Bayesian copula selection

David Huard, Guillaume Évin, Anne-Catherine Favre

Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Qué., Canada G1K 9A9

Received 1 March 2005; received in revised form 23 June 2005; accepted 30 August 2005
Available online 6 October 2005

Abstract

In recent years, the use of copulas has grown extremely fast and with it, the need for a simple and reliable method to choose the right copula family. Existing methods pose numerous difficulties and none is entirely satisfactory. We propose a Bayesian method to select the most probable copula family among a given set. The copula parameters are treated as nuisance variables, and hence do not have to be estimated. Furthermore, by a parameterization of the copula density in terms of Kendall’s $\tau$, the prior on the parameter is replaced by a prior on $\tau$, conceptually more meaningful. The prior on $\tau$, common to all families in the set of tested copulas, serves as a basis for their comparison. Using simulated data sets, we study the reliability of the method and observe the following: (1) the frequency of successful identification approaches 100% as the sample size increases, (2) for weakly correlated variables, larger samples are necessary for reliable identification.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Copulas; Model selection; Bayes’ theorem; Goodness-of-fit test; Kendall’s tau; Pseudo-likelihood

0. Introduction

In order to extrapolate extreme quantiles from data sets, or to generate random variables, it is usually necessary to select a distribution function matching the available data. The choice of the best distribution is not an exact science and relies on guesswork and testing of multiple hypotheses. Since each hypothesis comes with its particular test, the whole procedure is too complicated for end-users and generally left to experts, along with the interpretation of the results. Furthermore, existing methods cannot compare distributions without specifying an optimal parameter set for each one of them. The selection of the best distribution is thus intertwined with the estimation of parameters, a non-trivial problem itself.

The situation is even worse in the case of two-dimensional distributions, for which even more parameters need to be estimated. Fortunately, the elegant concept of copulas greatly simplifies matters. Copulas are multivariate distributions modeling the dependence structure between variables, irrespective of their marginal distribution. They allow to choose completely different margins, the dependence structure given by the copula, and merge the margins into a genuine multivariate distribution. The choice of the best bivariate distribution can then be done in two steps: choose the optimal margins, and then choose the optimal copula. In this paper, we introduce a simple Bayesian method to choose the “best” copula, given some bivariate data expressed by quantiles.
The structure of the paper is as follows: Section 1 introduces the main ideas of copula theory. Section 2 reviews existing approaches to select copulas and highlights salient features. Section 3 describes the proposed method and its derivation from Bayes’ theorem. Results from numerical simulations are shown in Section 4, along with their analysis. Finally, we draw conclusions on the overall performance of the method and propose ideas for future work.

1. Copula theory

The concept of copula has been introduced by Sklar (1959) in the following way

**Copula Definition.** A copula is a joint distribution function of standard uniform random variables. That is,

$$ C(u_1, \ldots, u_p) = \Pr \{ U_1 \leq u_1, \ldots, U_p \leq u_p \}, $$

where $U_i \sim U(0, 1)$ for $i = 1, \ldots, p$.

For a more formal definition of copulas, the reader is referred to Nelsen (1999). Using the probability integral transformation, it is straightforward to see that a copula computed at $p$-dimensional generalization. The popular Archimedean 2-copulas (Genest and MacKay, 1986) for $p$-copula $C$ such that for all $x$ in $\mathbb{R}^p$,

$$ F(x_1, x_2, \ldots, x_p) = C(F_1(x_1), F_2(x_2), \ldots, F_p(x_p)), $$

where $\mathbb{R}^p$ denotes the extended real line $[-\infty, \infty]$.

According to Sklar’s theorem, copulas separate marginal behavior, as represented by the $F_i$’s, from the dependence structure. This constitutes one great advantage of copulas. In the usual representation of joint probabilities via multivariate distribution functions, the two cannot be separated. The general theory about copulas is summarized in Joe (1997), Nelsen (1999) or more recently in Cherubini et al. (2004). Copulas have been widely used in financial mathematics to determine the Value at Risk (see for example, Embrechts et al., 2002, 2003; Bouyé et al., 2000). Other fields of applications involve lifetime data analysis (Bagdonavicius et al., 1999), actuarial science (Frees and Valdez, 1998), and more recently, hydrology (De Michele and Salvadori, 2003; Favre et al., 2004).

Most copula applications are concerned with bivariate data. One reason for this is that relatively few copula families have practical $p$-dimensional generalization. The popular Archimedean 2-copulas (Genest and MacKay, 1986) for instance, have two known generalizations, both of them afflicted by serious shortcomings. Archimedean 2-copulas are defined as

$$ C(u_1, u_2) = \begin{cases} \varphi^{-1}(\varphi(u_1) + \varphi(u_2)) & \text{if } \sum_{i=1}^{2} \varphi(u_i) \leq \varphi(0), \\ 0 & \text{otherwise} \end{cases} $$

with $\varphi(u)$ a $\varphi^{-2}$ function satisfying $\varphi(1) = 0$, $\varphi'(u) < 0$ ($\varphi$ is decreasing) and $\varphi''(u) > 0$ ($\varphi$ is convex) for all $0 \leq u \leq 1$. $\varphi(u)$ is called the generator of the copula. The first generalization, termed symmetric (Joe, 1997), uses the same generator, thus the same dependence, for all variables

$$ C(u_1, \ldots, u_p) = \varphi^{-1}(\varphi(u_1) + \cdots + \varphi(u_p)) . $$

Since all variables are described by the same dependence, this generalization is too simplistic for most real life applications. The second generalization, termed asymmetric (Whelan, 2004), uses $(p-1)$ generators. For $p = 3$, the
Although copulas (Clayton, Ali–Mikhail–Haq (AMH), Gumbel, Frank, Joe, A12, A14) belong to the Archimedean class. Note they model a wide variety of dependence structure and cover most applications found in the literature. The first seven when the copula is expressed using the normal cdf:

For Archimedean copulas, Kendall’s tau can be written as

\[ \tau = \text{Pr}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \text{Pr}[(X_1 - X_2)(Y_1 - Y_2) < 0]. \]

Kendall’s tau may also be expressed simply in term of the copula function, as the expected value of the function \( C(U, V) \) of uniform (0, 1) random variables \( U \) and \( V \), whose joint distribution function is \( C \), i.e.

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

For Archimedean copulas, Kendall’s tau can be written as \( \tau = 1 + 4 \int_0^1 \left( \phi(t)/\varphi(t) \right) \, dt. \)

The nine copula families used in this paper, chosen for their analytical properties, are defined in Table 1. Put together, they model a wide variety of dependence structure and cover most applications found in the literature. The first seven copulas (Clayton, Ali–Mikhail–Haq (AMH), Gumbel, Frank, Joe, A12, A14) belong to the Archimedean class. Note that A12 and A14 are coined from the order of appearance in Nelsen (1999). We also consider two non-Archimedean copulas: the Farlie–Gumbel–Morgenstern (FGM), a copula with a quadratic section, and Gauss copula, an elliptical copula. Note that the link between Gauss copula and the classical multivariate normal distribution is made explicit when the copula is expressed using the normal cdf: \( C(u, v|\theta) = \Phi_0 \left( \Phi^{-1}(u), \Phi^{-1}(v) \right) \). Form more details about elliptical

| Copula | \( C(u, v|\theta) \) | \( \theta \in \Omega \) |
|--------|------------------|------------------|
| Clayton | \( \left( \frac{u^{-\theta} + v^{-\theta} - 1}{uv} \right)^{-1/\theta} \) | \([0, \infty]\) |
| AMH | \( 1 - \frac{\theta(1-u)(1-v)}{1 - \theta(1-u)(1-v)} \) | \([-1, 1]\) |
| Gumbel | \( \exp \left\{ -\left[ (-\ln u)^{\theta} + (-\ln v)^{\theta} \right]^{1/\theta} \right\} \) | \([1, \infty]\) |
| Frank | \( \frac{1}{\theta} \ln \left( \frac{1 + \left( e^{-\theta u} - 1 \right) \left( e^{-\theta v} - 1 \right)}{e^{-\theta} - 1} \right) \) | \([-1, 1] \setminus \{0\}\) |
| Joe | \( 1 - \left[ \left( 1 - u \right)^{\theta} + \left( 1 - v \right)^{\theta} - \left( 1 - u \right)^{\theta} \left( 1 - v \right)^{\theta} \right]^{1/\theta} \) | \([1, \infty]\) |
| A12 | \( \frac{1 + \left[ (u^{-1} - 1)^{\theta} + (v^{-1} - 1)^{\theta} \right]^{1/\theta}}{uv + \theta uv(1-u)(1-v)} \) | \([1, \infty]\) |
| A14 | \( \frac{1 + \left[ (u^{-1/\theta} - 1)^{\theta} + (v^{-1/\theta} - 1)^{\theta} \right]^{1/\theta}}{uv + \theta uv(1-u)(1-v)} \) | \([1, \infty]\) |
| FGM | | \([-1, 1]\) |
| Gauss | \( \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp \left( \frac{2\theta \omega - s^2 - \omega^2}{2(1-\theta^2)} \right) \, d\omega \) | \([-1, 1]\) |
copulas, one of the few families with practical $p$-dimensional generalizations, the reader is referred to Frahm et al. (2003).

2. Review of copula selection methods

The most commonly employed methods to select the best copula are based on a likelihood approach, which is used to define indicators of performance, as for example, the Akaike Information Criteria (AIC). Chen and Fan (2005) propose pseudo-likelihood ratio tests for selecting semi parametric multivariate copula models in which the marginal distributions are unspecified. For Archimedean copulas, Genest and Rivest (1993) proposed to compare the one-dimensional function

$$K_\theta(t) = \Pr(C(u, v|\theta) < t),$$

with its non-parametric estimation $K_n$, given by

$$K_n(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(e_{jn} \leq t),$$

where $e_{jn} = (1/n) \sum_{k=1}^{n} \mathbb{1}(X_{1k} \leq X_{1j}, \ldots, X_{pk} \leq X_{pj})$. The best copula is then the one for which the function $K_\theta$ is closest to $K_n$. Durrleman et al. (2000) suggested to choose the copula minimizing the distance ($L^2$-norm, Kolmogorov, etc.) from $K_\theta$ to the non-parametric estimation $K_n$. Using the same idea, they computed a distance based on the discrete $L^2$-norm. This distance is calculated between an empirical copula of Deheuvels (1979) and the tested copulas. For an exhaustive presentation of these procedures, we refer to the working paper of Durrleman et al. (2000).

In recent years, various authors developed goodness-of-fit tests (GOF tests) for copulas. Genest et al. (2005) proposed an user-friendly and powerful tool, a GOF test statistic with a non-truncated version of Kendall’s process

$$\mathbb{K}_n(t) = \sqrt{n} \left\{ K_n(t) - K_{\theta_n}(t) \right\},$$

where $\theta_n$ denotes a robust estimation of $\theta$. The expression for the statistic is straightforward and the test has nice properties. For example, it sustains the prescribed error probability of the first kind under the null hypothesis, even with small sample sizes. Nevertheless, an explicit expression is needed for $K_\theta$, which limits the set of copulas for which the GOF test statistic can be computed.

An easy way to construct GOF tests for copulas is to consider $p$-dimensional $\chi^2$ tests. The methodology is presented in Pollard (1979). Dobrić and Schmidt (2004) recently used this method in a financial application. The main criticism about this approach concerns the arbitrary choice of the subsets that divide the $p$-dimensional space $[0, 1]^p$ (Kendall and Stuart, 1983). Also, the calculation of the empirical critical values proves to be troublesome.

Several authors employed the transformation of Rosenblatt (1952) to test whether the transformed random variables set is composed of independent uniformly distributed variables, as is the case under the null hypothesis (Justel et al., 1997). Using Rosenblatt’s transformation, Chen et al. (2003) compared the kernel density estimation of their transformed random variables to the uniform density. As noted in Fermanian (2005), Rosenblatt’s transformation may be a tedious preliminary task, especially for high dimensions. Therefore, Fermanian (2005) presents a GOF test based directly on the kernel density estimation of the original multivariate data. This test, however, requires some heavy numerical integration. Moreover, it was noted through simulations that with small sample sizes, it is difficult to sustain the prescribed error probability of the first kind.

The comments in this section are based on a survey (Evin, 2004) that compared the results of the GOF tests presented above. Numerous simulations led the authors to conclude that the test of Genest et al. (in press) is the only one to be unbiased, and by far the most powerful. All these tests, however, rely on previous estimation of an optimal parameter set. Strictly speaking, comparisons are made between copulas with given parameters, and not between copula families. We suggest that model selection methods should be independent of the parameter choice.
3. Selection of the right copula

We present in this section a Bayesian model selection method that attributes weights to copula families. It does not rely on parameter estimation and to our knowledge, may be applied to all known copulas.

Let \( \mathcal{C} \) denote the set of all copulas. From this set, we select a finite subset \( \mathcal{C}_Q \subset \mathcal{C} \) of copulas to be included in the proposed method. Each family of copulas in \( \mathcal{C}_Q \) is identified by \( C_l \), with \( l = 1, \ldots, Q \). The proposed method consists simply in defining \( Q \) hypotheses:

\[ H_l : \text{The data come from copula } C_l, \quad l = 1, \ldots, Q \]

and then computing \( \Pr(H_l|D) \), the probability of each hypothesis given the data \( D \). We will suppose that the data set \( D \) is composed of \( n \) mutually independent pairs of quantiles \( (u_i, v_i), \quad i = 1, \ldots, n \). Note that the independence assumption may not hold if quantiles are computed empirically using ranks. In that case, the method must be thought of as approximative and used with caution. Applying Bayes’ theorem, we get for each family:

\[
\Pr(H_l|D, I) = \frac{\Pr(D|H_l, I) \Pr(H_l|I)}{\Pr(D|I)}, \tag{2}
\]

where \( \Pr(D|H_l, I) \) is the likelihood, \( \Pr(H_l|I) \) is the prior on the copula family, \( \Pr(D|I) \) the normalization constant and \( I \) stands for any relevant additional knowledge. The “right” copula is then simply the copula with the highest \( \Pr(H_l|D, I) \).

3.1. Likelihood

The likelihood \( \Pr(D|H_l, I) \) in Eq. (2) is the probability of “drawing” data \( D \) from the \( l \)th copula. However, for most copulas, there exists no such explicit expression because the copula density depends on a parameter \( \theta \) (cf. Table A.2). Note also that the parameterization is arbitrary, in the sense that we could choose any function \( \beta = g(\theta) \) and replace \( c(u, v|\theta) \) by \( \tilde{c}(u, v|\beta) = c(u, v|g^{-1}(\beta)) \).

There is, of course, no wrong choice for the parameter, and for reasons that will soon be clear, we will choose Kendall’s tau \( \tau = g_l(\theta) \) to be the common parameter for all copulas in \( \mathcal{C}_Q \) (cf. Table 2).

We introduce Kendall’s tau in Eq. (2) as a nuisance variable:

\[
\Pr(H_l|D, I) = \int_{-1}^1 \Pr(H_l, \tau|D, I) \, d\tau
= \int_{-1}^1 \frac{\Pr(D|H_l, \tau, I) \Pr(H_l|\tau, I) \Pr(\tau|I) \, d\tau}{\Pr(D|I)}, \tag{3}
\]

where \( \Pr(H_l|\tau, I) \) is the prior on the family hypothesis and \( \Pr(\tau|I) \) is the prior density on Kendall’s tau. The likelihood \( \Pr(D|H_l, \tau, I) \) now depends on \( \tau \) and, if all \( n \) data pairs are mutually independent, can be computed from the copula density:

\[
\Pr(D|H_l, \tau, I) = \prod_{i=1}^n \Pr(u_i, v_i|\tau, I)
= \prod_{i=1}^n c_l(u_i, v_i|g_l^{-1}(\tau)), \tag{4}
\]

where \( c_l(u_i, v_i|g_l^{-1}(\tau)) \) is the density of the \( l \)th copula (cf. Table A.2). If quantiles are computed using ranks, mutual independence cannot be guaranteed and Eq. (4) rather describes a pseudo-likelihood. In that case, results obtained should be considered as approximations. Fortunately, this approximation is thought to improve with increasing sample size, as the dependence between ranks decreases.
Table 2
Kendall’s tau and its domain of definition for the copulas used in this paper

<table>
<thead>
<tr>
<th>Copula</th>
<th>( \tau = g(\theta) )</th>
<th>( \tau \in \Omega )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>( 1 - \frac{2}{2 + \theta} )</td>
<td>([0, 1]) (\setminus) ([0])</td>
</tr>
<tr>
<td>AMH</td>
<td>( 1 - \frac{2}{3} \frac{\theta^2 \ln(1 - \theta) - 2 \theta \ln(1 - \theta) + \theta + \ln(1 - \theta)}{\theta^2} )</td>
<td>([-0.181726, \frac{1}{3}])</td>
</tr>
<tr>
<td>Gumbel</td>
<td>( 1 - \theta^{-1} )</td>
<td>([0, 1])</td>
</tr>
<tr>
<td>Frank</td>
<td>( 1 - \frac{4}{\theta} \left( 1 - \frac{1}{\theta} \int_0^{\theta} \frac{t}{\sqrt{t^2 - 1}} , dt \right) )</td>
<td>([-1, 1]) (\setminus) ([0])</td>
</tr>
<tr>
<td>Joe</td>
<td>No closed form</td>
<td>Discontinuities</td>
</tr>
<tr>
<td>A12</td>
<td>( 1 - \frac{2}{3\theta} )</td>
<td>([\frac{1}{3}, 1])</td>
</tr>
<tr>
<td>A14</td>
<td>( 1 - \frac{2}{1 + 2\theta} )</td>
<td>([\frac{1}{3}, 1])</td>
</tr>
<tr>
<td>FGM</td>
<td>( \frac{2\theta}{\theta} )</td>
<td>([-\frac{2}{5}, \frac{2}{5}])</td>
</tr>
<tr>
<td>Gauss</td>
<td>( \frac{2}{\pi} \arcsin(\theta) )</td>
<td>([-1, 1])</td>
</tr>
</tbody>
</table>

3.2. Priors

To select priors, a host of methods exists (Kass and Wasserman, 1996), some said objective, others said subjective, but the choice of the method itself remains subjective and open to debate. Our approach consists in stating desiderata precise enough to define uniquely the prior on \( \tau \) as well as the prior on the family. The subjectivity of the prior’s choice is then confined to these desiderata and put in evidence for criticism. Let us state these basic desiderata, the additional information denoted earlier by \( I \):

\((I_1)\)  Kendall’s tau belongs to the set \( A \) and each outcome of \( \tau \in A \) is equally likely;

\((I_2)\) for a given \( \tau \), all families satisfying \( \tau \in \Omega_l \) are equally probable,

where \( \Omega_l \) is \( \tau \)’s domain for the \( l \)th copula (see Table 2). The purpose of \( A \) is to give the user the possibility to include additional knowledge about the correlation between the variables. For example, if the correlation is known to be positive, we may assume that \( A = [0, 1] \). In the case where no information is available, \( A \) is simply put equal to \([-1, 1]\).

Desideratum \((I_2)\) determines the prior on the family. Indeed, since all families are equally probable with respect to a given \( \tau \) (for \( \tau \in \Omega_l \)),

\[
\Pr(H_l | \tau, I_2) \propto 1 \ (\tau \in \Omega_l) .
\]  

Similarly, Desideratum \((I_1)\) specifies the prior on \( \tau \):

\[
\Pr(\tau | I_1) = \begin{cases} \frac{1}{\lambda(A)} & \tau \in A, \\ 0 & \text{otherwise}, \end{cases}
\]

where \( \lambda(\cdot) \) denotes the Lebesgue measure, here the width of the interval spanned by \( A \).

These priors, chosen mainly for testing purposes, reflect complete ignorance of the correlation between the variables as well as no preference over the copula family. However, in real cases, cogent information would probably be available and should be included in the calculation via an informative prior. For example, if \( \tau \) is known to lie around a certain
value, a beta distribution, extended to the range \([-1, 1]\), could provide an effective way to describe the prior on \(\tau\):

\[
\Pr(\tau) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left( \frac{1 + \tau}{2} \right)^{\alpha-1} \left( \frac{1 - \tau}{2} \right)^{\beta-1}.
\]

To estimate the parameters, the easiest way would probably be to vary \(\alpha, \beta\) until the shape of the distribution agrees with our intuitive perception of the probability. Information on the copula family may also be included, by modifying the weight given to each family. For example, if we knew beforehand that the type of data to analyse exhibits correlated extreme events, we could increase the weight given to copulas with strong tail dependence.

### 3.3. Normalization

Plugging Eqs. (4), (5) and (6) into (3), we find

\[
\Pr(H_l|D, I) = \frac{1}{\Pr(D|I)} \int_{-1}^{1} \prod_{i=1}^{n} c_{l}(u_i, v_i|g^{-1}_l(\tau)) \cdot 1(\tau \in \Omega_l) \cdot \frac{1(\tau \in A)}{\lambda(A)} \, d\tau
\]

In general, the normalization constant \(\Pr(D|I)\) in Eq. (7) is computed using the sum rule (Jaynes and Bretthorst, 2003):

\[
\Pr(D|I) = \sum_{l=1}^{Q} \Pr(D|H_l, I) \, \Pr(H_l|I).
\]

However, the sum rule is only true if the hypotheses \(H_l\) are mutually exclusive and the set exhaustive. In our case, arguments invalidate both claims. First, if the data come from a copula not in \(C_Q\), the set of hypotheses is clearly not exhaustive. Second, if the set contains two or more copulas that are very similar, the hypotheses should not be considered completely exclusive. Solutions to insure exhaustiveness and take non-exclusivity into account are discussed in Section 4.3, their application, however, is beyond the scope of this article. We will hence limit the computation to the weights \(W_l\):

\[
W_l = \frac{1}{\lambda(A)} \int_{\Omega_l \cap A} \prod_{i=1}^{n} c_{l}(u_i, v_i|g^{-1}_l(\tau)) \, d\tau.
\]

Note that in the figures shown below, the weights are normalized for convenience.

### 4. Simulations and analysis

To assess the performance of the method, we select eight one-parameter copulas to form the subset \(C_Q\): Clayton, AMH, Gumbel, Frank, A12, A14, FGM and Gauss. Those copulas are chosen because analytical formulas exist both for the density (cf. Table A.2) and Kendall’s tau (cf. Table 2). These copulas are then used to generate data sets of different sizes and correlations. We study the cases of small negative dependence \(\tau = -0.2\) and small positive dependence \(\tau = 0.2\), using samples of sizes \(n = 30, 300\) and 600, and the cases of medium dependence \(\tau = 0.5\) and large dependence \(\tau = 0.7\), using samples of sizes \(n = 30, 100\) and 300. For each copula, for each \(n\) and for each \(\tau\), 1000 data sets are generated over which Eq. (9) for \(l = 1\) to \(Q\) is computed.

#### 4.1. Main results

Once the weights are computed from (9), we count the number of times the right copula is chosen by the method, that is, the number of times it attains the highest weight among copulas from the set. The results are presented in
Table 3. Number of successful identifications over 1000 trials

<table>
<thead>
<tr>
<th>Copula</th>
<th>( \tau )</th>
<th>(-0.2)</th>
<th>0.2</th>
<th>0.5</th>
<th>0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n )</td>
<td>30</td>
<td>300</td>
<td>600</td>
<td>30</td>
</tr>
<tr>
<td>Clayton</td>
<td>466</td>
<td>865</td>
<td>951</td>
<td>774</td>
<td>917</td>
</tr>
<tr>
<td>AMH</td>
<td>127</td>
<td>495</td>
<td>734</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gumbel</td>
<td>411</td>
<td>850</td>
<td>954</td>
<td>667</td>
<td>853</td>
</tr>
<tr>
<td>Frank</td>
<td>567</td>
<td>627</td>
<td>696</td>
<td>409</td>
<td>484</td>
</tr>
<tr>
<td>A12</td>
<td>291</td>
<td>610</td>
<td>795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A14</td>
<td>216</td>
<td>521</td>
<td>786</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGM</td>
<td>52</td>
<td>478</td>
<td>637</td>
<td>52</td>
<td>438</td>
</tr>
<tr>
<td>Gauss</td>
<td>549</td>
<td>750</td>
<td>832</td>
<td>330</td>
<td>576</td>
</tr>
</tbody>
</table>

Table 3. We note the following:

- As expected, the method becomes more accurate as the sample size grows.
- As \( \tau \) approaches zero, larger samples are required for a successful identification.
- Some copulas are easier to identify than others.

Regarding the first comment, results show that the method converges to the right copula. That is, we suspect that as \( n \) increases, the probability of a successful identification approaches one.

The second comment may be explained by noting that certain copulas show a similar behavior when \( \tau \) approaches zero. That is, copulas cluster into classes defined by an identical asymptotic density (Genest et al., 2005a). For example, FGM, AMH and Frank copula are part of the same class, whose asymptotic density is given by \( c(u, v|\tau \to 0) \propto uv(1-u)(1-v) \). This similarity between copulas is more clearly seen in Fig. 2, displaying weights computed by the method, averaged over 1000 trials.

Concerning the third comment, the ease with which copulas are identified is related to the precision of similar copulas. Copulas that look alike obtain similar weights and hence, are difficult to identify. Conversely, copulas with peculiar densities stand out and are easily identifiable. The Clayton, for instance, was shown in Saïd (2004) to display a behavior different from other copulas. This fact is also pointed out in Genest and Verret (2005), where the authors showed that in the context of locally most powerful rank tests of independence, against alternatives expressed by copula models, Clayton differs from other families. This explains why in all cases, Clayton is the copula most often successfully identified.

To gain a better feeling of the reliability of the method, we compute the average weight obtained for each copula. Figs. 1 and 2 illustrate those weights, along with the weight obtained for the independent copula \( c(u, v) = 1 \), as a measure of comparison. Fig. 2 shows clearly the connection between AMH, Frank and FGM. Fig. 1 suggests an unexpected connection between Gumbel and A14. Comparison of the two figures also shows how identification becomes easier when variables are strongly correlated. Note also that the independent copula obtains a relatively high weight for \( \tau = 0.2 \) and \( n = 30 \). In that case, the independent copula is selected around 50% of the time, signifying that for such a small \( \tau \), there is not sufficient data to distinguish from independence (see Table A.1).

4.2. Additional results and comments

In order to complete the analysis, further simulations are done to explore different questions. In particular, we inquire the impact of empirically computed quantiles on the reliability of the method, discuss what happens when the data come from an “unknown” copula, provide some directions about how to include copulas with multidimensional parameters and suggest a way to normalize the weights.

4.2.1. Empirical quantiles

As noted earlier, if \( \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \) are not known exactly, the method is approximate. To evaluate the effect of this approximation, we use pairs \((u_i, v_i)\) generated from the copulas to compute \((x_i, y_i)\) by \( x_i = F_1^{-1}(u_i) \) and \( y_i = F_2^{-1}(v_i) \), where \( F_1 \)
Clayton AMH Gumbel Frank A12 A14 FGM Gauss Ind

$\tau = 0.7$

![Bar chart showing weights assigned to copulas](image)

Fig. 1. Weights assigned to copulas, averaged over 1000 trials, for a Kendall’s $\tau$ of 0.7 and samples sizes $n = 30$, 100 and 300. Bold font indicates the copula family that generated the samples.

and $F_2$ are asymmetrical Gamma distributions. Empirical quantiles $(\hat{u}_i, \hat{v}_i)$ are then estimated with $\hat{u}_i = 1/(n - 1)#\{j \neq i : x_j \leq x_i\}$. The weights are computed using the empirical quantiles and compared to those obtained with the theoretical quantiles. Table 4 compares the number of successful identification in both cases. As expected, the uncertainty induced by the empirical quantiles reduces the number of successful identifications. The effect, however, weakens for large samples.

4.2.2. Unknown copula hypothesis

To understand what happens when a sample comes from a copula not in the set $\mathcal{C}_Q$, we run simulations using samples generated by Joe copula, with $\tau = 0.5$ and $n = 30$, 100 and 300. In the vast majority of cases, Gumbel copula obtains the highest weight. Although this result is hardly surprising since Joe and Gumbel have similar shapes, it highlights
the fact that no warning signal is given when the data are coming from a copula not in $\mathcal{C}_Q$. To rectify the situation, a solution would be to include an additional hypothesis:

$H_{Q+1}$: The data come from an “unknown” copula.

The difficulty is, of course, to define the density of an “unknown” copula. A solution proposed by Bretthorst (1996) in the context of radar target identification, is to describe the “unknown” hypothesis by an over-parameterized (o.-p.) model. This o.-p. model must have enough parameters to capture virtually every conceivable behavior. Due to its high flexibility, it should reach likelihoods equivalent to those of the right copula, but its priors on the extra parameters would reduce its overall posterior probability. Hence, it would obtain weights lower than the right copula, but higher than false copulas. The o.-p. model would then only be selected when the true model is not in $\mathcal{C}_Q$. 

Fig. 2. Weights assigned to copulas, averaged over 1000 trials, for a Kendall’s $\tau$ of 0.2 and samples sizes $n = 30, 300$ and 600. Bold font indicates the copula family that generated the samples.
Table 4
Comparison of the number of successful identifications over 1000 trials for theoretical and empirical quantiles, using \( \tau = 0.5 \)

<table>
<thead>
<tr>
<th>Copula</th>
<th>Theoretical</th>
<th>Empirical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30</td>
<td>300</td>
</tr>
<tr>
<td>Clayton</td>
<td>742</td>
<td>903</td>
</tr>
<tr>
<td>Gumbel</td>
<td>667</td>
<td>860</td>
</tr>
<tr>
<td>Frank</td>
<td>571</td>
<td>842</td>
</tr>
<tr>
<td>A12</td>
<td>322</td>
<td>604</td>
</tr>
<tr>
<td>A14</td>
<td>201</td>
<td>484</td>
</tr>
<tr>
<td>Gauss</td>
<td>404</td>
<td>685</td>
</tr>
</tbody>
</table>

4.2.3. Two parameters copula

In principle, the method can be applied to copulas with any number of parameters. In practice, however, difficulties appear since it implies high-dimensional integration and the definition of multidimensional priors. To see how it could be done, let us consider the case of a copula with two parameters \( \hat{\theta} = (\theta_1, \theta_2) \). To apply the method, we parameterize the copula in terms of \( \theta_1 \) and \( \theta_2 \), where \( \theta_2 \) is a relevant quantity, uniquely determined by \( \theta_1, \theta_2 \) = \( g (\hat{\theta}) \). One such quantity could be the tail dependence, a measure of the probability of correlated extreme events (Juri and Wüthrich, 2003; Schmidt, 2002). A bivariate prior \( \pi (\theta_1, \theta_2) \) would then have to be specified, and the integration performed over \( \theta_1 \) and \( \theta_2 \).

Note that the computation of \( g^{-1} (\theta_1, \theta_2) \) might not be efficient. A better and equivalent solution consists in transforming the prior on \( \theta_1, \theta_2 \) into a prior on \( \hat{\theta} \) using the usual formula for variable substitution:

\[
\pi_\theta (\hat{\theta}) = \pi_{\theta_2} (g (\hat{\theta})) |J|,
\]

where \( J \) is the determinant of the Jacobian of the transformation. This method has been used in the computation for this paper. It has the advantage that \( J \) generally has a closed form, which is often not the case for \( g^{-1} \).

4.3. Normalization of the weights

As discussed in Section 3.3, the basic sum rule should not be applied since the set of hypotheses may not be exhaustive \( \Pr (H_1 + \cdots + H_Q) \neq 1 \) (+ stands for the logical OR operator), and some of the hypotheses might not be mutually exclusive \( \Pr (H_i, H_j) \neq 0, \quad i \neq j \). However, normalization is possible, although potentially tedious, if one introduces the unknown copula hypothesis and applies the extended sum rule. Indeed, adding the unknown copula hypothesis makes the set becomes exhaustive by definition, and the extended sum rule \( \Pr (A + B) = \Pr (A) + \Pr (B) - \Pr (A, B) \) allows for non-exclusive hypotheses. If the set contains three hypotheses, for instance, we would reckon:

\[
\Pr (D) = \Pr (D, H_1 + H_2 + H_3) \\
= \Pr (D|H_1) \Pr (H_1) + \Pr (D|H_2) \Pr (H_2) + \Pr (D|H_3) \Pr (H_3) \\
- \Pr (D|H_1, H_2) \Pr (H_1, H_2) - \Pr (D|H_1, H_3) \Pr (H_1, H_3) \\
- \Pr (D|H_2, H_3) \Pr (H_2, H_3) + \Pr (D|H_1, H_2, H_3) \Pr (H_1, H_2, H_3).
\]

Thus, in principle, it is possible to compute genuine probabilities instead of weights.

5. Conclusion and future work

We presented a novel method to select the “right” copula given a data set. This method, built on a straightforward application of Bayesian analysis, provides interesting advantages over the commonly used statistical tests: it has a simple interpretation, it is independent of parameter choices, it is easy to implement numerically, and, as judged from our simulations, provides reliable identification, even for small samples. Also, it has the conceptual advantage of being
Table A.1
Number of times the right copula is identified compared with the number of times the independent copula is identified for \( \tau = 0.2 \) over 1000 trials

<table>
<thead>
<tr>
<th>Copula</th>
<th>Right copula</th>
<th></th>
<th></th>
<th></th>
<th>[30</th>
<th>[300</th>
<th>[600</th>
<th>Independent</th>
<th></th>
<th>[30</th>
<th>[300</th>
<th>[600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>311</td>
<td>865</td>
<td>951</td>
<td>451</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMH</td>
<td>19</td>
<td>495</td>
<td>734</td>
<td>553</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gumbel</td>
<td>266</td>
<td>849</td>
<td>954</td>
<td>495</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>197</td>
<td>483</td>
<td>618</td>
<td>547</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGM</td>
<td>0</td>
<td>437</td>
<td>635</td>
<td>585</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gauss</td>
<td>137</td>
<td>575</td>
<td>760</td>
<td>504</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.2
Density of the copulas used in this paper

| Copula | \( c(u, v|\theta) = \frac{\partial^2 C(u, v|\theta)}{\partial u \partial v} \) |
|---|---|
| Clayton | \( (1 + \theta)u^{-1-\theta}v^{-1-\theta}(-1 + u^{-\theta} + v^{-\theta})^{-2-1/\theta} \) |
| AMH | \( [-1 + \theta^2(-1 + u + v - uv) - \theta(-2 + u + v + uv)] \) |
| Gumbel | \( \frac{-\log(u)^{1+\theta}}{\exp(-\theta) - \exp(-\theta(1 + u)) - \exp(-\theta(u + v))} -\frac{1}{uv} \) |
| Frank | \( \frac{\theta \exp(\theta(1 + u + v)(-1 + \exp(\theta))]}{(\exp(\theta) - \exp(\theta(1 + u)) - \exp(\theta(u + v)) + \exp(\theta(u + v))]^2} \) |
| Joe | \( (1 - u)^{-1+\theta} \left[ \theta \left[ -1 + (1 - u) \right] \left[ -1 + (1 - v) \right] \right] \left[ (1 - u)^{\theta} + (1 - v)^{\theta} - (1 - u)^{\theta(1 - v)^{\theta}} \right] \) |
| A12 | \( \left[ (1 + \frac{1}{u})^\theta - 1 + \theta + (\theta + 1) \left[ (1 + \frac{1}{u})^\theta + (1 + \frac{1}{v})^\theta \right] \right] \) |
| A14 | \( \left[ (1 + \frac{1}{u})^\theta - 1 + \theta + 2 \left[ (1 + \frac{1}{u})^\theta + (1 + \frac{1}{v})^\theta \right] \right] \left[ (1 + \frac{1}{u})^\theta + (1 + \frac{1}{v})^\theta \right] \) |
| FGM | \( 1 + \theta(1 - 2u)(1 - 2v) \) |
| Gauss | \( \frac{1}{\sqrt{1 - \rho^2}} \exp \left( \frac{\phi^{-1}(u)^2 + \phi^{-1}(v)^2}{2} \right) \exp \left( \frac{2\rho\phi^{-1}(u)\phi^{-1}(v) - \phi^{-1}(u)^2 - \phi^{-1}(v)^2}{2(1 - \rho^2)} \right) \) |

A genuine model selection method, in the sense that it does not depend on the choice of an optimal parameter. The framework allows naturally for copulas with any number of parameters and for higher dimensional copulas. Furthermore, it can be applied to any copula, as long as the copula density and Kendall’s tau can be computed numerically.
Future work will concern the following topics: theoretical Bayesian framework to demonstrate the convergence of the method, over-parameterized copula to model the “unknown copula” hypothesis, higher dimensional copulas and the development of a similar method to select margins.

Acknowledgements

Codes are available upon request. Thanks to Peter Perkins (The Mathworks) for his useful set of copula related functions and to Christian Genest and Alain Mailhot for helpful comments and suggestions. The authors are indebted to Eva-Maria Restle for her critical review of the paper, as well as to both referees for their scrutiny.

Appendix A

The number of times the right copula is identified compared with the number of times the independent copula is identified for $\tau = 0.2$ over 1000 trials is given in Table A.1. The density of the copulas used in this paper is given in Table A.2.

References


