Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask

Christian Genest[1] and Anne-Catherine Favre[2]

Abstract: This paper presents an introduction to inference for copula models, based on rank methods. By working out in detail a small, fictitious numerical example, the writers exhibit the various steps involved in investigating the dependence between two random variables and in modeling it using copulas. Simple graphical tools and numerical techniques are presented for selecting an appropriate model, estimating its parameters, and checking its goodness-of-fit. A larger, realistic application of the methodology to hydrological data is then presented.

DOI: 10.1061/(ASCE)1084-0699(2007)12:4(347)

CE Database subject headings: Frequency analysis; Distribution functions; Risk management; Statistical models.

Introduction

Hydrological phenomena are often multidimensional and hence require the joint modeling of several random variables. Traditionally, the pairwise dependence between variables such as depth, volume, and duration of flows has been described using classical families of bivariate distributions. Perhaps the most common models occurring in this context are the bivariate normal, lognormal, gamma, and extreme-value distributions. The main limitation of this approach is that the individual behavior of the two variables (or transformations thereof) must then be characterized by the same parametric family of univariate distributions.

Copula models, which avoid this restriction, are just beginning to make their way into the hydrological literature; see, e.g., De Michele and Salvadori (2002), Favre et al. (2004), Salvadori and De Michele (2004), and De Michele et al. (2005). Restricting attention to the bivariate case for the sake of simplicity, the copula approach to dependence modeling is rooted in a representation theorem due to Sklar (1959). The latter states that the joint cumulative distribution function (c.d.f.) $H(x,y)$ of any pair (X,Y) of continuous random variables may be written in the form

$$H(x,y) = C\{F(x),G(y)\}, \quad x,y \in \mathbb{R} \tag{1}$$

where $F(x)$ and $G(y)$=marginal distributions; and C:$[0,1]^2 \rightarrow [0,1]$=copula.

While Sklar (1959) showed that C, F, and G are uniquely determined when H is known, a valid model for (X,Y) arises from Eq. (1) whenever the three "ingredients" are chosen from given parametric families of distributions, viz.

$$F \in (F_\delta), \quad G \in (G_\eta), \quad C \in (C_\theta)$$

Thus, for example, F might be normal with (bivariate) parameter $\delta=(\mu,\sigma^2)$; G might be gamma with parameter $\eta=(\alpha,\lambda)$; and C might be taken from the Farlie–Gumbel–Morgenstern family of copulas, defined for each $\theta \in [-1,1]$ by

$$C_\theta(u,v) = uv + \theta uv(1-u)(1-v), \quad u,v \in [0,1] \tag{2}$$

The main advantage provided to the hydrologist by this approach is that the selection of an appropriate model for the dependence between X and Y, represented by the copula, can then proceed independently from the choice of the marginal distributions.

For an introduction to the theory of copulas and a large selection of related models, the reader may refer, e.g., to the monographs by Joe (1997) and Nelsen (1999), or to reviews such as Frees and Valdez (1998) and Cherubini et al. (2004), in which actuarial and financial applications are considered. While the theoretical properties of these objects are now fairly well understood, inference for copula models is, to an extent, still under development. The literature on the subject is yet to be collated, and most of it is not written with the end user in mind, making it difficult to decipher except for the most mathematically inclined.

The aim of this paper is to present, in the simplest terms possible, the successive steps required to build a copula model for hydrological purposes. To this end, a fictitious data set of (very) small size will be used to illustrate the diagnostic and inferential tools currently available. Although intuition will be given for the various techniques to be presented, emphasis will be put on their implementation, rather than on their theoretical foundation. Therefore, computations will be presented in more detail than usual, at the expense of exhaustive mathematical exposition, for which the reader will only be given appropriate references.

The pedagogical data set to be used throughout the paper is introduced in the "Dependence and Ranks" section, where it will be explained why statistical inference concerning dependence structures should always be based on ranks. This will lead, in the "Measuring Dependence" section, to the description of classical nonparametric measures of dependence and tests of independence. Exploratory tools for uncovering dependence and measuring it will be reviewed in the "Additional Graphical Tools for Detecting Dependence" section. Point and interval estimation for

[1]Professor, Dépt. de mathématiques et de statistique, Univ. Laval, Québec QC, Canada G1K 7P4.

[2]Professor, Chaire en Hydrologie Statistique, INRS, Eau, Terre et Environnement, Québec QC, Canada G1K 9A9.

Note. Discussion open until December 1, 2007. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on August 29, 2006; approved on August 29, 2006. This paper is part of the ***Journal of Hydrologic Engineering***, Vol. 12, No. 4, July 1, 2007. ©ASCE, ISSN 1084-0699/2007/4-347–368/$25.00.

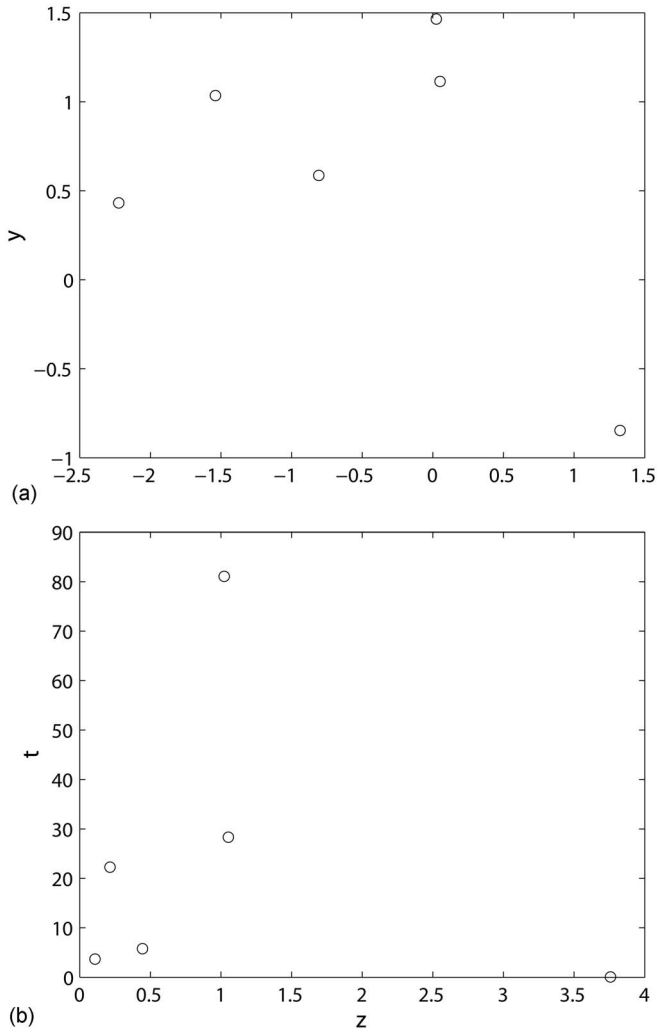


Fig. 1. (a) Conventional scatter plot of the pairs (X_i, Y_i); (b) corresponding scatter plot of the pairs $(Z_i, T_i) = (e^{X_i}, e^{3Y_i})$

dependence parameters from copula models will then be presented in the "Estimation" section. Recent goodness-of-fit techniques will be illustrated in the "Goodness-of-Fit Tests" section. The "Application" section will discuss in detail a concrete hydrological implementation of this methodology. This will lead to the consideration of additional tools for the treatment of extreme-value dependence structures in the "Graphical Diagnostics for Bivariate Extreme-Value Copulas" section. Final remarks will then be made in the "Conclusion" section.

Dependence and Ranks

Suppose that a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is given from some pair (X, Y) of continuous variables, and that it is desired to identify the bivariate distribution $H(x, y)$ that characterizes their joint behavior. In view of Sklar's representation theorem, there exists a unique copula C for which identity, Eq. (1), holds. Therefore, just as $F(x)$ and $G(y)$ give an exhaustive description of X and Y taken separately, the joint dependence between these variables is fully and uniquely characterized by C.

It is easy to see, for example, that X and Y are stochastically independent if and only if $C = \Pi$, where $\Pi(u, v) = uv$ for all $u, v \in [0, 1]$. At the other extreme, it can also be shown that in order for Y to be a deterministic function of X, C must be either one of the two copulas

$$W(u,v) = \max(0, u + v - 1) \quad \text{or} \quad M(u,v) = \min(u,v)$$

which are usually referred to as the Fréchet–Hoeffding bounds in the statistical literature; see, e.g., Fréchet (1951) or Nelsen (1999, p. 9). When $C = W$, Y is a decreasing function of X, while Y is monotone increasing in X when $C = M$. More generally, any copula C represents a model of dependence that lies somewhere between these two extremes, a fact that translates into the inequalities

$$W(u,v) \leqslant C(u,v) \leqslant M(u,v), \quad u,v \in [0,1]$$

To get a feeling for the dependence between X and Y, it is traditional to look at the scatter plot of the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. Such a representation is given in Fig. 1(a) for the following fictitious random sample of size $n = 6$ from the bivariate standard normal distribution with zero correlation. This example will be used for illustration purposes throughout the paper.

Table 1. Learning Data Set

i	1	2	3	4	5	6
X_i	−2.224	−1.538	−0.807	0.024	0.052	1.324
Y_i	0.431	1.035	0.586	1.465	1.115	−0.847

Learning Data Set

Table 1 shows six independent pairs of mutually independent observations X_i, Y_i generated from the standard $\mathcal{N}(0, 1)$ distribution using the statistical freeware R (R Development Core Team 2004). For simplicity, and without loss of generality, the pairs were labeled in such a way that $X_1 < \cdots < X_6$.

While there is nothing fundamentally wrong with looking at the pattern of the pairs (X_i, Y_i) (for example, to look for linear association), it must be realized that this picture does not only incorporate information about the dependence between X and Y, but also about their marginal behavior. To drive this point home, consider the transformed pairs

$$Z_i = \exp(X_i), \quad T_i = \exp(3Y_i), \quad 1 \leqslant i \leqslant 6$$

whose scatter plot, shown in Fig. 1(b), is drastically different from the original one.

In effect, both pictures are distortions of the dependence between the pairs (X, Y) and (Z, T), which is characterized by the *same* copula, C, whatever it may be. More generally, if φ and ψ are two increasing transformations with inverses φ^{-1} and ψ^{-1}, the copula of the pair (Z, T) with $Z = \varphi(X)$ and $T = \psi(Y)$ is the *same* as that of (X, Y). Let

$$H^*(z,t) = C^*\{F^*(z), G^*(t)\} \tag{3}$$

be the Sklar representation of the joint distribution of the pair (Z, T). Since the marginal distributions of Z and T are given by

$$F^*(z) = \mathrm{P}(Z \leqslant z) = \mathrm{P}\{X \leqslant \varphi^{-1}(z)\} = F\{\varphi^{-1}(z)\}$$

and

$$G^*(t) = \mathrm{P}(T \leqslant t) = \mathrm{P}\{Y \leqslant \psi^{-1}(t)\} = G\{\psi^{-1}(t)\}$$

Table 2. Ranks for the Learning Data Set of Table 1

i	1	2	3	4	5	6
R_i	1	2	3	4	5	6
S_i	2	4	3	6	5	1

one has

$$H^*(z,t) = \mathrm{P}(Z \leqslant z, T \leqslant t) = \mathrm{P}\{X \leqslant \varphi^{-1}(z), Y \leqslant \psi^{-1}(t)\}$$
$$= H\{\varphi^{-1}(z), \psi^{-1}(t)\} = C[F\{\varphi^{-1}(z)\}, G\{\psi^{-1}(t)\}]$$
$$= C\{F^*(z), G^*(t)\} \quad (4)$$

for all choices of $z,t \in \mathbb{R}$. It follows at once from the comparison of Eqs. (3) and (4) that $C^* = C$.

Expressed in different terms, the above development means that the unique copula associated with a random pair (X,Y) is *invariant* by monotone increasing transformations of the marginals. Since the dependence between X and Y is characterized by this copula, a faithful graphical representation of dependence should exhibit the same invariance property. Among functions of the data that meet this requirement, it can be seen easily that the pairs of ranks

$$(R_1,S_1), \ldots, (R_n,S_n)$$

associated with the sample are the statistics that retain the greatest amount of information; see, e.g., Oakes (1982). Here, R_i stands for the rank of X_i among $X_1, \ldots, X_n$, and S_i stands for the rank of Y_i among $Y_1, \ldots, Y_n$. These ranks are unambiguously defined, because ties occur with probability zero under the assumption of continuity for X and Y. Pairs of ranks corresponding to the learning data set are given in Table 2.

Displayed in Fig. 2(a) is the scatter plot of the pairs (R_i,S_i) corresponding to these (X_i,Y_i). Fig. 2(b) shows the graph of the pairs (R_i^*,S_i^*) associated with the (Z_i,T_i). The result is obviously the same. It is the most judicious representation of the copula C that one could hope for. Upon rescaling of the axes by a factor of $1/(n+1)$, one gets a set of points in the unit square $[0,1]^2$, which form the domain of the so-called *empirical copula* (Deheuvels 1979), formally defined by

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\frac{R_i}{n+1} \leqslant u, \frac{S_i}{n+1} \leqslant v\right)$$

with $1(A)$ denoting the indicator function of set A. For any given pair (u,v), it may be shown that $C_n(u,v)$ is a rank-based estimator of the unknown quantity $C(u,v)$ whose large-sample distribution is centered at $C(u,v)$ and normal.

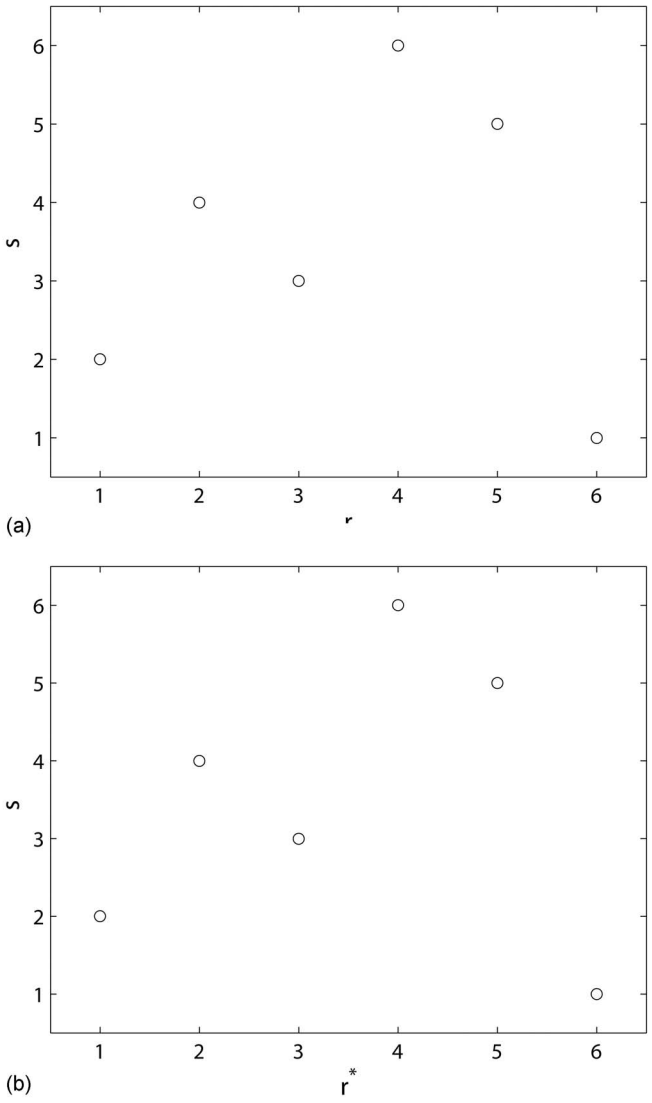


Fig. 2. Displayed in (a) is a scatter plot of the pairs (R_i,S_i) of ranks derived from the learning data set (X_i,Y_i), $1 \leqslant i \leqslant 6$. As for (b), it shows a scatter plot of the pairs (R_i^*,S_i^*) of ranks derived from the transformed data $(Z_i,T_i)=(\exp(X_i),\exp(3Y_i))$, $1 \leqslant i \leqslant 6$. For obvious reasons, the two graphs are actually identical.

Measuring Dependence

It was argued above that the empirical copula C_n is the best sample-based representation of the copula C, which is itself a characterization of the dependence in a pair (X,Y). It would make sense, therefore, to measure dependence, both empirically and theoretically, using C_n and C, respectively. It will now be explained how this leads to two well-known nonparametric measures of dependence, namely Spearman's rho and Kendall's tau.

Spearman's Rho

Mimicking the familiar approach of Pearson to the measurement of dependence, a natural idea is to compute the correlation between the pairs (R_i,S_i) of ranks, or equivalently between the points $(R_i/(n+1), S_i/(n+1))$ forming the support of C_n. This leads directly to Spearman's rho, viz.

$$\rho_n = \frac{\sum_{i=1}^{n}(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\bar{R})^2 \sum_{i=1}^{n}(S_i-\bar{S})^2}} \in [-1,1]$$

where

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{n+1}{2} = \frac{1}{n}\sum_{i=1}^{n} S_i = \bar{S}$$

This coefficient, which may be expressed more conveniently in the form

$$\rho_n = \frac{12}{n(n+1)(n-1)} \sum_{i=1}^{n} R_i S_i - 3\frac{n+1}{n-1}$$

shares with Pearson's classical correlation coefficient, r_n, the property that its expectation vanishes when the variables are independent. However, ρ_n is theoretically far superior to r_n, in that

1. $E(\rho_n) = \pm 1$ occurs if and only if X and Y are *functionally* dependent, i.e., whenever their underlying copula is one of the two Fréchet–Hoeffding bounds, M or W;
2. In contrast, $E(r_n) = \pm 1$ if and only if X and Y are *linear* functions of one another, which is much more restrictive; and
3. ρ_n estimates a population parameter that is *always* well defined, whereas there are heavy-tailed distributions (such as the Cauchy, for example) for which a theoretical value of Pearson's correlation does not exist.

For additional discussion on these points, see, e.g., Embrechts et al. (2002).

As it turns out, ρ_n is an asymptotically unbiased estimator of

$$\rho = 12 \int_{[0,1]^2} uv dC(u,v) - 3 = 12 \int_{[0,1]^2} C(u,v) dv du - 3$$

where the second equality is an identity originally proven by Hoeffding (1940) and extended by Quesada-Molina (1992). To show this, one may use the fact that

$$12 \int_{[0,1]^2} uv dC_n(u,v) - 3 = \frac{12}{n} \sum_{i=1}^{n} \frac{R_i}{n+1} \frac{S_i}{n+1} - 3 = \frac{n-1}{n+1} \rho_n$$

and that $C_n \to C$ as $n \to \infty$. For more precise conditions under which this result holds, see, e.g., Hoeffding (1948).

Note in passing that under the null hypothesis $H_0: C = \Pi$ of independence between X and Y, the distribution of ρ_n is close to normal with zero mean and variance $1/(n-1)$, so that one may reject H_0 at approximate level $\alpha = 5\%$, for instance, if $\sqrt{n-1}\,|\rho_n| > z_{\alpha/2} = 1.96$.

Example

For the observations from the learning data set, a simple calculation yields $\rho_n = 1/35 = 0.028$, while $r_n = -0.397$. Here, there is no reason to reject the null hypothesis of independence. For, if $\mathcal{Z}$ is a standard normal random variable, the P-value of the test based on ρ_n is $2\Pr(\mathcal{Z} > \sqrt{5}/35) = 94.9\%$.

Given a family (C_θ) of copulas indexed by a real parameter, the theoretical value of ρ is, typically, a monotone increasing function of θ. A sufficient condition for this is that the copulas be ordered by positive quadrant dependence (PQD), which means that the implication $\theta < \theta' \Rightarrow C_\theta(u,v) \leq C_{\theta'}(u,v)$ is true for all $u, v \in [0,1]$. The original definition of PQD as a concept of dependence goes back to Lehmann (1966); the same ordering, rediscovered by Dhaene and Goovaerts (1996) in an actuarial context, is often referred to as the correlation or concordance ordering in that field.

In the Farlie–Gumbel–Morgenstern model, for example, one has

$$\int_{[0,1]^2} uv dC_\theta(u,v) = \int_0^1 \int_0^1 uv c_\theta(u,v) dv du$$

where

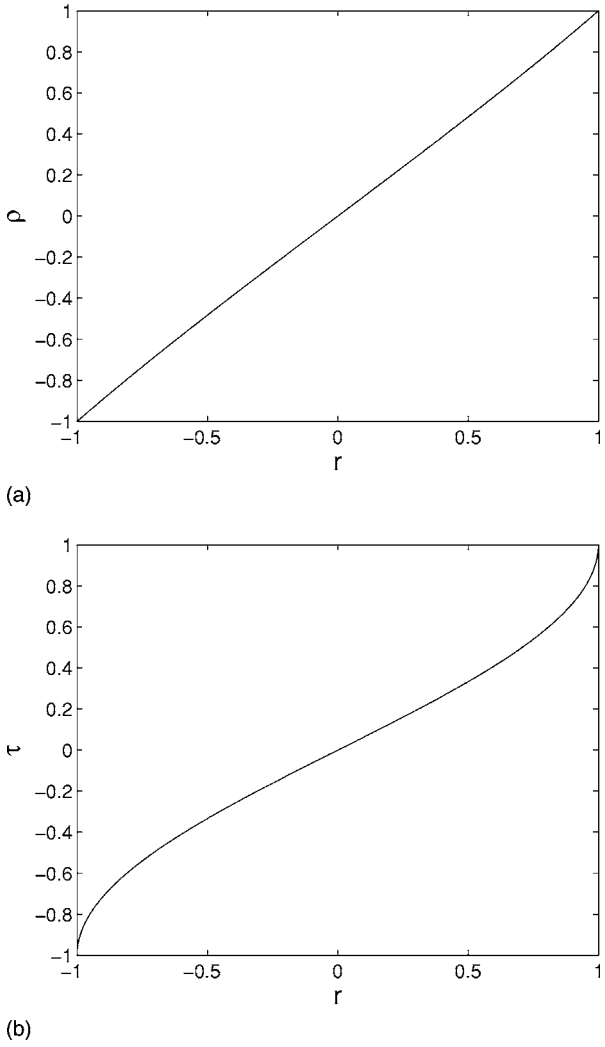


Fig. 3. Spearman's rho (a) and Kendall's tau (b) as a function of Pearson's correlation in the bivariate normal model

$$c_\theta(u,v) = \frac{\partial^2}{\partial u \partial v} C_\theta(u,v) = 1 + \theta(1-2u)(1-2v)$$

since C_θ is absolutely continuous in this case. A simple calculation then yields

$$\int_0^1 \int_0^1 uv c_\theta(u,v) dv du = \frac{1}{4} + \frac{\theta}{36}$$

and, hence, $\rho = \theta/3$, as initially shown by Schucany et al. (1978).

As a second example, if (X,Y) follows a bivariate normal distribution with correlation r, a somewhat intricate calculation to be found, e.g., in Kruskal (1958), shows that

$$\rho = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x) G(y) dH(x,y) - 3 = \frac{6}{\pi} \arcsin\left(\frac{r}{2}\right)$$

For those people accustomed to thinking in terms of r, the above formula may suggest that a serious effort would be required to think of correlation in terms of Spearman's rho in the traditional bivariate normal model. As shown in Fig. 3(a), however, the difference between ρ and r is minimal in this context.

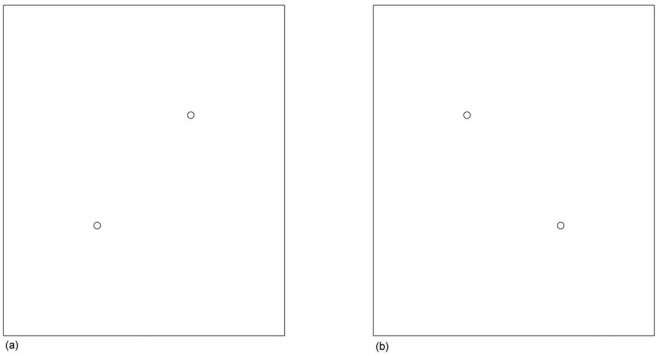


Fig. 4. Two pairs of concordant (a) and discordant (b) observations

Kendall's Tau

A second, well-known measure of dependence based on ranks is Kendall's tau, whose empirical version is given by

$$\tau_n = \frac{P_n - Q_n}{\binom{n}{2}} = \frac{4}{n(n-1)} P_n - 1 \tag{5}$$

where P_n and Q_n=number of concordant and discordant pairs, respectively. Here, two pairs (X_i, Y_i), (X_j, Y_j) are said to be concordant when $(X_i - X_j)(Y_i - Y_j) > 0$, and discordant when $(X_i - X_j)(Y_i - Y_j) < 0$. One need not worry about ties, since the borderline case $(X_i - X_j)(Y_i - Y_j) = 0$ occurs with probability zero under the assumption that X and Y are continuous. The characteristic patterns of concordant and discordant pairs are displayed in Fig. 4.

It is obvious that τ_n is a function of the ranks of the observations only, since $(X_i - X_j)(Y_i - Y_j) > 0$ if and only if $(R_i - R_j) \times (S_i - S_j) > 0$. Accordingly, τ_n is also a function of C_n. To make the connection, introduce

$$I_{ij} = \begin{cases} 1 & \text{if } X_j < X_i, Y_j < Y_i \\ 0 & \text{otherwise} \end{cases}$$

for arbitrary $i \neq j$, and let $I_{ii} = 1$ for all $i \in \{1, \ldots, n\}$. Observe that

$$P_n = \frac{1}{2} \sum_{i=1}^{n} \sum_{j \neq i} (I_{ij} + I_{ji}) = \sum_{i=1}^{n} \sum_{j \neq i} I_{ij} = -n + \sum_{i=1}^{n} \sum_{j=1}^{n} I_{ij}$$

since $I_{ij} + I_{ji} = 1$ if and only if the pairs (X_i, Y_i) and (X_j, Y_j) are concordant.

Now write

$$W_i = \frac{1}{n} \sum_{j=1}^{n} I_{ij} = \frac{1}{n} \#\{j : X_j \leq X_i, Y_j \leq Y_i\}$$

so that if $\bar{W} = (W_1 + \cdots + W_n)/n$, then $P_n = -n + n^2 \bar{W}$ and

$$\tau_n = 4 \frac{n}{n-1} \bar{W} - \frac{n+3}{n-1} \tag{6}$$

The connection with C_n then comes from the fact that by definition

$$W_i = C_n\left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right)$$

hence

$$\bar{W} = \int_{[0,1]^2} C_n(u,v) dC_n(u,v)$$

Using Eq. (6) and the fact that under suitable regularity conditions, $C_n \to C$ as $n \to \infty$, one can conclude [with Hoeffding (1948)] that τ_n is an asymptotically unbiased estimator of the population version of Kendall's tau, given by

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

An alternative test of independence can be based on τ_n, since under H_0, this statistic is close to normal with zero mean and variance $2(2n+5)/\{9n(n-1)\}$. Thus, H_0 would be rejected at approximate level $\alpha = 5\%$ if

$$\sqrt{\frac{9n(n-1)}{2(2n+5)}} |\tau_n| > 1.96$$

Example (Continued)

For the observations from the learning data set, a simple calculation yields $\tau_n = 1/15 = 0.067$. Here, there is no reason to reject the null hypothesis of independence. For, if $\mathcal{Z}$ is a standard normal random variable, the P-value of the test based on τ_n is $2\Pr(\mathcal{Z} > 0.188) = 85.1\%$.

As for Spearman's rho, the theoretical value of Kendall's tau is a monotone increasing function of the real parameter θ whenever a family (C_θ) of copulas is ordered by positive quadrant dependence. In the Farlie–Gumbel–Morgenstern model, for example, one has

$$\int_{[0,1]^2} C_\theta(u,v) dC_\theta(u,v) = \int_0^1 \int_0^1 C_\theta(u,v) c_\theta(u,v) dv du$$

which reduces to $\theta/18 + 1/4$, hence $\tau = 2\theta/9$, as per Schucany et al. (1978).

For the bivariate normal model with correlation r, Kruskal (1958) has shown that

$$\tau = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x,y) dH(x,y) - 1 = \frac{2}{\pi} \arcsin(r)$$

As shown in Fig. 3(b), τ is also nearly a linear function of r in this special case.

Other Measures and Tests of Dependence

Although Spearman's rho and Kendall's tau are the two most common statistics with which dependence is measured and tested, many alternative rank-based procedures have been proposed in the statistical literature. Most of them are based on expressions of the form

$$\int J(u,v) dC_n(u,v)$$

where J is some (suitably regular) score function. Thus, while $J(u,v) = uv$ is the basis of Spearman's statistic, as seen earlier, the choice $J(u,v) = \Phi^{-1}(u)\Phi^{-1}(v)$, e.g., yields the van der Waerden statistic. Genest and Verret (2005), who review this literature, explain how each J should be chosen so as to yield the most powerful testing procedure against a specific class of copula alternatives.

In the absence of privileged information about the suspected departure from independence, however, omnibus procedures such as those based on ρ_n and τ_n usually perform well. See Deheuvels (1981) or Genest and Rémillard (2004) for other general tests based on the empirical copula process $\mathbb{C}_n=\sqrt{n}(C_n-C)$.

Additional Graphical Tools for Detecting Dependence

Besides the scatter plot of ranks, two graphical tools for detecting dependence have recently been proposed in the literature, namely, chi-plots and K-plots. These will be briefly described in turn.

Chi-Plots

Chi-plots were originally proposed by Fisher and Switzer (1985) and more fully illustrated in Fisher and Switzer (2001). Their construction is inspired from control charts and based on the chi-square statistic for independence in a two-way table. Specifically, introduce

$$H_i=\frac{1}{n-1}\#\{j\neq i:X_j\leq X_i,Y_j\leq Y_i\}=\frac{nW_i-1}{n-1}$$

$$F_i=\frac{1}{n-1}\#\{j\neq i:X_j\leq X_i\}$$

and

$$G_i=\frac{1}{n-1}\#\{j\neq i:Y_j\leq Y_i\}$$

Noting that these quantities depend exclusively on the ranks of the observations, Fisher and Switzer propose to plot the pairs (λ_i,χ_i), where

$$\chi_i=\frac{H_i-F_iG_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}}$$

and

$$\lambda_i=4\ \text{sign}\ (\tilde{F}_i\tilde{G}_i)\ \max\ (\tilde{F}_i^2,\tilde{G}_i^2)$$

where $\tilde{F}_i=F_i-1/2$, $\tilde{G}_i=G_i-1/2$ for $i\in\{1,\ldots,n\}$. To avoid outliers, they recommend that what should be plotted are only the pairs for which

$$|\lambda_i|\leq 4\left(\frac{1}{n-1}-\frac{1}{2}\right)^2$$

Fig. 5 shows the resulting graph for the learning data set of Tables 1 and 2. The coordinates of the points and the intermediate calculations that lead to them are summarized in Table 3. Note that, in general, between two and four points may be lost due to division by zero; such is the case here for three points. Given that the original data set consisted of six observations only, this leaves only $6-3=3$ points on the graph, which is obviously not particularly revealing. However, the real-life applications considered in the "Application" section and by Fisher and Switzer (1985, 2001) provide more convincing evidence of the usefulness of this tool.

Fisher and Switzer (1985, 2001) argue that λ_i, $\chi_i\in[-1,1]$. While λ_i=measure of distance between the pair (X_i,Y_i) and the center of the scatter plot, $\sqrt{n}\chi_i$=(signed) square root of the traditional chi-square test statistic for independence in the two-way table generated by counting points in the four regions delineated by the lines $x=X_i$ and $y=Y_i$.

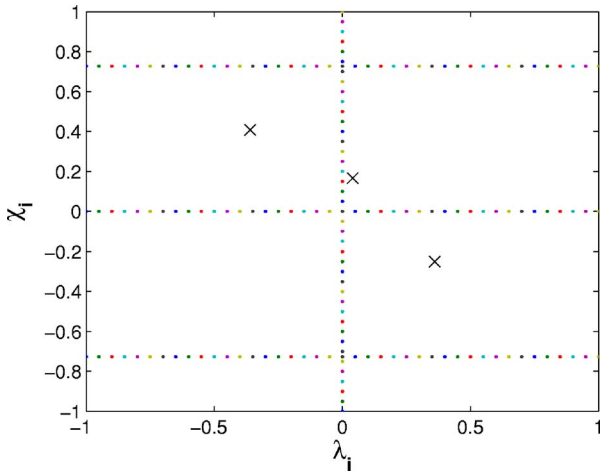


Fig. 5. Chi-plot for the learning data set

Since one would expect $H_i\approx F_i\times G_i$ for all i under independence, values of χ_i that fall too far from zero are indicative of departures from that hypothesis. To help identify such departures, Fisher and Switzer (1985, 2001) suggest that "control limits" be drawn at $\pm c_p/\sqrt{n}$, where c_p is selected so that approximately $100p\%$ of the pairs (λ_i,χ_i) lie between the lines. Through simulations, they found that the c_p values 1.54, 1.78, and 2.18 correspond to $p=0.9$, 0.95, and 0.99, respectively.

K-Plots

Another rank-based graphical tool for visualizing dependence was recently proposed by Genest and Boies (2003). It is inspired by the familiar notion of QQ-plot. Specifically, their technique consists in plotting the pairs $(W_{i:n},H_{(i)})$ for $i\in\{1,\ldots,n\}$, where

$$H_{(1)}<\cdots<H_{(n)}$$

are the order statistics associated with the quantities $H_1,\ldots,H_n$ introduced in the "Chi-Plots" subsection. As for $W_{i:n}$, it is the expected value of the ith statistic from a random sample of size n from the random variable $W=C(U,V)=H(X,Y)$ under the null hypothesis of independence between U and V (or between X and Y, which is the same). The latter is given by

$$W_{i:n}=n\binom{n-1}{i-1}\int_0^1 wk_0(w)\{K_0(w)\}^{i-1}\{1-K_0(w)\}^{n-i}dw$$

where

Table 3. Computations Required for Drawing the Chi-Plot Associated with the Learning Data Set of Table 1

i	1	2	3	4	5	6
$5H_i$	0	1	1	3	3	0
$5F_i$	0	1	2	3	4	5
$5G_i$	1	3	2	5	4	0
χ_i	—	0.408	0.167	—	−0.25	—
λ_i	1	−0.36	0.04	1	0.36	−1

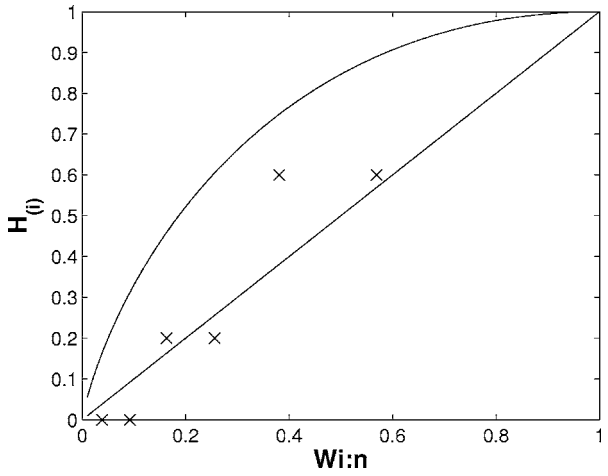


Fig. 6. K-plot for the learning data set. Superimposed on the graph are a straight line corresponding to the case of independence and a smooth curve $K_0(w)$ associated with perfect positive dependence.

$$K_0(w) = \mathrm{P}(UV \leqslant w) = \int_0^1 \mathrm{P}\left(U \leqslant \frac{w}{v}\right) dv$$

$$= \int_0^w 1\,dv + \int_w^1 \frac{w}{v} dv = w - w\log(w)$$

and k_0=corresponding density.

The values of $W_{1:6}, \ldots, W_{6:6}$ required to produce Fig. 6 can be readily computed using any symbolic calculator, such as Maple. They are given in Table 4. The interpretation of K-plots is similar to that of QQ-plots: just as curvature is problematic, e.g., in a normal QQ-plot, any deviation from the main diagonal is a sign of dependence in K-plots. Positive or negative dependence may be suspected in the data, depending whether the curve is located above or below the line $y=x$. Roughly speaking, "the further the distance, the greater the dependence." In this construction, perfect negative dependence (i.e., $C=W$) would translate into a string of data points aligned on the x-axis. As for perfect positive dependence (i.e., $C=M$), it would materialize into data aligned on the curve $K_0(w)$ shown on the graph.

As for the chi-plot, the linearity (or lack thereof) in the K-plot displayed in Fig. 6 is hard to detect, given the extremely small size of the learning data set. However, see the "Application" section and Genest and Boies (2003) for more compelling illustrartions of K-plots.

Estimation

Now suppose that a parametric family (C_θ) of copulas is being considered as a model for the dependence between two random variables X and Y. Given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from (X, Y), how should θ be estimated? This section reviews different nonparametric strategies for tackling this problem, depending on whether θ is real or multidimensional.

Only rank-based estimators are considered in the sequel. This methodological choice is justified by the fact, highlighted earlier, that the dependence structure captured by a copula has nothing to do with the individual behavior of the variables. A fortiori, any inference about the parameter indexing a family of copulas should thus rely only on the ranks of the observations, which are the best summary of the joint behavior of the random pairs.

Table 4. Coordinates of Points Displayed on the K-Plot Associated with the Learning Data Set of Table 1

i	1	2	3	4	5	6
$W_{i:6}$	0.038	0.092	0.163	0.256	0.381	0.569
$H_{(i)}$	0.0	0.2	0.2	0.6	0.6	0.0

Estimate Based on Kendall's Tau

To fix ideas, suppose that the underlying dependence structure of a random pair (X, Y) is appropriately modeled by the Farlie–Gumbel–Morgenstern family (C_θ) defined in Eq. (2). In this case, θ is real and as seen in the "Kendall's Tau" subsection there exists an immediate relation in this model between the parameter θ and the population value τ of Kendall's tau, namely

$$\tau = \frac{2}{9}\theta$$

Given a sample value τ_n of τ computed from Eq. (5) or (6), a simple and intuitive approach to estimating θ would then consist of taking

$$\tilde{\theta}_n = \frac{9}{2}\tau_n$$

Since τ_n is rank-based, this estimation strategy may be construed as a nonparametric adaptation of the celebrated method of moments.

More generally, if $\theta=g(\tau)$ for some smooth function g, then $\tilde{\theta}_n=g(\tau_n)$ may be referred to as the Kendall-based estimator of θ. A small adaptation of Proposition 3.1 of Genest and Rivest (1993) implies that

$$\sqrt{n}\frac{\tau_n - \tau}{4S} \approx \mathcal{N}(0,1)$$

where

$$S^2 = \frac{1}{n}\sum_{i=1}^n (W_i + \tilde{W}_i - 2\bar{W})^2$$

and

$$\tilde{W}_i = \frac{1}{n}\sum_{j=1}^n I_{ji} = \frac{1}{n}\#\{j{:}X_i \leqslant X_j, Y_i \leqslant Y_j\}$$

Therefore, an application of Slutsky's theorem, also known as the "Delta method," implies that as $n \to \infty$

$$\tilde{\theta}_n \approx \mathcal{N}\left[\theta, \frac{1}{n}\{4Sg'(\tau_n)\}^2\right]$$

Accordingly, an approximate $100\times(1-\alpha)\%$ confidence interval for θ is given by

$$\tilde{\theta}_n \pm z_{\alpha/2}\frac{1}{\sqrt{n}}4S|g'(\tau_n)|$$

For an alternative consistent estimator of the asymptotic variance of τ_n, see for instance, Samara and Randles (1988).

Table 5. Intermediate Values Required for the Computation of the Standard Error Associated with Kendall's Tau

i	1	2	3	4	5	6
$6W_i$	1	2	2	4	4	1
$6\widetilde{W}_i$	5	3	3	1	1	1

Example (Continued)
For the learning data set of Table 1, it was seen earlier that $\tau_n=1/15$, hence $\widetilde{\theta}_n=0.3$. Using the intermediate quantities summarized in Table 5, one finds $S^2=0.043$, hence an approximate 95% confidence interval for this estimation is $[-1,1]$, since $g'(\tau)\equiv 9/2$, and hence, $1.96\times 4S|g'(\tau_n)|/\sqrt{n}=2.99$. While the size of the standard error may appear exceedingly conservative, this result is not surprising, considering that the sample size is $n=6$.

The popularity of $\widetilde{\theta}_n$ as an estimator of the dependence parameter stems in part from the fact that closed-form expressions for the population value of Kendall's tau are available for many common parametric copula models. Such is the case, in particular, for several Archimedean families of copulas, e.g., those of Ali et al. (1978), Clayton (1978), Frank (1979), Gumbel–Hougaard (Gumbel 1960), etc. Specifically, a copula C is said to be Archimedean if there exists a convex, decreasing function $\varphi:(0,1]\rightarrow[0,\infty)$ such that $\varphi(1)=0$ and

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

is valid for all $u,v\in(0,1)$. As shown by Genest and MacKay (1986)

$$\tau=1+4\int_0^1\frac{\varphi(t)}{\varphi'(t)}dt \tag{7}$$

Table 6 gives the generator φ and an expression for τ for the three most common Archimedean models. Algebraically closed formulas are available for various other dependence models, e.g., extreme-value or Archimax copulas. See, for example, Ghoudi et al. (1998) or Capéraà et al. (2000).

Estimate Based on Spearman's Rho

When the dependence parameter θ is real, an alternative rank-based estimator that remains in the spirit of the method of moments consists of taking

$$\breve{\theta}_n=h(\rho_n)$$

where $\theta=h(\rho)$ represents the relationship between the parameter and the population value of Spearman's rho. In the context of the Farlie–Gumbel–Morgenstern family of copulas, for example, it was seen earlier that $\rho=\theta/3$, so that $\breve{\theta}_n=3\rho_n$ would be an alternative nonparametric estimator to $\widetilde{\theta}_n=9\tau_n/2$.

Now it follows from standard convergence results about empirical processes to be found, e.g., in Chapter 5 of Gaenssler and Stute (1987), that

$$\rho_n\approx\mathcal{N}\left(\rho,\frac{\sigma^2}{n}\right)$$

where the asymptotic variance σ^2 depends on the underlying copula C in a way that has been described in detail by Borkowf (2002). Arguing along the same lines as in the "Estimate Based on Kendall's Tau" subsection, it can then be seen that under suitable regularity conditions on h

$$\breve{\theta}_n\approx\mathcal{N}\left[\theta,\frac{1}{n}\{\sigma_n h'(\rho_n)\}^2\right]$$

where σ_n^2=suitable estimator of σ^2. An approximate $100\times(1-\alpha)\%$ confidence interval for θ is then given by

$$\breve{\theta}_n\pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sigma_n|h'(\rho_n)|$$

Substituting C_n for C in the expressions reported by Borkowf (2002), a very natural, consistent estimate for σ^2 is given by

$$\sigma_n^2=144(-9\mathcal{A}_n^2+\mathcal{B}_n+2\mathcal{C}_n+2\mathcal{D}_n+2\mathcal{E}_n)$$

where

$$\mathcal{A}_n=\frac{1}{n}\sum_{i=1}^n\frac{R_i}{n+1}\frac{S_i}{n+1}$$

$$\mathcal{B}_n=\frac{1}{n}\sum_{i=1}^n\left(\frac{R_i}{n+1}\right)^2\left(\frac{S_i}{n+1}\right)^2$$

$$\mathcal{C}_n=\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\frac{R_i}{n+1}\frac{S_i}{n+1}1(R_k\leq R_i,S_k\leq S_j)+\frac{1}{4}-A_n$$

$$\mathcal{D}_n=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\frac{S_i}{n+1}\frac{S_j}{n+1}\max\left(\frac{R_i}{n+1},\frac{R_j}{n+1}\right)$$

and

$$\mathcal{E}_n=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\frac{R_i}{n+1}\frac{R_j}{n+1}\max\left(\frac{S_i}{n+1},\frac{S_j}{n+1}\right)$$

Table 6. Three Common Families of Archimedean Copulas, Their Generator, Their Parameter Space, and an Expression for the Population Value of Kendall's Tau

Family	Generator	Parameter	Kendall's tau
Clayton	$(t^{-\theta}-1)/\theta$	$\theta\geqslant-1$	$\theta/(\theta+2)$
Frank	$-\log\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$	$\theta\in\mathbb{R}$	$1-4/\theta+4D_1(\theta)/\theta$
Gumbel–Hougaard	$\lvert\log(t)\rvert^\theta$	$\theta\geqslant 1$	$1-1/\theta$

Note: Here, $D_1(\theta)=\int_0^\theta(x/\theta)/(e^x-1)dx$ is the first Debye function.

Example (Continued)
For the learning data set of Table 1, it was seen earlier that $\rho_n=1/35$, hence $\breve{\theta}_n=3/35\approx 0.086$. Burdensome but simple calculations yield $\sigma_n=7.77$, hence an approximate 95% confidence interval for this estimation is $[-1,1]$, since $h'(\tau)\equiv 3$, and hence, $1.96\times\sigma_n|h'(\rho_n)|/\sqrt{n}=18.66$. Here again, the size of the standard error is quite large, as might be expected given that $n=6$.

Maximum Pseudolikelihood Estimator

In classical statistics, maximum likelihood estimation is a well-known alternative to the method of moments that is usually more

efficient, particularly when θ is multidimensional. In the present context, an adaptation of this approach to estimation is required if inference concerning dependence parameters is to be based exclusively on ranks. Such an adaptation was described in broad terms by Oakes (1994) and was later formalized and studied by Genest et al. (1995) and by Shih and Louis (1995).

The method of maximum pseudolikelihood, which requires that C_θ be absolutely continuous with density c_θ, simply involves maximizing a rank-based, log-likelihood of the form

$$\ell(\theta) = \sum_{i=1}^{n} \log\left\{ c_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right) \right\} \tag{8}$$

The latter is exactly the expression one gets when the unknown marginal distributions F and G in the classical log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log[c_\theta\{F(X_i), G(Y_i)\}]$$

are replaced by rescaled versions of their empirical counterparts, i.e.

$$F_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} 1(X_i \leq x)$$

and

$$G_n(y) = \frac{1}{n+1} \sum_{i=1}^{n} 1(Y_i \leq y)$$

That this substitution yields formula (8) is immediate, once it is realized that $F_n(X_i) = R_i/(n+1)$ and $G_n(Y_i) = S_i/(n+1)$ for all $i \in \{1, \ldots, n\}$.

This method may seem superficially less attractive than the inversion of Kendall's tau or Spearman's rho, both because it involves numerical work and requires the existence of a density c_θ. At the same time, however, it is much more generally applicable than the other methods, since it does not require the dependence parameter to be real. The procedure for estimating a multivariate θ and computing associated approximate confidence region is described by Genest et al. (1995). For simplicity, it is only presented here in the case where θ is real; however, see the "Application" section for the bivariate case.

Letting $\dot{c}_\theta(u,v) = \partial c_\theta(u,v)/\partial\theta$, Genest et al. (1995) show under mild regularity conditions that the root $\hat{\theta}_n$ of the equation

$$\dot{\ell}(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \sum_{i=1}^{n} \frac{\dot{c}_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right)}{c_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right)} = 0$$

is unique. Furthermore

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{\nu^2}{n} \right)$$

where ν^2 depends exclusively on the true underlying copula C_θ as per Proposition 2.1 of Genest et al. (1995). As mentioned by these authors, a consistent estimate of ν^2 is given by

$$\hat{\nu}_n^2 = \hat{\sigma}_n^2 / \hat{\beta}_n^2$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (M_i - \bar{M})^2$$

and

$$\hat{\beta}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (N_i - \bar{N})^2$$

are sample variances computed from two sets of pseudo-observations with means $\bar{M} = (M_1 + \cdots + M_n)/n$ and $\bar{N} = (N_1 + \cdots + N_n)/n$, respectively.

To compute the pseudo-observations M_i and N_i, one should proceed as follows:

- Step 1: Relabel the original data $(X_1, Y_1), \ldots, (X_n, Y_n)$ in such a way that $X_1 < \cdots < X_n$; as a consequence one then has $R_1 = 1, \ldots, R_n = n$.
- Step 2: Write $L(\theta, u, v) = \log c_\theta(u, v)$ and compute L_θ, L_u, and L_v, which are the derivatives of L with respect to θ, u, and v, respectively.
- Step 3: For $i \in \{1, \ldots, n\}$, set

$$N_i = L_\theta\left(\hat{\theta}_n, \frac{i}{n+1}, \frac{S_i}{n+1} \right)$$

- Step 4: For $i \in \{1, \ldots, n\}$, let also

$$M_i = N_i - \frac{1}{n} \sum_{j=i}^{n} L_\theta\left(\hat{\theta}_n, \frac{j}{n+1}, \frac{S_j}{n+1} \right) L_u\left(\hat{\theta}_n, \frac{j}{n+1}, \frac{S_j}{n+1} \right) - \frac{1}{n} \sum_{S_j \geq S_i} L_\theta\left(\hat{\theta}_n, \frac{j}{n+1}, \frac{S_j}{n+1} \right) L_v\left(\hat{\theta}_n, \frac{j}{n+1}, \frac{S_j}{n+1} \right)$$

Example (Continued)

Suppose that a Farlie–Gumbel–Morgenstern copula model is being considered for the learning data set of Table 1. In this case

$$c_\theta(u,v) = 1 + \theta(1-2u)(1-2v)$$

and

$$\frac{\dot{c}_\theta(u,v)}{c_\theta(u,v)} = \frac{(1-2u)(1-2v)}{1 + \theta(1-2u)(1-2v)}$$

Accordingly, the log-pseudolikelihood associated with this model is given by

$$\ell(\theta) = \sum_{i=1}^{n} \log\left\{ 1 + \theta\left(1 - \frac{2R_i}{n+1} \right)\left(1 - \frac{2S_i}{n+1} \right) \right\}$$

and the corresponding pseudoscore function is

$$\dot{\ell}(\theta) = \sum_{i=1}^{n} \frac{\left(1 - 2\frac{R_i}{n+1} \right)\left(1 - 2\frac{S_i}{n+1} \right)}{1 + \theta\left(1 - 2\frac{R_i}{n+1} \right)\left(1 - 2\frac{S_i}{n+1} \right)} = \sum_{i=1}^{n} \frac{(n+1-2R_i)(n+1-2S_i)}{(n+1)^2 + \theta(n+1-2R_i)(n+1-2S_i)}$$

These two functions are plotted in Fig. 7 with $n=6$ and the values of R_i and S_i given in Table 2.

Upon substitution, one gets $\hat{\theta}_n = 0.0989$ as the unique root of the equation

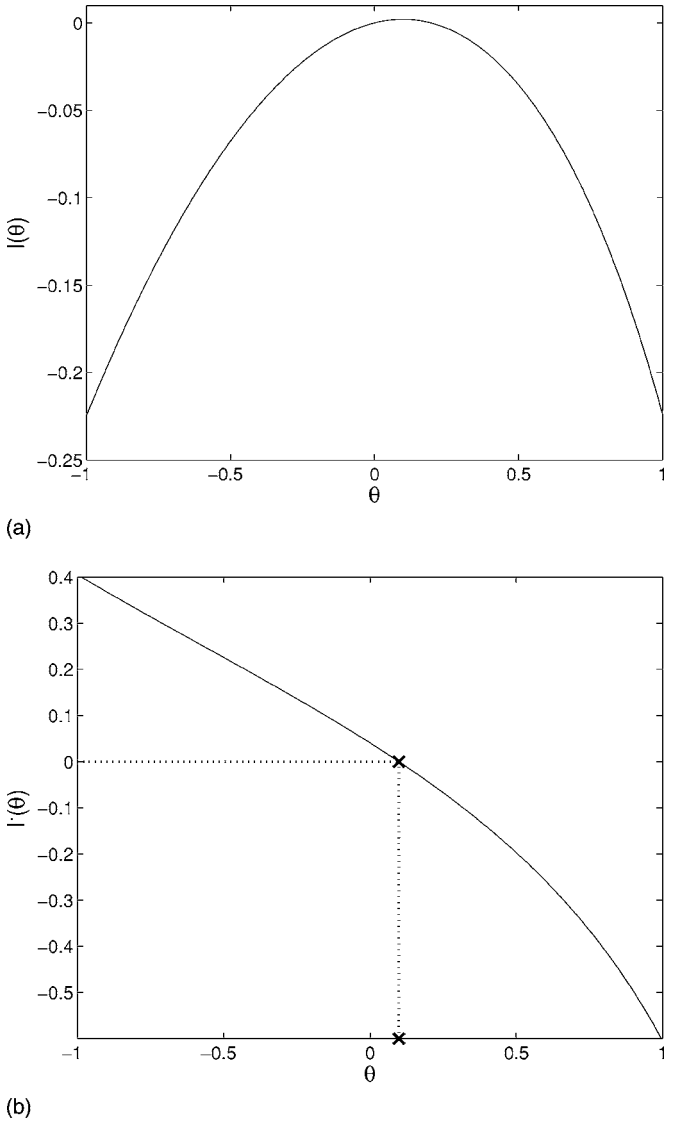


Fig. 7. Graphs of $\ell(\theta)$ (a) and $\dot{\ell}(\theta)$ (b) for the learning data set of Table 1 when the assumed model is the Farlie–Gumbel–Morgenstern family of copulas

$$\dot{\ell}(\theta) = \frac{15}{49+15\theta} - \frac{3}{49-3\theta} + \frac{1}{49+\theta} + \frac{5}{49+5\theta} + \frac{9}{49+9\theta} - \frac{25}{49-25\theta} = 0$$

In the present case

$$L_\theta(\theta,u,v) = \frac{(1-2u)(1-2v)}{1+\theta(1-2u)(1-2v)}$$

$$L_u(\theta,u,v) = \frac{-2\theta(1-2v)}{1+\theta(1-2u)(1-2v)}$$

$$L_v(\theta,u,v) = \frac{-2\theta(1-2u)}{1+\theta(1-2u)(1-2v)}$$

Using the intermediate calculations summarized in Table 7, one gets $\hat{v}_n^2 = 0.0677/0.0707 = 0.958$ and $1.96 \times \hat{v}_n/\sqrt{n} = 0.783$. The confidence interval for the maximum likelihood estimator is given by $[-0.684, 0.882]$.

Table 7. Values of the Constants N_i and M_i Required to Compute an Approximate Confidence Interval for the Maximum of Pseudolikelihood Estimator $\hat{\theta}_n$

i	1	2	3	4	5	6
N_i	0.297	−0.0616	0.0204	0.101	0.180	−0.537
M_i	0.286	−0.0832	−0.00147	0.0824	0.162	−0.534

Other Estimation Methods

Although they are the most common, estimators based on the maximization of the pseudolikelihood and on the inversion of either Kendall's tau or Spearman's rho are not the only rank-based procedures available for selecting appropriate values of dependence parameters in a copula-based model. Tsukahara (2005), for example, recently investigated the behavior and performance of two new classes of estimators derived from minimum-distance criteria and an estimating-equation approach. In his simulations, however, the maximum pseudolikelihood estimator turned out to have the smallest mean-squared error. Circumstances under which the latter approach is asymptotically semiparametrically efficient were delineated by Klaassen and Wellner (1997) and by Genest and Werker (2002). See Biau and Wegkamp (2005) for another rank-based, minimum-distance method for dependence parameter estimation.

In all fairness, it should be mentioned that the exclusive reliance on ranks for copula parameter estimation advocated here does not make complete consensus in the statistical community. In his book, Joe (1997, Chap. 10) recommends a parametric two-step procedure often referred to as the "inference from margins" or IFM method. As in the pseudolikelihood approach described above, the estimate of θ is obtained through the maximization of a function of the form

$$\ell(\theta) = \sum_{i=1}^{n} \log[c_\theta\{\hat{F}(X_i), \hat{G}(Y_i)\}]$$

However, while the rank-based method takes $\hat{F} = F_n$ and $\hat{G} = G_n$, Joe (1997) substitutes $\hat{F} = F_{\delta_n}$ and $\hat{G} = G_{\eta_n}$, where (F_δ) and (G_η) = suitable parametric families for the margins, and δ_n and η_n = standard maximum likelihood estimates of their parameters, derived from the observed values of X and Y, respectively. Cherubini et al. (2004, Section 5.3) point out that the IFM method may be viewed as a special case of the generalized method of moments with an identity weight matrix. Joe (2005) quantifies the asymptotic efficiency of the approach in different circumstances. Although they usually perform well, the estimates of the association parameters derived by the IFM technique clearly depend on the choice of F and G, and thus always run the risk of being unduly affected if the models selected for the margins turn out to be inappropriate [see e.g., Kim et al. (2007)].

For completeness, it may be worth mentioning that another developing body of literature proposes the use of kernel methods to derive a smooth estimate of a copula or its density, without assuming any specific parametric form for it. See, e.g., Gijbels and Mielniczuk (1990) or Fermanian and Scaillet (2003).

Goodness-of-Fit Tests

In typical modeling exercises, the user has a choice between several different dependence structures for the data at hand. To keep things simple, suppose that two copulas C_{θ_n} and D_{ξ_n} were fitted by some arbitrary method. It is then natural to ask which of the two models provides the best fit to the observations. Both informal and formal ways of tackling this question will be discussed in turn.

Graphical Diagnostics

When dealing with bivariate data, possibly the most natural way of checking the adequacy of a copula model would be to compare a scatter plot of the pairs $(R_i/(n+1), S_i/(n+1))$ (i.e., the support of the empirical copula C_n) with an artificial data set of the same size generated from C_{θ_n}. To avoid arbitrariness induced by sampling variability, however, a better strategy consists of generating a large sample from C_{θ_n}, which effectively amounts to portraying the associated copula density in two dimensions.

Simple simulation algorithms are available for most copula models; see, e.g., Devroye (1986, Chap. 11), or Whelan (2004) for Archimedean copulas. In the bivariate case, a good strategy for generating a pair (U, V) from a copula C proceeds as follows:

- Step 1: Generate U from a uniform distribution on the interval (0, 1).
- Step 2: Given $U=u$, generate V from the conditional distribution

$$Q_u(v) = \mathrm{P}(V \leqslant v | U = u) = \frac{\partial}{\partial u} C(u, v)$$

by setting $V=Q_u^{-1}(U^*)$, where U^*=another observation from the uniform distribution on the interval (0, 1). When an explicit formula does not exist for Q_u^{-1}, the value $v=Q_u^{-1}(u^*)$ can be determined by trial and error or more effectively using the bisection method; see Devroye (1986, Chap. 2).

Thus, for the Farlie–Gumbel–Morgenstern family of copulas, one finds

$$Q_u(v) = v + \theta v(1-v)(1-2u)$$

for all $u, v \in [0,1]$, and hence

$$Q_u^{-1}(u^*) = \begin{cases} u^* & \text{if } b = \theta(1-2u) = 0 \\ \dfrac{(b+1) - \sqrt{(b+1)^2 - 4bu^*}}{2b} & \text{if } b = \theta(1-2u) \neq 0 \end{cases}$$

Fig. 8(a) displays 100 pairs (U_i, V_i) generated with this algorithm, taking $\theta=\hat{\theta}_n=0.0989$ as deduced from the method of maximum pseudolikelihood. The six points of the learning data set, represented by crosses, are superimposed. Given the small size of the data set, it is hard to tell from this graph whether the selected model accurately reproduces the dependence structure revealed by the six observations. To show the effectiveness of the procedure, the same exercise was repeated in Fig. 8(b), using a Clayton copula with $\theta=10$. Here, the inappropriateness of the model is apparent, as might have been expected from the fact that $\tau=5/6$ for this copula, while $\tau_n=1/15$.

Another option, which is related to K-plots, consists of comparing the empirical distribution K_n of the variables $W_1, \ldots, W_n$ introduced previously with K_{θ_n}, i.e., the theoretical distribution of $W=C_{\theta_n}(U,V)$, where the pair (U,V) is drawn from C_{θ_n}. One possibility is to plot K_n and K_{θ_n} on the same graph to see how well they agree. Alternatively, a QQ-plot can be derived from the order statistics $W_{(1)} \leqslant \cdots \leqslant W_{(n)}$ by plotting the pairs $(W_{i:n}, W_{(i)})$ for $i \in \{1, \ldots, n\}$. In this case, however, $W_{i:n}$ is the expected value of the ith order statistic from a random sample of size n from K_{θ_n}, rather than from K_0, as was the case in the K-plot. In other words

$$W_{i:n} = n\binom{n-1}{i-1}\int_0^1 w k_{\theta_n}(w)\{K_{\theta_n}(w)\}^{i-1}\{1-K_{\theta_n}(w)\}^{n-i}dw \quad (9)$$

where $K_{\theta_n}(w)=\mathrm{P}\{C_{\theta_n}(U,V) \leqslant w\}$ and $k_{\theta_n}=dK_{\theta_n}(w)/dw$.

These two graphs are presented in Fig. 9 for the learning data set and Clayton's copula with parameter $\theta_n=\hat{\theta}_n=0.449$, obtained by the method of maximum pseudolikelihood. As implied by the data in Table 5, K_n is a scale function with steps of

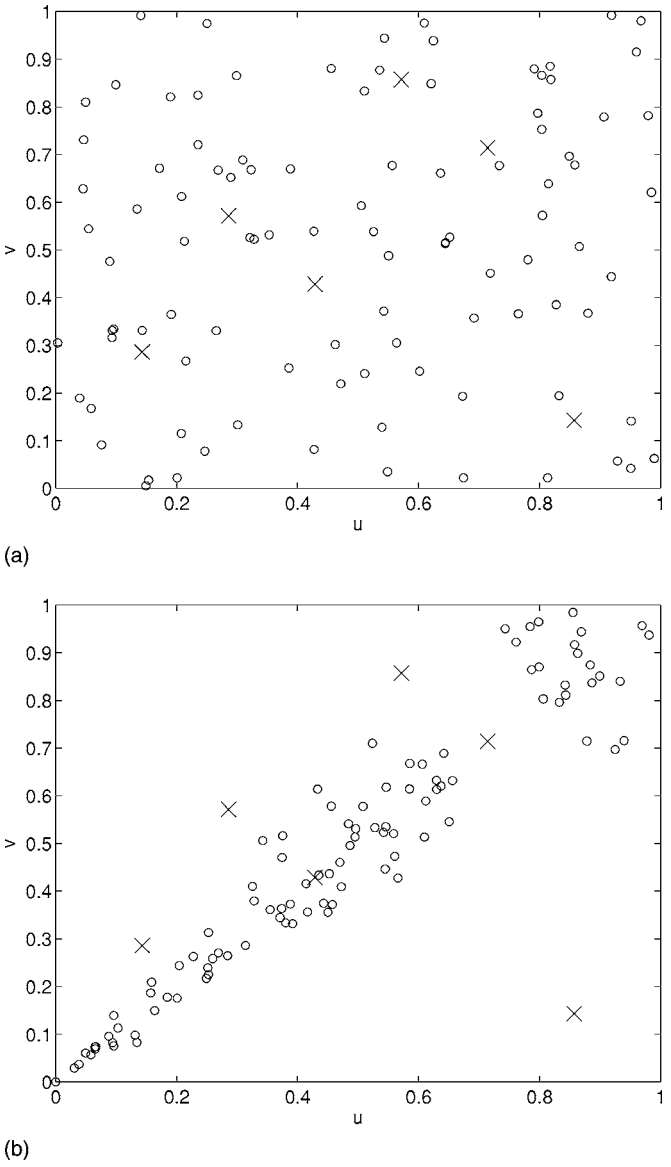


Fig. 8. (a) Scatter plot of 100 pairs (U_i, V_i) simulated from the Farlie–Gumbel–Morgenstern with parameter $\theta=0.0989$. (b) Similar plot, generated from the Clayton copula with $\tau=5/6$. On both graphs, the six points of the learning data set are indicated with a cross.

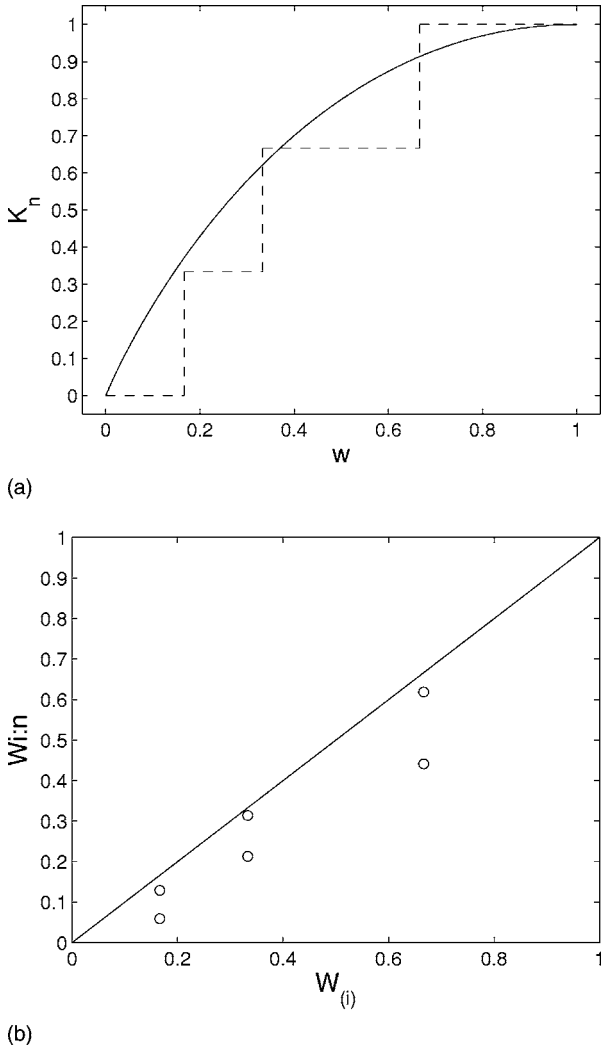


Fig. 9. (a) Graphs of K_n and K_{θ_n} for the learning data set and Clayton's copula with $\theta_n=\hat{\theta}_n=0.449$. (b) Generalized K-plot providing a visual check of the goodness-of-fit of the same model on these data.

height 1/3 at $w=1/6$, 2/6, and 4/6. This is portrayed in dotted lines in Fig. 9(a). The solid line which is superimposed is

$$K_{\theta_n}(w) = w + \frac{w}{\theta_n}(1 - w^{\theta_n}), \quad w \in (0,1) \tag{10}$$

Since $K_{\theta_n} \to K_\theta$ and $K_n \to K$ as shown by Genest and Rivest (1993), the two curves should look very similar when the data are sufficiently abundant and the model is good, i.e., when $K=K_\theta$. More generally, see Barbe et al. (1996) for a study of the large-sample behavior of the empirical process $\sqrt{n}(K_n-K)$.

In the present case, the formula for K_{θ_n} is easily deduced from the fact, established by Genest and Rivest (1993), that if C is an Archimedean copula with generator φ, the distribution function of $W=C(U,V)=H(X,Y)$, called the bivariate probability integral transform (BIPIT), is given by

$$K(w) = w - \frac{\varphi(w)}{\varphi'(w)}, \quad w \in (0,1)$$

It may be observed in passing that identity (7) is a straightforward consequence of this result and the fact that $\mathrm{E}(W)=(\tau+1)/4$. For additional information about the BIPIT and its properties and applications, refer to Genest and Rivest (2001) and Nelsen et al. (2003).

Table 8. Coordinates of the QQ-Plot Displayed in Fig. 9(b)

i	1	2	3	4	5	6
$W_{i:n}$	0.059	0.129	0.213	0.314	0.441	0.619
$W_{(i)}$	1/6	1/6	2/6	2/6	4/6	4/6

Example (Continued)

Fig. 9(b) shows a QQ-plot for visual assessment of the adequacy of the Clayton model for the learning data set. The coordinates of the points on the graph are given in Table 8. The y-coordinates were obtained by numerical integration, upon substitution of the specific choice of K_{θ_n} given in Eq. (10) into the general formula (9). By construction, this generalized K-plot is designed to yield an approximate straight line, when the model is adequate and the data sufficiently numerous to make a visual assessment. The effectiveness of the two diagnostic tools described above will be demonstrated more convincingly in the "Application" section.

Formal Tests of Goodness-of-Fit

Formal methodology for testing the goodness-of-fit of copula models is just emerging. To the writers' knowledge, the first serious effort to develop such a procedure was made by Wang and Wells (2000) in the context of Archimedean models. Inspired by Genest and Rivest (1993), these authors proposed to compute a Cramér–von Mises statistic of the form

$$S_{n\xi} = n \int_\xi^1 \{K_n(w) - K_{\theta_n}(w)\}^2 dw$$

where $\xi \in (0,1)$ is an arbitrary cutoff point. While Theorem 3 in their paper identifies the limiting distribution of $S_{n\xi}$, the latter is analytically unwieldy. Furthermore, the bootstrap procedure they propose in replacement is, of their own admission, ineffective. As a result, P-values for the statistic cannot be computed. When faced with a choice between several copulas, therefore, Wang and Wells (2000) thus end up recommending that the model yielding the smallest value of $S_{n\xi}$ be selected.

Recently, Genest et al. (2006) introduced two variants of the $S_{n\xi}$ statistic and of the bootstrap procedure of Wang and Wells (2000) that allow overcoming these limitations. In addition to being much simpler to compute than $S_{n\xi}$ and independent of the choice of ξ, the statistics proposed by Genest et al. (2006) can be used to test the adequacy of any copula model, whether Archimedean or not. More importantly still, P-values associated with these statistics are relatively easy to obtain by bootstrapping.

Specifically, the statistics considered by Genest et al. (2006) are of the form

$$\mathcal{S}_n = \int_0^1 |\mathbb{K}_n(w)|^2 k_{\theta_n}(w) dw$$

and

Table 9. *P*-Values Estimated by Parametric Bootstrap for Testing the Goodness-of-Fit of the Clayton Copula Model on the Learning Data Set Using the Cramér–von Mises and the Kolmogorov–Smirnov Statistics $\mathcal{S}_n$ and $\mathcal{T}_n$

	P-value based on a run of			
Statistic	$N=100{,}000$	$N=10{,}000$	$N=100$	$N=100$
$\mathcal{S}_n$	0.266	0.262	0.45	0.39
$\mathcal{T}_n$	0.494	0.489	0.58	0.49

$$\mathcal{T}_n = \sup_{0\leq w\leq 1} |\mathbb{K}_n(w)|$$

where $\mathbb{K}_n(w)=\sqrt{n}\{K_n(w)-K_{\theta_n}(w)\}$. Although prima facie these expressions seem just as complicated as $S_{n\xi}$, Genest et al. (2006) show that in fact, they can be easily computed as follows:

$$\mathcal{S}_n = \frac{n}{3} + n\sum_{j=1}^{n-1} K_n^2\left(\frac{j}{n}\right)\left\{K_{\theta_n}\left(\frac{j+1}{n}\right) - K_{\theta_n}\left(\frac{j}{n}\right)\right\} - n\sum_{j=1}^{n-1} K_n\left(\frac{j}{n}\right)\left\{K_{\theta_n}^2\left(\frac{j+1}{n}\right) - K_{\theta_n}^2\left(\frac{j}{n}\right)\right\}$$

and

$$\mathcal{T}_n = \sqrt{n}\max_{i=0,1;0\leq j\leq n-1}\left\{\left|K_n\left(\frac{j}{n}\right) - K_{\theta_n}\left(\frac{j+i}{n}\right)\right|\right\}$$

The bootstrap methodology required to compute associated *P*-values proceeds as follows, say in the case of $\mathcal{S}_n$:

- Step 1: Estimate θ by a consistent estimator θ_n.
- Step 2: Generate N random samples of size n from C_{θ_n} and, for each of these samples, estimate θ by the same method as before and determine the value of the test statistic.
- Step 3: If $\mathcal{S}_{1:N}^* \leq \cdots \leq \mathcal{S}_{N:N}^*$ denote the ordered values of the test statistics calculated in Step 2, an estimate of the critical value of the test at level α based on $\mathcal{S}_n$ is given by

$$\mathcal{S}_{\lfloor(1-\alpha)N\rfloor:N}^*$$

and

$$\frac{1}{N}\#\{j:\mathcal{S}_j^* \geq \mathcal{S}_n\}$$

yields an estimate of the *P*-value associated with the observed value $\mathcal{S}_n$ of the statistic. Here, $\lfloor x\rfloor$ simply refers to the integer part of $x\in\mathbb{R}$.

Obviously, the larger N, the better. In practice, $N=10{,}000$ seems perfectly adequate, although one could certainly get by with less, if limited in time or computing power. An additional complication occurs when K_θ cannot be written in algebraic form. In that case, a double bootstrap procedure must be called upon, for which the reader is referred to Genest and Rémillard (2005).

Example (Continued)

Suppose that Clayton's copula model has been fitted to the learning data set using some consistent estimator θ_n. To test the adequacy of this dependence structure, one could then compute the "distance" between K_n and

$$K_{\theta_n}(w) = w + \frac{w}{\theta_n}(1-w^{\theta_n})$$

using either $\mathcal{S}_n$ or $\mathcal{T}_n$. The corresponding *P*-values could then be found via the parametric bootstrap procedure described above. In order to get valid results, however, note that the same estimation method must be used at every iteration of this numerical algorithm. To reduce the intensity of the computing effort, the estimator $\tilde{\theta}_n$ obtained through the inversion of Kendall's tau is often the most convenient choice, particularly for Archimedean models.

When the dependence parameter of Clayton's model is estimated in this fashion, one gets $\theta_n=\tilde{\theta}_n=0.143$. The observed values of these statistics are then easily found to be

$$\mathcal{S}_n = 0.272, \quad \mathcal{T}_n = 1.053$$

Table 9 reports the simulated *P*-values obtained via parametric bootstrapping for one run of $N=100{,}000$, one run of $N=10{,}000$, and two runs of $N=100$. The discrepancy observed between *P*-values derived from the two runs at $N=100$ illustrates the importance of taking N large enough to insure reliable conclusions. As can be seen from Table 9, taking $N=100{,}000$ instead of $N=10{,}000$ did not change the estimated *P*-values much, which is reassuring. Notwithstanding these differences, neither of the two tests leads to the rejection of Clayton's model. Given the sample size, this is of course unsurprising.

One drawback of this general strategy to goodness-of-fit testing is that as the number of variables increases, the univariate summary represented by the probability integral transformation $W=C(U_1,\ldots,U_d)=H(X_1,\ldots,X_d)$ and its distribution function $K(w)$ is less and less representative of the multivariate dependence structure embodied in C.

For bivariate or trivariate applications such as are common in hydrology, there is, however, another more serious difficulty associated with a test based on $S_{n\xi}$, $\mathcal{S}_n$, or $\mathcal{T}_n$. This arises from the fact that a given theoretical distribution K can sometimes correspond to two different copulas. In other words, it may happen that K is not only the distribution function of $W=C(U,V)$ but also that of $W^\dagger=C^\dagger(U^\dagger,V^\dagger)$, where $(U^\dagger,V^\dagger)$ is distributed as $C^\dagger$. In fact, Nelsen et al. (2003) show that unless C belongs to the Bertino family of copulas (Bertino 1977; Fredricks and Nelsen 2002), there always exists $C^\dagger$ in that class such that $K=K^\dagger$ and $C\neq C^\dagger$.

To illustrate the difficulties associated with the lack of uniqueness of K, consider the class of bivariate extreme-value copulas, which are of the form

$$C(u,v) = \exp\left[\log(uv)A\left\{\frac{\log(u)}{\log(uv)}\right\}\right] \tag{11}$$

where $A:[0,1]\rightarrow[1/2,1]=$some convex mapping such that $A(t)\geq\max(t,1-t)$ for all $t\in[0,1]$. See, e.g., Geoffroy (1958), Sibuya (1960), or Ghoudi et al. (1998). The population value of Spearman's rho for this class of copulas can be written as

$$\rho_A = 12\int_0^1 \{A(w)+1\}^{-2}dw - 3$$

Also, as shown by Ghoudi et al. (1998), the distribution function of $W=C(U,V)$ for C in this class is given by

$$K_A(w) = w - (1-\tau_A)w\log(w)$$

where

$$\tau_A = \int_0^1 \frac{w(1-w)}{A(w)} A''(w)dw$$

whenever the second derivative of A is continuous. In particular, note that K_A does not depend on the whole function A, but only on the population value of Kendall's tau induced by A. For this reason, formal and informal goodness-of-fit procedures based on a comparison of K_n and K_{θ_n} could not possibly distinguish, e.g., between two extreme-value copulas whose real-valued parameters would be estimated through inversion of Kendall's tau. In statistical parlance, the above-mentioned tests are not consistent.

As already mentioned by Fermanian (2005) and by Genest et al. (2006), an obvious way to circumvent the consistency issue would be to base a goodness-of-fit test directly on the distance between C_n and C_{θ_n}. Since the limiting distribution of the process $\sqrt{n}(C_n - C_{\theta_n})$ is very complex, however, this strategy could only be implemented through an intensive use of the parametric bootstrap. For additional information in this regard, refer to the "Application" section and to Genest and Rémillard (2005).

The only other general solution available to date involves kernel estimation of the copula density, as developed in Fermanian (2005). An advantage of his statistic is that it has a standard chi-square distribution in the limit. The implementation of the procedure, however, involves arbitrary choices of a kernel, its window, and a weight function. As a result, some objectivity is lost.

Finally, since extreme-value copula models are likely to be useful in frequency analysis; diagnostic and selection tools specifically suited to that case will be discussed in the context of the hydrological application to be considered next.

Application

The Harricana watershed is located in the northwest region of the province of Québec. The Harricana River originates from several lakes near Val d'Or and empties into James Bay about 553 km north. The name of the river takes its origin from the Algonquin word "Nanikana" meaning "the main way." The daily discharges of the Harricana River at Amos (measured at Environment Canada Station Number 04NA001) have been used several times in the hydrology literature since the data are available from 1914 to present; see, e.g., Bobée and Ashkar (1991) and Bâ et al. (2001). The main characteristics of the watershed are the following: drainage area of 3,680 km^2 at the gauging station, mean altitude 380 m, 23% of lakes and swamp, and 72% of forest. Spring represents the high flow season due to the contribution of seasonal snowmelt to river runoff. Generally, a combination of snowmelt and rainfall events generates the annual floods.

Data

The data considered for the application consist of the maximum annual flow X (in m^3/s) and the corresponding volume Y (in hm^3) for $n=85$ consecutive years, starting in 1915 and ending in 1999. Using standard univariate modeling techniques, the present writers came to the conclusion that the annual flow X could be appropriately modeled by a Gumbel extreme-value distribution $\hat{F}$ with mean 189 m^3/s and standard error 51.5 m^3/s. As for volume Y, it is faithfully described by a gamma distribution $\hat{G}$ with mean 1,034.88 hm^3 and standard error 234.93 hm^3. Fig. 10(a and b) show QQ-plots attesting to the good fit of these marginal distributions to the observed values of X and Y, respectively. Since the focus of the present study is on modeling the dependence between the two variables, nothing further will be said about the fit of their marginal distributions. As previously emphasized, the choice of margins is immaterial anyway, at least insofar as inference on the dependence structure of the data is based on ranks.

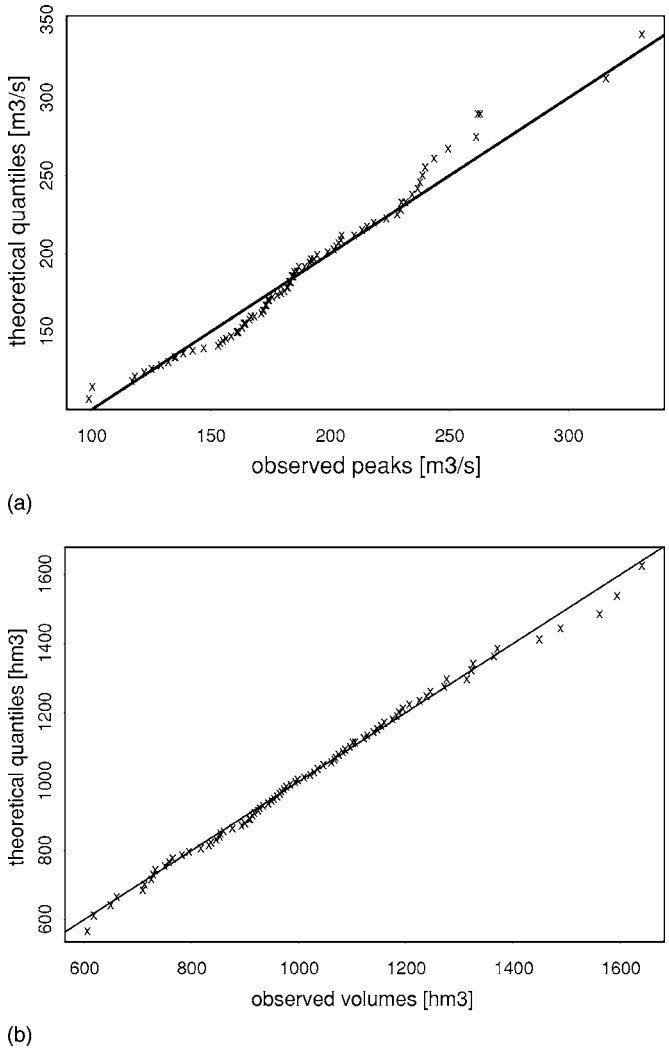


Fig. 10. QQ-plots showing the fit of marginal models for peak (a) and volume (b) for the Harricana River data

Assessment of Dependence

Before a copula model for the pair (X, Y) is sought, visual tools were used to check for the presence of dependence. The scatter plot of (normalized) ranks shown in Fig. 11 suggests the presence of positive association between peak flow and volume, as might be expected. This is confirmed by the chi-plot and the K-plot, reproduced in Fig. 12(a) and 12(b), respectively. As can be seen, most of the points fall outside the "confidence band" of the chi-plot. An obvious curvature is also apparent in the K-plot. Both graphs point to the existence of a positive relationship between the two variables.

To quantify the degree of dependence in the pair (X, Y), sample values of Spearman's rho and Kendall's tau were

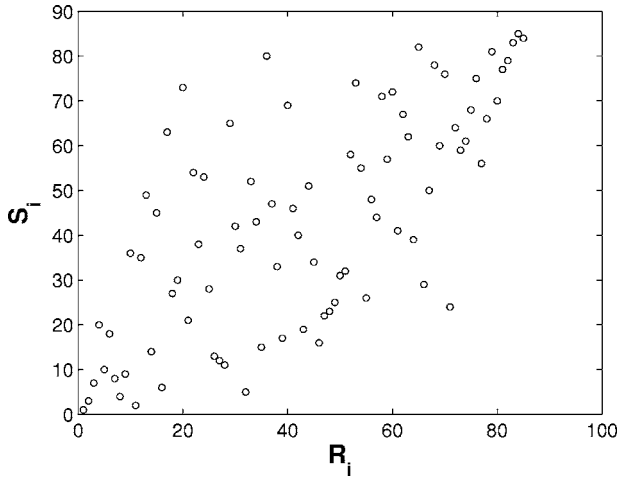


Fig. 11. Scatter plot of the ranks for the Harricana River data

computed, along with the P-values of the associated tests of independence. For ρ, one finds

$$\rho_n = 0.696 \quad \text{and} \quad \sqrt{n-1}|\rho_n| = 6.38$$

so that the P-value of the test is $2\Pr(\mathcal{Z} > \sqrt{84} \times 0.696) = 2\Pr(\mathcal{Z} > 6.38) \approx 0\%$, where $\mathcal{Z}$ continues to denote a normal random variate with zero mean and unit variance. For τ, one gets

$$\tau_n = 0.522 \quad \text{and} \quad \sqrt{\frac{9n(n-1)}{2(2n+5)}}|\tau_n| = 7.073$$

so that the P-value of this test is even smaller: $2\Pr(\mathcal{Z} > 7.073) \approx 0\%$.

Choice of Models

In order to model the dependence between the annual peak (X) and the volume (Y) of the Harricana River, some 20 families of copulas were considered, which could be classified into four broad categories:

1. One-, two-, and three-parameter Archimedean models, including the traditional Ali–Mikhail–Haq, Clayton, Frank (Nelsen 1986; Genest 1987), and Gumbel–Hougaard families of copulas and their extension described by Genest et al. (1998), but also the system of Kimeldorf and Sampson (1975), the class of Joe (1993), and the BB1–BB3 and BB6–BB7 classes described in the book of Joe (1997, pp. 150–153);
2. Extreme-value copulas, including (besides the Gumbel–Hougaard system mentioned just above) Joe's BB5 family and the classes of copulas introduced by Galambos (1975), Hüsler and Reiss (1989), and Tawn (1988);
3. Meta-elliptical copulas described, e.g., in Fang et al. (2002) or Abdous et al. (2005), most notably the normal, the Student, and the Cauchy copulas; and
4. Other miscellaneous families of copulas, such as those of Farlie–Gumbel–Morgenstern and Plackett (1965).

Some of these families of copulas could be eliminated off hand, given that the degrees of dependence they span were insufficient to account for the association observed in the data set. This was the case, e.g., for the Ali–Mikhail–Haq and Farlie–Gumbel–Morgenstern systems. To help sieve through the remaining models, the use was made of tools described in the "Graphical Diagnostics" subsection. Given a family (C_θ) of copulas, an estimate θ_n of its parameter was first obtained by the method of maximum pseudolikelihood, and then 10,000 pairs of points were generated from C_{θ_n}. Fig. 13 shows the five best contenders along with the traditional bivariate normal model.

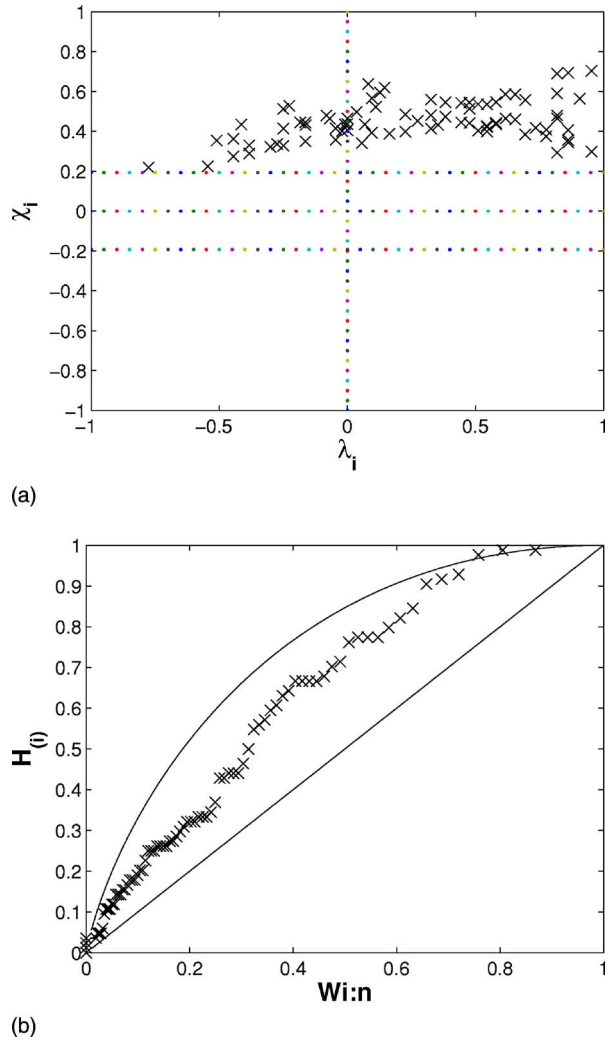


Fig. 12. Chi-plot (a) and K-plot (b) for the annual peaks and corresponding volumes of the Harricana River

As a further graphical check, the margins of the 10,000 random pairs (U_i, V_i) from each of the six estimated copula models C_{θ_n} were transformed back into the original units using the marginal distributions $\hat{F}$ and $\hat{G}$ identified in the "Data" subsection for volume and peak. The resulting scatter plots of pairs $(X_i, Y_i) = (\hat{F}^{-1}(U_i), \hat{G}^{-1}(V_i))$ are displayed in Fig. 14, along with the actual observations.

While Fig. 13 provides a graphical test of the goodness-of-fit of the dependence structure taken in isolation, Fig. 14 makes it possible to judge globally the viability of the complete model for frequency analysis. Keeping in mind the predictive ability that the final model must have, it was decided to discard the bivariate normal copula structure, due to the obvious lack of fit of the resulting model in the upper part of the distribution.

Hence, of the five dependence structures retained for the finer analysis, four were extreme-value copulas, i.e., the BB5, Galambos, Gumbel–Hougaard, and Hüsler–Reiss families. The fifth was the two-parameter BB1 Archimedean model. As can be seen from Table 10, the Gumbel–Hougaard family obtains as a

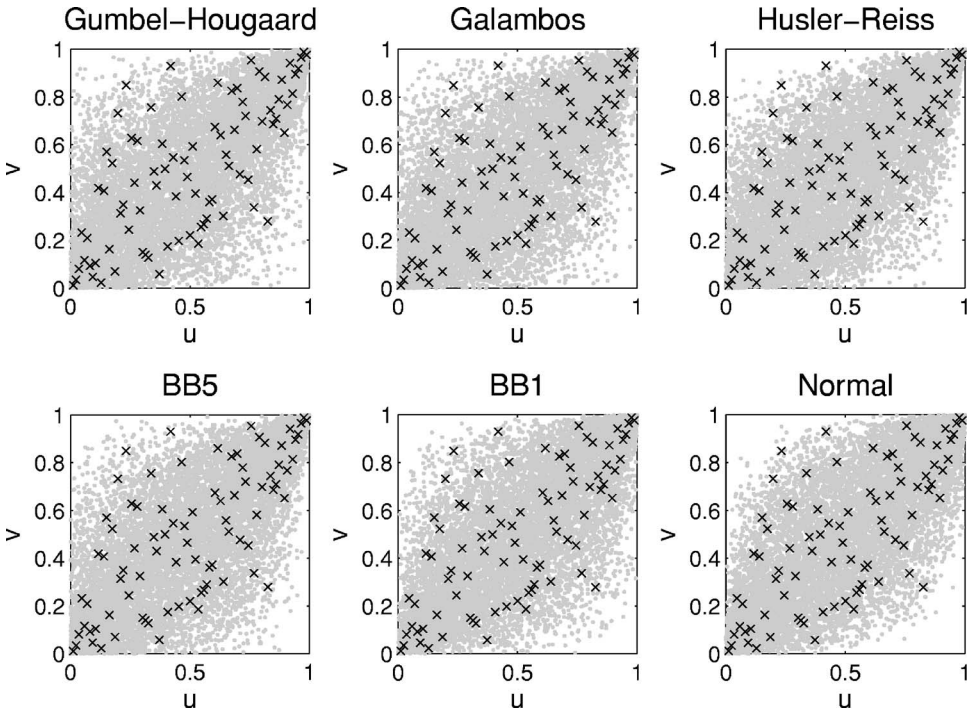


Fig. 13. Simulated random sample of size 10,000 from six chosen families (C_θ) of copulas with parameter θ estimated by the method of maximum pseudolikelihood using the peak–volume Harricana River data, whose pairs of ranks are indicated by an "X"

special case of the BB1 system when $\theta_1 \to 0$, while setting $\theta_2 = 1$ actually yields the Kimeldorf–Sampson family. Likewise, the Galambos and Gumbel–Hougaard distributions are special cases of Family BB5 corresponding, respectively, to $\theta_1 = 1$ and $\theta_2 \to 0$.

Estimation

Table 11 gives parameter values for each of the five models, based on maximum pseudolikelihood. For one-parameter models, 95% confidence intervals were computed as explained above. For

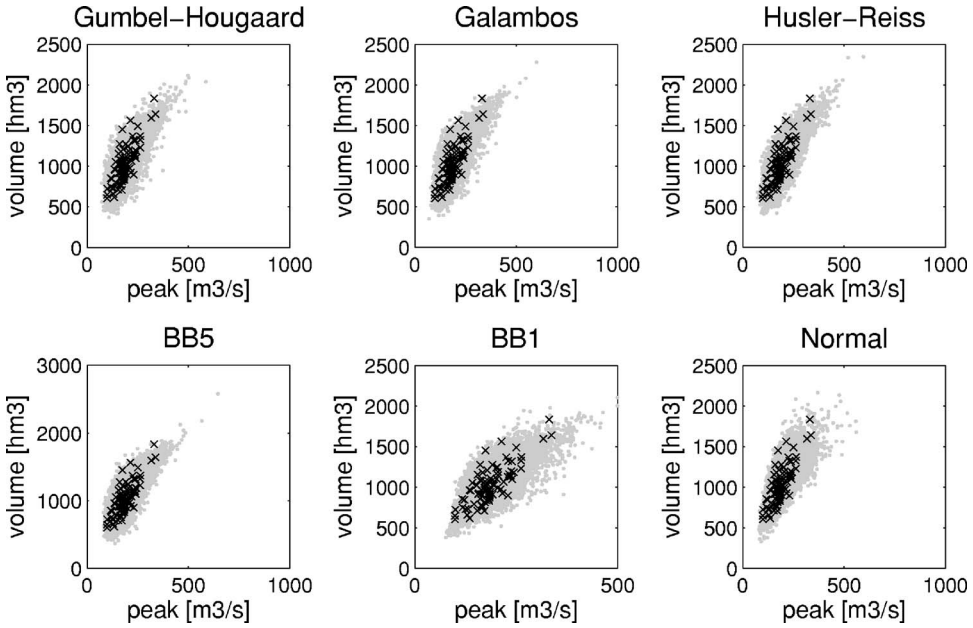


Fig. 14. Same data as in Fig. 13, upon transformation of the marginal distributions as per the selected models for the peak and the volume of the Harricana River data, whose pairs of observations are indicated by an "X"

Table 10. Definition of the Five Chosen Families of Copulas with Their Parameter Space

Copula	$C_\theta(u,v)$	Parameter(s)
Gumbel–Hougaard	$\exp\{-(\tilde{u}^\theta+\tilde{v}^\theta)^{1/\theta}\}$	$\theta \geqslant 1$
Galambos	$uv\exp\{(\tilde{u}^{-\theta}+\tilde{v}^{-\theta})^{-1/\theta}\}$	$\theta \geqslant 0$
Hüsler–Reiss	$\exp\left[-\tilde{u}\Phi\left\{\frac{1}{\theta}+\frac{\theta}{2}\log\left(\frac{\tilde{u}}{\tilde{v}}\right)\right\}-\tilde{v}\Phi\left\{\frac{1}{\theta}+\frac{\theta}{2}\log\left(\frac{\tilde{v}}{\tilde{u}}\right)\right\}\right]$	$\theta \geqslant 0$
BB1	$[1+\{(u^{-\theta_1}-1)^{\theta_2}+(v^{-\theta_1}-1)^{\theta_2}\}^{1/\theta_2}]^{-1/\theta_1}$	$\theta_1>0,\ \theta_2\geqslant 1$
BB5	$\exp[-\{\tilde{u}^{\theta_1}+\tilde{v}^{\theta_1}-(\tilde{u}^{-\theta_1\theta_2}+\tilde{v}^{-\theta_1\theta_2})^{-1/\theta_2}\}^{1/\theta_1}]$	$\theta_1\geqslant 1,\ \theta_2>0$

Note: With $\tilde{u}=-\log(u)$, $\tilde{v}=-\log(v)$ and Φ standing for the cumulative distribution function of the standard normal.

q-parameter models with $q\geqslant 2$, the determination of the confidence regions relies on an estimation of the limiting variance–covariance matrix $B^{-1}\Sigma B^{-1}$ of the estimator $\hat{\theta}_n=(\hat{\theta}_{1n},\ldots,\hat{\theta}_{qn})$ of $\theta=(\theta_1,\ldots,\theta_q)$.

Following Genest et al. (1995), the estimate of B is simply the empirical $q\times q$ variance–covariance matrix of the variables $N_1,\ldots,N_q$, for which a set of n pseudo-observations is available, namely

$$N_{pi}=L_{\theta_p}\left(\hat{\theta}_n,\frac{i}{n+1}\frac{S_i}{n+1}\right),\quad i\in\{1,\ldots,n\}$$

where L_{θ_p} denotes the derivative of $L(\theta,u,v)=\log\{c_\theta(u,v)\}$ with respect to θ_p. Here, it is assumed that the original data have been relabeled so that $X_1<\cdots<X_n$. Likewise, $\Sigma=q\times q$ variance–covariance matrix of the variables $M_1,\ldots,M_q$, for which the pseudo-observations are

$$M_{pi}=N_{pi}-\frac{1}{n}\sum_{j=i}^{n}L_{\theta_p}\left(\hat{\theta}_n,\frac{j}{n+1},\frac{S_j}{n+1}\right)L_u\left(\hat{\theta}_n,\frac{j}{n+1},\frac{S_j}{n+1}\right)$$
$$-\frac{1}{n}\sum_{S_j\geqslant S_i}L_{\theta_p}\left(\hat{\theta}_n,\frac{j}{n+1},\frac{S_j}{n+1}\right)L_v\left(\hat{\theta}_n,\frac{j}{n+1},\frac{S_j}{n+1}\right)$$

for $i\in\{1,\ldots,n\}$.

An alternative, possibly more efficient way of estimating the information matrix B is given by the Hessian matrix associated with $L(\theta,u,v)$ at $\hat{\theta}_n$, namely, the $q\times q$ matrix whose (p,r) entry is given by

$$-\frac{1}{n}\sum_{i=1}^{n}L_{\theta_p\theta_r}\left(\hat{\theta}_n,\frac{i}{n+1},\frac{S_i}{n+1}\right)$$

where $L_{\theta_p\theta_r}$ stands for the cross derivative of $L(\theta,u,v)$ with respect to both θ_p and θ_r. In Table 11, the confidence intervals for Models BB1 and BB5 were derived using the latter approach, as it produced somewhat narrower intervals.

Goodness-of-Fit Testing

As a second step towards model selection, one should look at the generalized K-plot corresponding to the five families under consideration. The graphs corresponding to the BB1 are displayed in Fig. 15. For reasons given in the "Graphical Diagnostics" subsection, the graphs corresponding to the BB5, Gumbel–Hougaard, Galambos, and Hüsler–Reiss copulas are identical, since they are all extreme-value dependence structures. The graphs appear in Fig. 16.

The plots displayed in Figs. 15 and 16 suggest that both the BB1 and extreme-value copula structures are adequate for the data at hand. A similar conclusion is drawn from the formal goodness-of-fit tests based on $\mathcal{S}_n$ and $\mathcal{T}_n$, as indicated in Table 12. Again, the generalized K-plot and the formal goodness-of-fit tests corresponding to the Galambos, Hüsler–Reiss, and BB5 extreme-value copula models yield *exactly the same* results, as evidenced in Table 12.

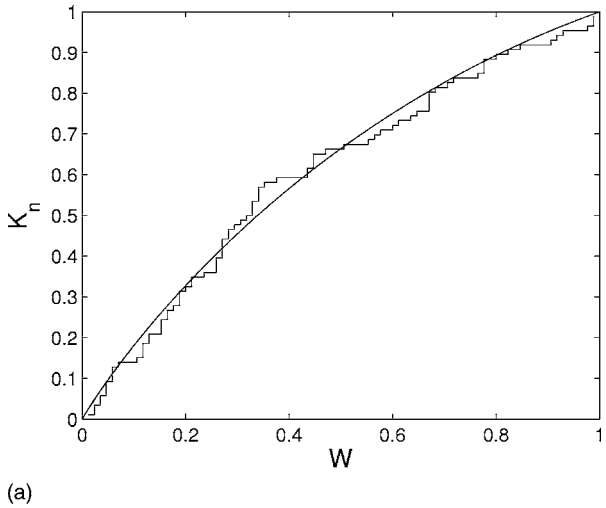
In an attempt to distinguish between the extreme-value copula structures, a consistent goodness-of-fit test could be constructed from the process $\sqrt{n}(C_n-C_{\theta_n})$, as evoked but dismissed by Fermanian (2005), on account of the unwieldy nature of its limit. However, this difficulty can be overcome easily with the use of a parametric (or double parametric) bootstrap, whose validity in this context has recently been established by Genest and Rémillard (2005). The bootstrap procedure is exactly the same as described in the "Formal Tests of Goodness-of-Fit" section, but with S_n replaced by the Cramér–von Mises statistic

$$\mathcal{CM}_n=n\sum_{i=1}^{n}\left\{C_n\left(\frac{R_i}{n+1},\frac{S_i}{n+1}\right)-C_{\theta_n}\left(\frac{R_i}{n+1},\frac{S_i}{n+1}\right)\right\}^2$$
$$=n\sum_{i=1}^{n}\left\{W_i-C_{\theta_n}\left(\frac{R_i}{n+1},\frac{S_i}{n+1}\right)\right\}^2$$

This bootstrap-based goodness-of-fit test was applied for each

Table 11. Maximum Pseudolikelihood Parameter Estimates and Corresponding 95% Confidence Interval for Five Families of Copulas, Based on the Harricana River Data

Copula	Estimate(s)	95% confidence interval (CI)
Gumbel–Hougaard	$\hat{\theta}_n=2.161$	CI=[1.867, 2.455]
Galambos	$\hat{\theta}_n=1.464$	CI=[1.162, 1.766]
Hüsler–Reiss	$\hat{\theta}_n=2.027$	CI=[1.778, 2.275]
BB1	$\hat{\theta}_{1n}=0.418,\ \hat{\theta}_{2n}=1.835$	CI=[0.022, 0.815]×[1.419, 2.251]
BB5	$\hat{\theta}_{1n}=1.034,\ \hat{\theta}_{2n}=1.244$	CI=[1.000, 1.498]×[0.774, 1.294]



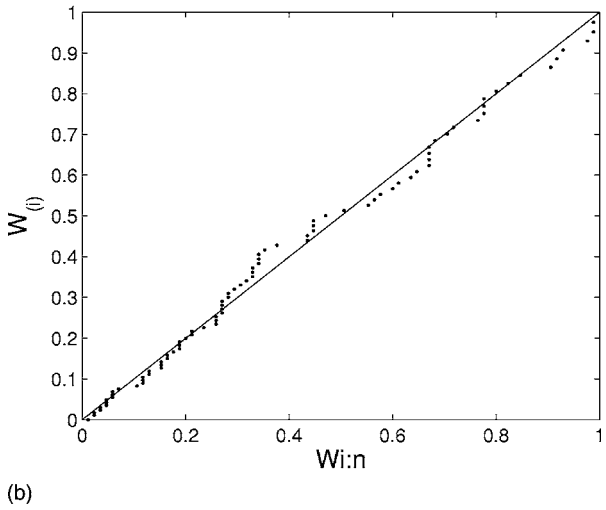
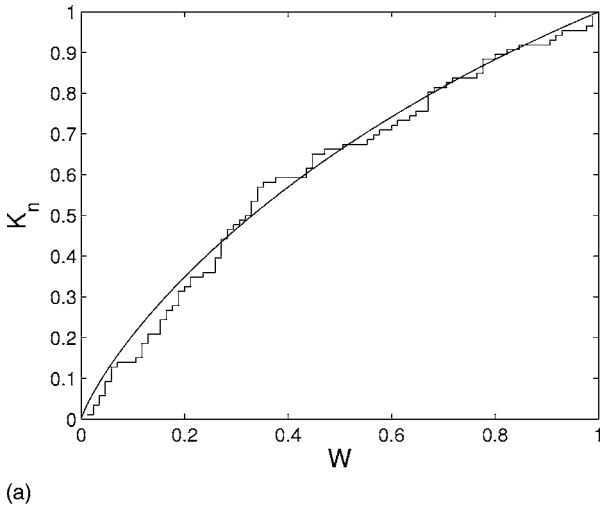


Fig. 15. (a) Graphs of K_n and K_{θ_n} for the BB1 copula with $\hat{\theta}_n=0.418$, $\hat{\delta}_n=1.835$, based on the Harricana River data. (b) Generalized K-plot providing a visual check of the goodness-of-fit of the same model for these data.



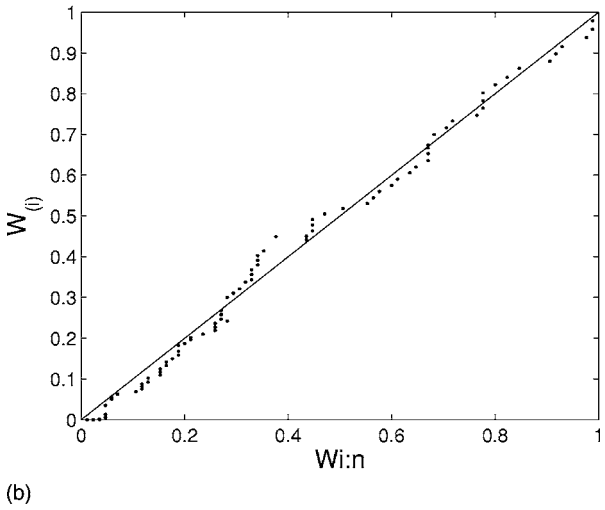


Fig. 16. (a) Graphs of K_n and K_{θ_n} for the Gumbel–Hougaard copula with $\theta_n=\hat{\theta}_n=2.161$, based on the Harricana River data. (b) Generalized K-plot providing a visual check of the goodness-of-fit of the same model for these data.

of the five families of copulas still under consideration. The results are summarized in Table 13. As it turned out, none of the models could be rejected on this basis either.

The Gumbel–Hougaard (and Galambos) copula(s) being embedded in two-parameter model(s) BB1 (and BB5), yet another option for choosing between them would be to call on a pseudo-likelihood ratio test procedure recently introduced by Chen and Fan (2005). Their approach, inspired by a semiparametric adaptation of the Akaike Information Criterion, makes it possible to measure the trade-off between goodness-of-fit and model parsimony.

Suppose it is desired to compare two nested copula models, say $\mathcal{C}=(C_{\theta,\lambda})$ and $\mathcal{D}=(C_{\theta,\lambda_0})$. Let $(\hat{\theta}_n,\hat{\lambda}_n)$ represent the maximum pseudolikelihood estimator of $(\theta,\lambda)\in\mathbb{R}^2$ under model $\mathcal{C}$, and write $\bar{\theta}_n$ for the maximum pseudolikelihood estimator of $\theta\in\mathbb{R}$ under the submodel $\mathcal{D}$. The test statistic proposed by Chen and Fan (2005) then rejects the null hypothesis $H_0:\lambda=\lambda_0$ that model $\mathcal{D}$ is preferable to model $\mathcal{C}$ whenever

$$\mathcal{CF}_n=2\sum_{i=1}^{n}\log\left\{\frac{c_{(\bar{\theta}_n,\lambda_0)}\left(\frac{R_i}{n+1},\frac{S_i}{n+1}\right)}{c_{(\hat{\theta}_n,\hat{\lambda}_n)}\left(\frac{R_i}{n+1},\frac{S_i}{n+1}\right)}\right\}$$

is sufficiently small. To determine a P-value for this test, one must resort to a nonparametric bootstrap procedure, which proceeds as follows. For some large integer N and each $k\in\{1,\ldots,N\}$, do the following:

- Step 1: Draw a bootstrap random sample $(X_1^\dagger,Y_1^\dagger),\ldots,(X_n^\dagger,Y_n^\dagger)$ with replacement from $(X_1,Y_1),\ldots,(X_n,Y_n)$.
- Step 2: Use the method of maximum pseudolikelihood to determine estimators $(\hat{\theta}_n^\dagger,\hat{\lambda}_n^\dagger)$ and $\bar{\theta}_n^\dagger$ of (θ,λ) and (θ,λ_0) under models $\mathcal{C}$ and $\mathcal{D}$, respectively.
- Step 3: Compute the Hessian matrices B_1 and B_2 associated with $\log\{c_{\theta,\lambda}(u,v)\}$ and $\log\{c_{\theta,\lambda_0}(u,v)\}$ at $(\hat{\theta}_n^\dagger,\hat{\lambda}_n^\dagger)$ and $\bar{\theta}_n^\dagger$, respectively.
- Step 4: Determine the value of

Table 12. Results of the Bootstrap Based on the Cramér–von Mises Statistic $\mathcal{S}_n$ and Kolmogorov–Smirnov Statistic $\mathcal{T}_n$: Observed Statistic, Critical Value $q(\alpha)$ Corresponding to $\alpha=5\%$ and Approximate P-Value, Based on $N=10{,}000$ Parametric Bootstrap Samples

Copula	$\mathcal{S}_n$	$q(95\%)$	P-value	$\mathcal{T}_n$	$q(95\%)$	P-value
Gumbel–Hougaard	1.563	2.285	0.119	1.842	2.081	0.091
Galambos	1.563	2.285	0.119	1.842	2.081	0.091
Hüsler–Reiss	1.563	2.285	0.119	1.842	2.081	0.091
BB1	1.499	1.987	0.135	1.849	2.059	0.123
BB5	1.563	2.285	0.119	1.842	2.081	0.091

$$\mathcal{CF}_{n,k} = (\bar{\theta}_n^{\dagger} - \bar{\theta}_n) B_2 (\bar{\theta}_n^{\dagger} - \bar{\theta}_n) - (\hat{\theta}_n^{\dagger} - \hat{\theta}_n, \hat{\lambda}_n^{\dagger} - \hat{\lambda}_n) B_1 (\hat{\theta}_n^{\dagger} - \hat{\theta}_n, \hat{\lambda}_n^{\dagger} - \hat{\lambda}_n)^{\mathsf{T}}$$

The P-value associated with the test of Chen and Fan (2005) is then given by

$$\frac{1}{N}\sum_{k=1}^{N} 1(\mathcal{CF}_{n,k} \geq \mathcal{CF}_n)$$

The conclusions drawn from this analysis (not reported here) are consistent with Table 11, which indicate that the interval estimates for the BB5 family are compatible with the Galambos model (since $\theta_1=1$ is a possible value) but not with the Gumbel–Hougaard (because $\theta_2=0$ is excluded from its 95% confidence interval). Likewise, the parameter intervals for Family BB1 suggest that neither the Gumbel–Hougaard nor the Kimeldorf–Sampson families are adequate for the data at hand. Additional tools that may help to distinguish between bivariate extreme-value models will be presented in the next section.

Graphical Diagnostics for Bivariate Extreme-Value Copulas

In the bivariate case, extreme-value copulas are characterized by the dependence function A, as in Eq. (11). When the marginal distributions F and G of H are known, a consistent estimator A_n of A has been proposed by Capéraà et al. (1997). It is given by

$$A_n(t) = \exp\left\{ \int_0^t \frac{H_n(z) - z}{z(1-z)} dz \right\}, \quad t \in [0,1]$$

where

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1(Z_i \leq t)$$

is the empirical distribution function of the random sample $Z_1, \ldots, Z_n$ with $Z_i = \log\{F(X_i)\}/\log\{F(X_i)G(Y_i)\}$ for $i \in \{1, \ldots, n\}$

Table 13. Results of the Bootstrap Based on the Cramér–von Mises Statistics $\mathcal{CM}_n$: Observed Statistic, Critical Value $q(\alpha)$ Corresponding to $\alpha=5\%$ and Approximate P-Value, Based on $N=10{,}000$ Parametric Bootstrap Samples

Copula	$\mathcal{CM}_n$	$q(95\%)$	P-value
Gumbel–Hougaard	1.632	3.213	0.3961
Galambos	1.581	3.154	0.4134
Hüsler–Reiss	1.522	3.264	0.5001
BB1	1.692	3.579	0.4160
BB5	1.573	4.140	0.6362

These authors showed that if $Z_{(1)} < \cdots < Z_{(n)}$ are the associated ordered statistics, then A_n can be written in closed form as

$$A_n(t) = \begin{cases} (1-t)Q_n^{1-p(t)} & \text{if } 0 \leq t \leq Z_{(1)} \\ t^{i/n}(1-t)^{1-i/n} Q_n^{1-p(t)} Q_i^{-1} & \text{if } Z_{(i)} \leq t \leq Z_{(i+1)} \\ t Q_n^{-p(t)} & \text{if } Z_{(n)} \leq t \leq 1 \end{cases}$$

in terms of a weight function p so that $p(0)=1-p(1)=1$ and quantities

$$Q_i = \left\{ \prod_{k=1}^{i} Z_{(k)}/(1 - Z_{(k)}) \right\}^{1/n}, \quad i \in \{1, \ldots, n\}$$

The asymptotic behavior of the process $\sqrt{n}\{\log(A_n) - \log(A)\}$ is given by Capéraà et al. (1997) under mild regularity conditions, and could be used to perform a goodness-of-fit test, say, using the Cramér–von Mises statistic

$$n\int_0^1 \log\{A_n(t)/A_{\theta_n}(t)\}^2 dt$$

When the margins are unknown, however, as is most often the case in practice, it would seem reasonable to use a variant $\hat{A}_n$ of the same estimator, with Z_i replaced by the pseudo-observation

$$\hat{Z}_i = \log\left(\frac{R_i}{n+1}\right)/\log\left(\frac{R_i}{n+1} \times \frac{S_i}{n+1}\right), \quad 1 \leq i \leq n$$

Before a proper test can be developed, it will be necessary to examine the asymptotic behavior of the process

$$\sqrt{n}[\log\{\hat{A}_n(t)\} - \log\{A_{\theta_n}(t)\}]$$

This may be the object of future work. For additional discussion on this general theme, refer to Abdous and Ghoudi (2005).

For the time being, a useful graphical diagnostic tool for extreme-value copulas may still consist of drawing $\hat{A}_n$ and A_{θ_n} on the same plot. Fig. 17 shows such a plot for the four families of extreme-value copulas retained for this study. Here, the weight function used was $p(t)=1-t$. The reason for which no goodness-of-fit test could distinguish between these models is obvious from the graph: the generators of the four families are not only fairly close to A_n, they are practically identical.

Conclusion

Using both a learning data set and 85 annual records of volume and peak from the Harricana watershed, this paper has illustrated the various issues involved in characterizing, measuring, and modeling dependence through copulas. The main emphasis was

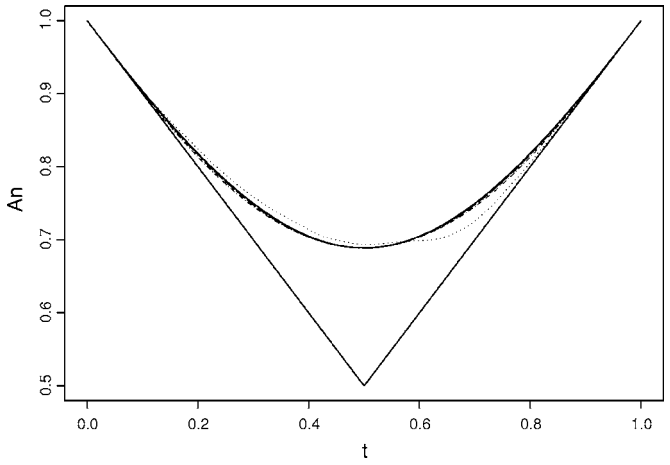


Fig. 17. Plot of A (solid lines) and A_n (dashed line) for the following families of extreme-value copulas: Gumbel–Hougaard, Galambos, Hüsler–Reiss, and BB5

put on inference and testing procedures, many of which have just been developed. Although the presentation was limited to the case of two variables, most of the tools described here extend to the multidimensional case. As the number of variables increases, of course, the intricacies of modeling become more complex, and even the construction of appropriate copula models still poses serious difficulties.

From the analysis presented here, it would appear that several copula families provide acceptable models of the dependence in the Harricana River data. Not surprisingly, several of them are of the extreme-value type, and it is unlikely that choosing between them would make any serious difference for prediction purposes. If forced to express a preference, an analyst would probably want to call on additional criteria, such as model parsimony, etc.

As evidenced by the material in the previous section, the theory surrounding goodness-of-fit testing for this particular class of copulas is still incomplete. For the data at hand, Fig. 17 suggests that an *asymmetric* extreme-value copula model might possibly provide a somewhat better fit. Examples of such models mentioned by Tawn (1988) are the asymmetric mixed and logistic models. In the former

$$A(t) = \phi t^3 + \theta t^2 - (\theta + \phi)t + 1$$

with $0 \leqslant \min(\theta, \theta+3\phi)$ and $\max(\theta+\phi, \theta+2\phi) \leqslant 1$; in the latter

$$A(t) = \{\theta^r(1-t)^r + \phi^r t^r\}^{1/r} + (\theta - \phi)t + 1 - \theta$$

where $\theta \geqslant 0$, $\phi \leqslant 1 \leqslant r$. Khoudraji's device, described among others in Genest et al. (1998), could be used to generate other asymmetric copula models (whether extreme-value or not). The problem of fitting an asymmetric copula to the Harricana River data is left to the reader as a "knowledge integration activity," and the data set is available from the writers for that purpose. Users should keep in mind, however, that in copula matters as in any other statistical modeling exercise, the pursuit of perfection is illusory and a balance should always be struck between fit and parsimony.

The statistical literature on copula modeling is still growing rapidly. In recent years, numerous successful applications of this evolving methodology have been made, most notably in survival analysis, actuarial science, and finance, but also quite recently in hydrology; see, e.g., De Michele and Salvadori (2002), Favre et al. (2004), Salvadori and De Michele (2004), De Michele et al. (2005), and some of the papers in the current issue of the *Journal of Hydrologic Engineering*. It is hoped that this special issue, and the present paper, in particular, can help foster the use of copula methodology in this field of science.

Acknowledgments

Partial funding in support of this work was provided by the Natural Sciences and Engineering Research Council of Canada, the fonds québécois de la recherche sur la nature et les technologies, the Institut de finance mathématique de Montréal, and Hydro-Québec.

References

Abdous, B., Genest, C., and Rémillard, B. (2005). "Dependence properties of meta-elliptical distributions." *Statistical modeling and analysis for complex data problems*, Springer, New York, 1–15.

Abdous, B., and Ghoudi, K. (2005). "Non-parametric estimators of multivariate extreme dependence functions." *J. Nonparam. Stat.*, 17(8), 915–935.

Ali, M. M., Mikhail, N. N., and Haq, M. S. (1978). "A class of bivariate distributions including the bivariate logistic." *J. Multivariate Anal.*, 8(3), 405–412.

Bâ, K. M., Díaz-Delgado, C., and Cârsteanu, A. (2001). "Confidence intervals of quantile in hydrology computed by an analytical method." *Natural Hazards*, 24(1), 1–12.

Barbe, P., Genest, C., Ghoudi, K., and Rémillard, B. (1996). "On Kendall's process." *J. Multivariate Anal.*, 58(2), 197–229.

Bertino, S. (1977). "Sulla dissomiglianza tra mutabili cicliche." *Metron*, 35(1–2), 53–88.

Biau, G., and Wegkamp, M. H. (2005). "A note on minimum distance estimation of copula densities." *Stat. Probab. Lett.*, 73(2), 105–114.

Bobée, B., and Ashkar, F. (1991). *The gamma family and derived distributions applied in hydrology*, Water Resources, Littleton, Colo.

Borkowf, C. B. (2002). "Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation." *Comput. Stat. Data Anal.*, 39(3), 271–286.

Capéraà, P., Fougères, A.-L., and Genest, C. (1997). "A nonparametric estimation procedure for bivariate extreme value copulas." *Biometrika*, 84(3), 567–577.

Capéraà, P., Fougères, A.-L., and Genest, C. (2000). "Bivariate distributions with given extreme value attractor." *J. Multivariate Anal.*, 72(1), 30–49.

Chen, X., and Fan, Y. (2005). "Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection." *Can. J. Stat.*, 33(3), 389–414.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*, Wiley, New York.

Clayton, D. G. (1978). "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence." *Biometrika*, 65(1), 141–151.

De Michele, C., and Salvadori, G. (2002). "A generalized Pareto intensity–duration model of storm rainfall exploiting 2-copulas." *J. Geophys. Res.*, 108(D2), 1–11.

De Michele, C., Salvadori, G., Canossi, M., Petaccia, A., and Rosso, R. (2005). "Bivariate statistical approach to check adequacy of dam spillway." *J. Hydrol. Eng.*, 10(1), 50–57.

Deheuvels, P. (1979). "La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance." *Bull. Cl. Sci., Acad. R. Belg.*, 65(6), 274–292.

Deheuvels, P. (1981). "A Kolmogorov–Smirnov type test for independence and multivariate samples." *Rev. Roum. Math. Pures Appl.*, 26(2), 213–226.

Devroye, L. (1986). *Nonuniform random variate generation*, Springer, New York.

Dhaene, J., and Goovaerts, M. J. (1996). "Dependency of risks and stop-loss order." *Astin Bull.*, 26(2), 201–212.

Embrechts, P., McNeil, A. J., and Straumann, D. (2002). "Correlation and dependence in risk management: Properties and pitfalls." *Risk management: Value at risk and beyond (Cambridge, 1998)*, Cambridge University Press, Cambridge, U.K., 176–223.

Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). "The meta-elliptical distributions with given marginals." *J. Multivariate Anal.*, 82(1), 1–16 [corr. *J. Multivariate Anal.*, 94(1), 222–223].

Favre, A.-C., El Adlouni, S., Perreault, L., Thiémonge, N., and Bobée, B. (2004). "Multivariate hydrological frequency analysis using copulas." *Water Resour. Res.*, 40(W01101), 1–12.

Fermanian, J.-D. (2005). "Goodness-of-fit tests for copulas." *J. Multivariate Anal.*, 95(1), 119–152.

Fermanian, J.-D., and Scaillet, O. (2003). "Nonparametric estimation of copulas for time series." *J. Risk*, 5(4), 25–54.

Fisher, N. I., and Switzer, P. (1985). "Chi-plots for assessing dependence." *Biometrika*, 72(2), 253–265.

Fisher, N. I., and Switzer, P. (2001). "Graphical assessment of dependence: Is a picture worth 100 tests?" *Am. Stat.*, 55(3), 233–239.

Frank, M. J. (1979). "On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$." *Aequ. Math.*, 19(2–3), 194–226.

Fréchet, M. (1951). "Sur les tableaux de corrélation dont les marges sont données." *Ann. Univ. Lyon Sect. A. (3)*, 14(3), 53–77.

Fredricks, G. A., and Nelsen, R. B. (2002). "The Bertino family of copulas." *Distributions with given marginals and statistical modelling*, Kluwer, Dordrecht, The Netherlands, 81–91.

Frees, E. W., and Valdez, E. A. (1998). "Understanding relationships using copulas." *North Am. Act. J.*, 2(1), 1–25.

Gaenssler, P., and Stute, W. (1987). *Seminar on empirical processes*, Vol. 9, Birkhäuser, Basel, Switzerland.

Galambos, J. (1975). "Order statistics of samples from multivariate distributions." *J. Am. Stat. Assoc.*, 70(3), 674–680.

Genest, C. (1987). "Frank's family of bivariate distributions." *Biometrika*, 74(3), 549–555.

Genest, C., and Boies, J.-C. (2003). "Detecting dependence with Kendall plots." *Am. Stat.*, 57(4), 275–284.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). "A semiparametric estimation procedure of dependence parameters in multivariate families of distribution." *Biometrika*, 82(3), 543–552.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1998). "Discussion of 'Understanding relationships using copulas' by E. W. Frees and E. A. Valdez." *North Am. Act. J.*, 2(3), 143–149.

Genest, C., and MacKay, R. J. (1986). "Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données." *Can. J. Stat.*, 14(2), 145–159.

Genest, C., Quessy, J.-F., and Rémillard, B. (2006). "Goodness-of-fit procedures for copula models based on the probability integral transformation." *Scand. J. Stat.*, 33(2), 337–366.

Genest, C., and Rémillard, B. (2004). "Tests of independence and randomness based on the empirical copula process." *Test*, 13(2), 335–369.

Genest, C., and Rémillard, B. (2005). "Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models." *Technical Rep. G-2005-51*, Groupe d'Études et de Recherche en Analyse des Décisions, Montréal.

Genest, C., and Rivest, L.-P. (1993). "Statistical inference procedures for bivariate Archimedean copulas." *J. Am. Stat. Assoc.*, 88(3), 1034–1043.

Genest, C., and Rivest, L.-P. (2001). "On the multivariate probability integral transformation." *Stat. Probab. Lett.*, 53(4), 391–399.

Genest, C., and Verret, F. (2005). "Locally most powerful rank tests of independence for copula models." *J. Nonparam. Stat.*, 17(5), 521–539.

Genest, C., and Werker, B. J. M. (2002). "Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models." *Distributions with given marginals and statistical modeling*, Kluwer, Dordrecht, The Netherlands, 103–112.

Geoffroy, J. (1958). "Contribution à la théorie des valeurs extrêmes." *Publ. Inst. Stat. Univ. Paris*, 7, 37–185.

Ghoudi, K., Khoudraji, A., and Rivest, L.-P. (1998). "Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles." *Can. J. Stat.*, 26(1), 187–197.

Gijbels, I., and Mielniczuk J. (1990). "Estimating the density of a copula function." *Commun. Stat: Theory Meth.*, 19(2), 445–464.

Gumbel, É. J. (1960). "Distributions des valeurs extrêmes en plusieurs dimensions." *Publ. Inst. Stat. Univ. Paris*, 9, 171–173.

Hoeffding, W. (1940). "Maszstabinvariante Korrelationstheorie." *Schriftenr. Math. Inst. Inst. Angew. Math. Univ. Berlin*, 5, 181–233.

Hoeffding, W. (1948). "A class of statistics with asymptotically normal distribution." *Ann. Math. Stat.*, 19(3), 293–325.

Hüsler, J., and Reiss, R.-D. (1989). "Maxima of normal random vectors: Between independence and complete dependence." *Stat. Probab. Lett.*, 7(4), 283–286.

Joe, H. (1993). "Parametric families of multivariate distributions with given margins." *J. Multivariate Anal.*, 46(2), 262–282.

Joe, H. (1997). *Multivariate models and dependence concepts*, Chapman and Hall, London.

Joe, H. (2005). "Asymptotic efficiency of the two-stage estimation method for copula-based models." *J. Multivariate Anal.*, 94(2), 401–419.

Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). "Comparison of semiparametric and parametric methods for estimating copulas." *Comp. Stat. Data Anal.*, 51(6), 2836–2850.

Kimeldorf, G., and Sampson, A. R. (1975). "Uniform representations of bivariate distributions." *Commun. Stat: Theory Meth.*, 4(7), 617–627.

Klaassen, C. A. J., and Wellner, J. A. (1997). "Efficient estimation in the bivariate normal copula model: Normal margins are least favorable." *Bernoulli*, 3(1), 55–77.

Kruskal, W. H. (1958). "Ordinal measures of association." *J. Am. Stat. Assoc.*, 53(4), 814–861.

Lehmann, E. L. (1966). "Some concepts of dependence." *Ann. Math. Stat.*, 37(5), 1137–1153.

Nelsen, R. B. (1986). "Properties of a one-parameter family of bivariate distributions with specified marginals." *Commun. Stat: Theory Meth.*, 15(11), 3277–3285.

Nelsen, R. B. (1999). *An introduction to copulas*, Springer, New York.

Nelsen, R. B., Quesada-Molina, J. J., Rodríguez-Lallena, J. A., and Úbeda-Flores, M. (2003). "Kendall distribution functions." *Stat. Probab. Lett.*, 65(3), 263–268.

Oakes, D. (1982). "A model for association in bivariate survival data." *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, 44(3), 414–422.

Oakes, D. (1994). "Multivariate survival distributions," *J. Nonparam. Stat.*, 3(3-4), 343–354.

Plackett, R. L. (1965). "A class of bivariate distributions." *J. Am. Stat. Assoc.*, 60(2), 516–522.

Quesada-Molina, J. J. (1992). "A generalization of an identity of Hoeffding and some applications." *J. Ital. Stat. Soc.*, 3, 405–411.

R Development Core Team. (2004), "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna.

Salvadori, G., and De Michele, C. (2004). "Frequency analysis via copulas: Theoretical aspects and applications to hydrological events." *Water Resour. Res.*, 40(W12511), 1–17.

Samara, B., and Randles, R. H. (1988). "A test for correlation based on Kendall's tau." *Commun. Stat: Theory Meth.*, 17(9), 3191–3205.

Schucany, W. R., Parr, W. C., and Boyer, J. E. (1978). "Correlation structure in Farlie–Gumbel–Morgenstern distributions." *Biometrika*, 65(3), 650–653.

Shih, J. H., and Louis, T. A. (1995). "Inferences on the association parameter in copula models for bivariate survival data." *Biometrics*,

51(4), 1384–1399.
Sibuya, M. (1960). "Bivariate extreme statistics. I." *Ann. Inst. Stat. Math. Tokyo*, 11, 195–210.
Sklar, A. (1959). "Fonctions de répartition à *n* dimensions et leurs marges." *Publ. Inst. Stat. Univ. Paris*, 8, 229–231.
Tawn, J. A. (1988). "Bivariate extreme value theory: Models and estimation." *Biometrika*, 75(3), 397–415.
Tsukahara, H. (2005). "Semiparametric estimation in copula models." *Can. J. Stat.*, 33(3), 357–375.
Wang, W., and Wells, M. T. (2000). "Model selection and semiparametric inference for bivariate failure-time data (with discussion)." *J. Am. Stat. Assoc.*, 95(1), 62–76.
Whelan, N. (2004). "Sampling from Archimedean copulas." *Quant. Finance*, 4(3), 339–352.