though we have described here other two methods, nonparametric and semi-parametric maximum likelihood, that dispose of rationality in computation. 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. As a highlight of this all, we have improved the copula model with linear convex combination of different pairs of copulas.

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?*, 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, pp. 329-384, 2001.

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

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

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

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

[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.

Contact address

Ing. Tomáš Bacigál, Stavebná fakulta STU Radlinského 11, 813 68 Bratislava, bacigal@math.sk.