A class of multivariate copulas based on products of bivariate copulas

Gildas Mazo, Stéphane Girard, Florence Forbes

A class of multivariate copulas based on products of bivariate copulas

Gildas Mazo (gildas.mazo@free.fr), Stéphane Girard and Florence Forbes

Inria and Laboratoire Jean Kuntzmann, Grenoble, France

Abstract

Copulas are a useful tool to model multivariate distributions. While there exist various families of bivariate copulas, much fewer has been done when the dimension is higher. In this paper we propose a class of multivariate copulas based on products of transformed bivariate copulas. No constraints on the parameters refrain the applicability of the proposed class. Furthermore the analytical forms of the copulas within this class allow to naturally associate a graphical structure which helps to visualize the dependencies and to compute the likelihood efficiently even in high dimension.

Keywords: maximum-likelihood inference, graphical models, message-passing algorithm, multivariate, copula.

1 Introduction

The modelling of random multivariate events is a central problem in various scientific domains and the construction of multivariate distributions able to properly model the variables at play is challenging. A useful tool to deal with this problem is the concept of copulas. Let $(X_1, \dots, X_d)$ be a random vector with distribution function F. Let F_i be the (continuous) marginal distribution function of X_i, $i = 1, \dots, d$. By Sklar's Theorem [17], there exists a unique function C such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)). \tag{1}$$

This function C is called the copula of F and is the d-dimensional distribution function of the random vector $(F_1(X_1), \dots, F_d(X_d))$. For a general account on copulas, see, e.g. [16]. Copulas are interesting since they permit to impose a dependence structure on pre-determined marginal distributions.

While there exist many copulas in the bivariate case, it is less clear how to construct copulas in higher dimension. In the presence of non-Gaussianity and/or tail dependence, various constructions have been adopted, such as, for instance, Archimedean copulas [9], Vines [1] or elliptical copulas [5]. Because Archimedean copulas possess only a few parameters, they lack flexibility in high-dimension. Vines, on the opposite, achieve greater flexibility but at the price of increased complexity in the modeling process. The use of elliptical copulas

goes together with assuming a similar dependence pattern among all pairs of variables. This may be undesirable in applications. Moreover, they have in general as many as $O(d^2)$ parameters and it is difficult to carry out maximum likelihood inference [3].

Another approach [14] aims at constructing a multivariate copula as a product of transformed bivariate copulas. This approach possesses several advantages. A probabilistic interpretation is available and thus the generation of random vectors is straightforward. The resulting copula is explicit, leading to explicit bounds on dependence coefficients of the bivariate marginals. The class of copulas which can be constructed from this approach is large and can cover a wide range of dependencies. Finally the analysis of extreme values can be performed by constructing extreme-value copulas.

However, although many copulas with different features can be built, the use of this approach for practical applications remains challenging. Indeed, two pitfalls render inference difficult: first, they are constraints on the parameters, and second, the product form complicates the computation of the density – hence, of the potential likelihood – even numerically.

The main contribution of this paper is to revisit the product of transformed copulas in order to propose a new multivariate copula model of practical interest. First, there are no constraints on the parameters anymore. Moreover, a graphical structure associated to the copulas within this class permits to visualize the dependencies and to efficiently compute the likelihood, even in high dimension.

The rest of this paper is organized as follows. Section 2 reviews the product of transformed copulas and important properties such as random generation and the ability to construct extreme-value models. Section 3 presents the new copula model and enlightens the link with the product of transformed copulas. Section 4 discusses the dependence properties of bivariate marginals of the proposed class by providing bounds on some of the most popular dependence coefficients such as the Spearman's rho, Kendall's tau, and tail dependence coefficients. In Section 5, we apply the proposed copula model to a simulated and a real dataset. The appendix gathers the proofs of this paper.

2 Product of transformed copulas

It is easily seen that a product of copulas is not a copula in general. Nonetheless the next theorem due to Liebscher [14] shows that, up to marginal transformations, a product of copulas can lead to a well defined copula.

Theorem 1. *Assume* $\tilde{C}_1, \ldots, \tilde{C}_K : [0,1]^d \to [0,1]$ *are copulas. Let* $g_{ei} : [0,1] \to [0,1]$ *for* $e = 1, \ldots, K$, $i = 1, \ldots, d$ *be functions with the property that each of them is strictly increasing or is identically equal to* 1. *Suppose that* $\prod_{e=1}^{K} g_{ei}(v) = v$ *for* $v \in [0,1], i = 1, \ldots, d$, *and* $\lim_{v \to 0} g_{ei}(v) = 0$ *for* $e = 1, \ldots, K$, $i = 1, \ldots, d$. *Then*

$$C(u_1, \ldots, u_d) = \prod_{e=1}^{K} \tilde{C}_e \left(g_{e1}(u_1), \ldots, g_{ed}(u_d) \right) \tag{2}$$

is also a copula.

The probabilistic interpretation of (2) is as follows. Let

$$(U_1^{(1)}, \ldots, U_d^{(1)}), \ldots, (U_1^{(K)}, \ldots, U_d^{(K)})$$

be K independent random vectors having distribution function $\tilde{C}_1, \ldots, \tilde{C}_K$ respectively. Let g_{ei}, $e = 1, \ldots, K$, $i = 1, \ldots, d$ be as in Theorem 1 and define $g_{ei}^{-1}(v) := 0$ for $v \leq g_{ei}(0)$ and $J_i = \{e \in \{1, \ldots, K\} : g_{ei} \neq 1\}$. Then C is the joint distribution function of the random vector

$$\left(\max_{e \in J_1} g_{e1}^{-1}(U_1^{(e)}), \ldots, \max_{e \in J_d} g_{ed}^{-1}(U_d^{(e)}) \right). \tag{3}$$

If there exists a random generation procedure for $\tilde{C}_e$, $e = 1, \ldots, K$ then thanks to (3) a random generation procedure for C can be derived as well.

The statistical analysis of extreme values should theoretically be carried out with the help of extreme-value copulas [8]. Recall that a copula $C_\#$ is an extreme-value copula if there exists a copula C such that

$$C_\#(u_1, \ldots, u_d) = \lim_{n \uparrow \infty} C^n(u_1^{1/n}, \ldots, u_d^{1/n}), \tag{4}$$

for every $(u_1, \ldots, u_d) \in [0,1]^d$. A copula $C_\#$ is said to be max-stable if for every integer $n \geq 1$ and every $(u_1, \ldots, u_d) \in [0,1]^d$

$$C_\#^n(u_1^{1/n}, \ldots, u_d^{1/n}) = C_\#(u_1, \ldots, u_d).$$

Extreme-value copulas correspond exactly to max-stable copulas [8]. Theorem 1 can be used to construct extreme-value copulas as shown in the next proposition due to [4].

Proposition 1. *In (2), let $g_{ei}(v) = v^{\theta_{ei}}$, $v \in [0,1]$ with $\theta_{ei} \in [0,1]$ and $\sum_{e=1}^K \theta_{ei} = 1$ for $i = 1, \ldots, d$. If $\tilde{C}_e$, $e = 1, \ldots, K$ is max-stable then so is C.*

Out of the context of extreme values, applications of Theorem 1 can be found, for instance, in the analysis of directional dependence [13] ($K = d = 2$), finance [2] ($K = d = 2$) and hydrology [4] ($K = 2$, $d = 3$).

We are not aware of applications of Theorem 1 in practice when $K > 2$ or $d > 3$. As pointed out in the introduction, the product form (2) renders the density $\frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d}$, hence the likelihood, complicated to compute even numerically. Furthermore, the constraints $\prod_{e=1}^K g_{ei}(v) = v$, $v \in [0,1]$, $i = 1, \ldots, d$ in Theorem 1 are not easy to deal with in practice.

The next section aims at overcoming these drawbacks.

3 Product of transformed copulas revisited

The product over $e \in \{1, \ldots, K\}$ in (2) can be taken over $e \in E$, where E is an arbitrary finite set, yielding

$$C(u_1, \ldots, u_d) = \prod_{e \in E} \tilde{C}_e \left(g_{e1}(u_1), \ldots, g_{ed}(u_d) \right). \tag{5}$$

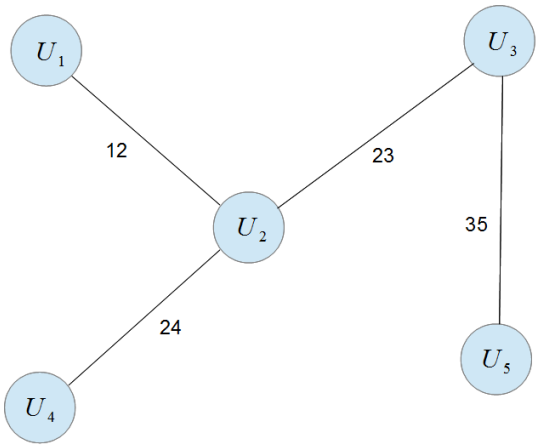

Figure 1: Graphical representation of the set $E = \{\{12\}, \{24\}, \{23\}, \{35\}\}$. $N(1) = \{\{12\}\}$, $N(2) = \{\{12\}, \{23\}, \{24\}\}$, $N(3) = \{\{23\}, \{35\}\}$, $N(4) = \{\{24\}\}$ and $N(5) = \{\{35\}\}$.

In particular, an element $e \in E$ can represent a pair of the variables at play. More precisely, let $U_1, \ldots, U_d$ be d standard uniform random variables. Denote by $\{ij\}$ the index of the pair (U_i, U_j) and let $E \subset \{\{ij\} : i, j = 1, \ldots, d,\ j > i\}$ be a subset of the set of the pair indices. The cardinal of E, denoted by $|E|$, is less or equal to $d(d-1)/2$. The pair index $e \in E$ is said to contain the variable index i if $e = \{ik\}$ for $k \neq i$. Let us introduce $N(i) = \{e \in E : e \text{ contains } i\}$. $N(i)$ is called the set of neighbors of i and has cardinal $|N(i)| = n_i$. It is natural to associate a graph to the set E as follows: an element $e = \{ij\} \in E$ is an edge linking U_i and U_j in the graph whose nodes are the variables $U_1, \ldots, U_d$. The example $E = \{\{12\}, \{24\}, \{23\}, \{35\}\}$ is illustrated in Figure 1. For $u = (u_1, \ldots, u_d) \in [0,1]^d$, consider the functional

$$C(u_1, \ldots, u_d) = \prod_{\{ij\} \in E} \tilde{C}_{ij}(u_i^{1/n_i}, u_j^{1/n_j}), \tag{6}$$

where the $\tilde{C}_{ij}$'s are bivariate copulas. Keeping in mind the graphical representation, C in (6) is a product over the edges. For instance, when $E = \{\{12\}, \{24\}, \{23\}, \{35\}\}$ as in Figure 1, (6) writes

$$C(u_1, u_2, u_3, u_4, u_5) = \tilde{C}_{12}(u_1, u_2^{1/3}) \tilde{C}_{24}(u_2^{1/3}, u_4) \tilde{C}_{23}(u_2^{1/3}, u_3^{1/2}) \tilde{C}_{35}(u_3^{1/2}, u_5).$$

In the following, (6) is referred to as the Product of Bivariate Copulas (PBC) copula, or PBC model. The next theorem establishes that (6) is a copula and makes the link with Theorem 1.

Theorem 2. *If in* (5)*:*

(i) for $e = \{ij\} \in E$, $\tilde{C}_e$ takes exactly two arguments non identically equal to one, namely, g_{ei} and g_{ej}, and

(ii) for $i = 1, \ldots, d$ and $e \in N(i)$, g_{ei} does not depend on e;

then the only copula which can be constructed from (5) *is the PBC model* (6)*, where* $\tilde{C}_{ij}$ *is defined by*

$$\tilde{C}_{ij}(u,v) = \tilde{C}_{\{ij\}}(1,\dots,1,u,1,\dots,1,v,1,\dots,1), \quad (u,v) \in [0,1]^2,$$

and where in $(1,\dots,1,u,1,\dots,1,v,1,\dots,1)$*,* u *and* v *are at the* i*-th and* j*-th positions respectively.*

Condition (i) in Theorem 2 simply means that only bivariate copulas are allowed in the construction. The simplification (ii) achieves two goals: first to reduce the number of parameters (an important feature in high-dimension), and second to intrinsically satisfy the constraints $\prod_{e\in E} g_{ei}(v) = v$, $v \in [0,1]$, $i = 1,\dots,d$ in the assumptions of Theorem 1. If assumption (ii) in Theorem 2 was not made, one could take $g_{ei}(v) = v^{\theta_{ei}}, e \in E$, $i = 1,\dots,d$, $\theta_{ei} \in [0,1]$ with the constraints

$$\sum_{e\in N(i)} \theta_{ei} = \sum_{k:\{ki\}\in E} \theta_{ki,i} = 1, \quad i = 1,\dots,d. \tag{7}$$

These constraints would be difficult to handle in practice, and, furthermore, the number of parameters would increase quadratically with the dimension. Indeed, one would have $(|E|-1)d$ parameters θ_{ei} plus an additional number $|E|$ of parameters for each copula $\tilde{C}_e$. If the graph associated to E is a tree, for instance, then $|E| = d-1$, yielding $O(d^2)$ parameters. As a comparison, in the PBC model (6), there are no constraints and only $O(d)$ parameters in total.

From (1), the PBC copula (6) is associated to a distribution function F with continuous marginals F_i, $i = 1,\dots,d$, such that

$$F(x_1,\dots,x_d) = C(F_1(x_1),\dots,F_d(x_d)), \quad (x_1,\dots,x_d) \in \mathbb{R}^d. \tag{8}$$

By substituting (6) into (8), it is easy to see that F writes

$$F(x_1,\dots,x_d) = \prod_{\{ij\}\in E} F_{ij}(x_i,x_j), \quad (x_1,\dots,x_d) \in \mathbb{R}^d, \tag{9}$$

where F_{ij}, $\{ij\} \in E$, is a bivariate distribution function such that the first (respectively the second) marginal $F_{ij,1}$ (respectively $F_{ij,2}$) only depends on i (respectively j). It is interesting to note that the converse is also true as stated in the following proposition.

Proposition 2. *The distribution function corresponding to the PBC copula* (6) *writes as* F *in* (9)*. Conversely, the copula corresponding to the distribution function* F *in* (9) *writes as the PBC copula* (6)*.*

4 Dependence properties and max-stability

Let C be the PBC copula (6). First the dependence properties of a pair (U_k, U_l) whose copula is the bivariate copula $C_{kl}(u_k,u_l) = C(1,\dots,1,u_k,1,\dots,1,u_l,1,\dots,1)$ are studied. The conditions under which the PBC model (6) is an extreme-value copula are given afterwards.

Proposition 3. *The bivariate marginal* C_{kl} *is given by*

$$C_{kl}(u_k,u_l) = \begin{cases} u_k^{(n_k-1)/n_k} u_l^{(n_l-1)/n_l} \tilde{C}_{kl}(u_k^{1/n_k}, u_l^{1/n_l}) & \text{if } \{kl\} \in E, \\ u_k u_l & \text{otherwise.} \end{cases} \tag{10}$$

Example 1. *If in* (10) $\tilde{C}_{kl}$ *is a Marshall-Olkin copula (see for instance [16], p.53) with parameters* $0 \leq \alpha, \beta \leq 1$ *(denoted by MO*(α, β)*), that is,*

$$\tilde{C}_{kl}(u_k, u_l) = \min(u_k^{1-\alpha} u_l, u_l^{1-\beta} u_k),$$

then C_{kl} *is MO*$(\alpha/n_k, \beta/n_l)$*. If* $\alpha = \beta$ *then* $\tilde{C}_{kl}$ *is a Cuadras-Augé copula and* C_{kl} *is MO*$(\alpha/n_k, \alpha/n_l)$*. If* $\alpha = \beta = 0$ *then both* $\tilde{C}_{kl}$ *and* C_{kl} *are the independence copula. If* $\alpha = \beta = 1$ *then* $\tilde{C}_{kl}$ *is the Fréchet upper bound copula and* C_{kl} *is MO*$(1/n_k, 1/n_l)$*.*

Remark 1. *If in* (10) *one puts* $\kappa = 1/n_k$ *and* $\lambda = 1/n_l$*, then the copulas take the form* $C_{kl}(u_k, u_l) = u_k^{1-\kappa} u_l^{1-\lambda} \tilde{C}_{kl}(u_k^{\kappa}, u_l^{\lambda})$*. This class of copulas, sometimes referred to as Khoudraji copulas, was proposed in [6] Proposition 2.*

Let (U, V) be a random vector with copula C. The dependence between U and V is positive if, roughly speaking, U and V tend to be large or small together. Below are recalled a few definitions of statistical concepts about positive dependence. The copula C has the TP2 (totally positive of order 2) property if and only if

$$C(u_1, u_2)C(v_1, v_2) \geq C(u_1, v_2)C(v_1, u_2), \text{ for all } u_1 < v_1 \text{ and } u_2 < v_2. \quad (11)$$

Also, C is said to be PQD (positive quadrant dependent) if $C(u, v) \geq uv$ for all $(u, v) \in [0, 1]^2$. The random variable V is said to be LTD (left tail decreasing) in U if for all $v \in [0, 1]$, the function $u \mapsto P(V \leq v | U \leq u)$ is decreasing in u. The dependence between U and V can be quantified through dependence measures such as the Kendall's tau or the Spearman's rho respectively given by

$$\tau = 4 \int_{[0,1]^2} C(u, v)\, dC(u, v) - 1, \quad (12)$$

$$\rho = 12 \int_{[0,1]^2} C(u, v)\, du\, dv - 3. \quad (13)$$

The dependence in the upper and lower tails can be respectively measured with

$$\lambda^{(U)} = \lim_{u \uparrow 1} \frac{1 - 2u + C(u, u)}{1 - u} \in [0, 1], \quad \lambda^{(L)} = \lim_{u \downarrow 0} \frac{C(u, u)}{u} \in [0, 1].$$

See [16] and [12] for further details about these concepts. Let us denote by τ_{kl}, ρ_{kl}, $\lambda_{kl}^{(U)}$ and $\lambda_{kl}^{(L)}$ the Kendall's tau, Spearman's rho, upper tail dependence coefficient and lower tail dependence coefficient of the copula C_{kl} in (10) respectively. As shown in Proposition 3, C_{kl} is a bivariate marginal of the PBC copula (6) and one may apply the results of [14] to obtain the following.

Proposition 4. *If in* (10) $\tilde{C}_{kl}$ *is TP2, LTD or PQD then* C_{kl} *is also TP2, LTD or PQD respectively.*

Explicit bounds in terms of the number of neighbors for the dependence coefficients of PBC bivariate marginals are given in the next proposition. The behavior of (10) when the number of neighbors tends to infinity is also studied.

Proposition 5. *We have* $\lambda_{kl}^{(L)} = 0$ *and* $\lambda_{kl}^{(U)} \leq \min(1/n_k, 1/n_l)$. *The lower and upper bounds for* ρ_{kl} *and* τ_{kl} *are respectively given by*

$$a_\rho(n_k, n_l) \leq \rho_{kl} \leq b_\rho(n_k, n_l),$$
$$a_\tau(n_k, n_l) \leq \tau_{kl} \leq b_\tau(n_k, n_l),$$

with

$$a_\rho(n_k, n_l) = \frac{6\beta(2n_k - 1, 2n_l - 1)n_k n_l}{(2n_k + 2n_l - 1)(n_k + n_l - 1)} - \frac{3}{(2n_k - 1)(2n_l - 1)},$$
$$b_\rho(n_k, n_l) = \frac{3}{2n_k + 2n_l - 1},$$
$$a_\tau(n_k, n_l) = \frac{\beta(2n_l - 1, 2n_k - 1)}{n_k + n_l - 1} - \frac{2}{(2n_k - 1)(2n_l - 1)},$$
$$b_\tau(n_k, n_l) = \frac{1}{n_k + n_l - 1},$$

where β *denotes the* β*-function,* $\beta(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$. *Furthermore, as* $\max(n_k, n_l) \to \infty$, *we have* $C_{kl}(u, v) \to uv$ *for all* $(u, v) \in [0, 1]^2$.

The above results show that we are facing a tradeoff: on the one hand, the larger the cardinal of E (or the more connected the graph associated to E), the less the pairs in E are able to model strong dependencies. On the other hand, the smaller the cardinal of E, the more there are independent pairs (since there are less pairs in E). To illustrate Proposition 5, numerical values of the bounds are computed in Table 1 for different numbers of neighbors (n_k, n_l).

coefficient	ρ_{kl}	τ_{kl}	λ_{kl}
(n_k, n_l)			
$(1, 2)$	$[-0.60, 0.60]$	$[-0.50, 0.50]$	$[0.00, 0.50]$
$(2, 2)$	$[-0.30, 0.43]$	$[-0.21, 0.33]$	$[0.00, 0.50]$
$(1, 3)$	$[-0.43, 0.43]$	$[-0.33, 0.33]$	$[0.00, 0.33]$
$(2, 3)$	$[-0.19, 0.33]$	$[-0.13, 0.25]$	$[0.00, 0.33]$
$(3, 3)$	$[-0.12, 0.27]$	$[-0.08, 0.20]$	$[0.00, 0.33]$

Table 1: Lower and upper bounds [lower, upper] for the Spearman'rho ρ_{kl}, Kendall's tau τ_{kl} and upper tail dependence coefficient λ_{kl} depending on the number of neighbors (n_k, n_l).

Finally, it is easy to construct extreme-value copulas belonging to the PBC class (6). Indeed, the following result follows from Proposition 1.

Proposition 6. *If in the PBC copula* (6), $\tilde{C}_{ij}$ *is an extreme-value copula for* $\{ij\} \in E$, *then* C *is also an extreme-value copula.*

All copulas C_{kl} in Example 1 are max-stable since Marshall-Olkin copulas are max-stable. Thus the associated PBC is an extreme-value copula. If $\tilde{C}_{kl}$ in (10) is a (max-stable) Gumbel copula, that is,

$$\tilde{C}_{kl}(u_k, u_l) = \exp - \left[(-\log u_k)^\theta + (-\log u_l)^\theta\right]^{1/\theta}, \quad \theta \geq 1, \tag{14}$$

then C_{kl} is also max-stable, hence, the PBC is an extreme-value copula.

5 Numerical applications to simulated and real datasets

In this section, PBC copulas models are applied to simulated and real datasets. The methods used to simulate and infer the copulas are presented in Section 5.1. The considered families for the bivariate copulas $\tilde{C}_{ij}$ in (6) are the following: the Ali-Mikhail-Haq (AMH), Farlie-Gumbel-Morgenstern (FGM), Frank, Gumbel, and Joe families. See [16] or [12] for details about these families. The corresponding PBC copula models (6) are therefore referred to as PBC AMH, PBC FGM, PBC Frank, PBC Gumbel and PBC Joe respectively. In Section 5.2, the two inference procedures presented in Section 5.1 are compared. Section 5.3 applies PBC models to an hydrological dataset.

5.1 Computational aspects

In this section, we assume that the copulas $\tilde{C}_{ij}$ of the PBC model (6) depend on parameters θ_{ij}'s and that we are given a sample of i.i.d data vectors from (6).

Data simulation from a PBC copula is straightforward thanks to the probabilistic interpretation given in (3). The generation procedure is given below.

- For all $\{ij\} \in E$, generate $(U_i^{(ij)}, U_j^{(ij)}) \sim \tilde{C}_{ij}$.
- For all $i = 1, \ldots, d$, compute $U_i = \max_{k \in \{1,\ldots,d\}:\{ki\} \in E} \left\{ \left(U_i^{(ki)} \right)^{n_i} \right\}$.

The resulting vector $(U_1, \ldots, U_d)$ has distribution (6).

The inference of PBC copulas can be performed by maximum-likelihood based methods. As it is well known, the estimators resulting from these methods have the advantage to be consistent and asymptotically unbiased under mild conditions. Properly scaled, their asymptotic distribution is Gaussian and confidence intervals or tests can be derived.

The first considered approach is the pairwise maximum-likelihood method [15]. This approach consists in maximizing the sum of the likelihoods corresponding to all the pairs of variables. In our case, it simplifies to maximizing $|E|$ univariate functions independently. However, unlike the full joint maximum likelihood estimator, the pairwise maximum-likelihood estimator is not guaranteed to be efficient.

The second considered approach is the standard full joint maximum-likelihood method. Indeed, it is possible to compute the full joint likelihood of a PBC copula when the graph associated to it is a tree thanks to a message-passing algorithm [11]. A brief explanation of how this algorithm works is given in Appendix B. The reader is referred to [11] for the complete algorithm and [10] for a detailed explanation. We implemented this algorithm in the `R` package `PBC` [18].

5.2 A simulation experiment to compare pairwise likelihood and full joint likelihood approaches

We generated 100 datasets of dimension $d = 9$ and size $n = 500$ according to a PBC copula whose tree graph is given in Figure 2. The amount of time required

to maximize the full joint likelihood for one dataset replication was 36, 21, 18, 21, and 21 seconds for PBC AMH, PBC FGM, PBC Frank, PBC Gumbel, and PBC Joe respectively with a 8 GiB memory and 3.20 GHz processor computer. The $d-1=8$ coordinates θ_i of the parameter vectors were chosen to be regularly spaced within the intervals $[-0.9, 0.9]$, $[-0.9, 0.9]$, $[-9, 11]$, $[2, 20]$ and $[1, 20]$ respectively.

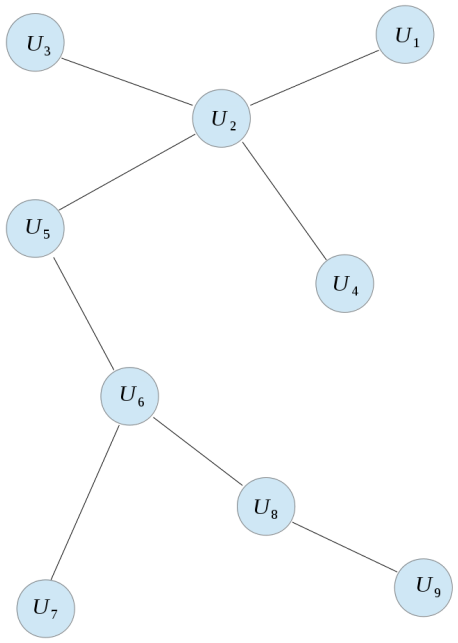

Figure 2: Tree graph associated to the simulated PBC copulas.

The following criteria were calculated in order to assess the results of the experiment. The variance ratio (VR) is defined as

$$VR = \frac{\sum_{e=1}^{d-1} \widehat{\text{Var}}\left(\hat{\theta}_e^{FULL}\right)}{\sum_{e=1}^{d-1} \widehat{\text{Var}}\left(\hat{\theta}_e^{PW}\right)},$$

where $\hat{\theta}_e^{FULL}$, $\hat{\theta}_e^{PW}$ is the coordinate estimated by maximization of the full joint likelihood, pairwise likelihood, respectively, and where $\widehat{\text{Var}}$ is the empirical variance on the replications. For each dataset replication, the mean absolute error on the Spearman's rho ρ (MAE_ρ) and the Kendall's tau τ (MAE_τ) is defined as

$$\text{MAE}_\rho = \frac{1}{d-1}\sum_{e=1}^{d-1} |\rho(\theta_e) - \rho(\hat{\theta}_e^{FULL})|, \quad \text{MAE}_\tau = \frac{1}{d-1}\sum_{e=1}^{d-1} |\tau(\theta_e) - \tau(\hat{\theta}_e^{FULL})|.$$

The MAEs were averaged over the 100 replications to get a single value per model.

Copula	VR	MAE_ρ	MAE_τ
PBC AMH	0.96	0.03	0.02
PBC FGM	0.98	0.03	0.02
PBC Frank	0.79	0.02	0.01
PBC Gumbel	0.68	0.00	0.00
PBC Joe	0.71	0.00	0.00

Table 2: Variance ratio (VR) and mean absolute errors (MAEs) for each of the tested PBC models. The MAEs were averaged over the dataset replications.

The results are reported in Table 2. It appears that for PBC AMH and PBC FGM the precision was not improved by maximizing the full joint likelihood relative to the pairwise approach: the VR for those models are close to 1. For the Frank, Gumbel and Joe families, however, the variance decreases by at least 20% in average. These families, in contrast to the AMH and FGM families, are comprehensive, meaning that they include the lower and upper bounds for copulas. The MAEs are quite low for all the models. This indicates that the maximization of the full joint likelihood with the message-passing algorithm of Appendix B performs well.

5.3 Application to an hydrological dataset

In this section, PBC copula models are applied to an hydrological dataset consisting of $d = 3$ stations and $n = 36$ observations of flow rate annual maxima. The sites are located on three french rivers at the following places: La Celle-en-Morvan on the river la Selle (S), Rigny-sur-Arroux on l'Arroux (A), and Isclades-et-Rieutord on la Loire (L). These rivers are embedded in the sense that Selle flows into Arroux which flows into Loire. Thus, the graph is naturally set up as

$$\text{S} - \text{A} - \text{L}.$$

The same models as in Section 5.2 were tested, that is, PBC AMH, PBC FGM, PBC Frank, PBC Gumbel, and PBC Joe. The Gumbel copula was also considered here as a benchmark. This family is standard in hydrology for fitting trivariate distributions [19]. The estimation of the parameters was performed by maximization of the full joint likelihood, as explained in Section 5.1. In order to assess the fit of the models, the empirical Spearman's rho and Kendall's tau coefficient estimates were compared to their counterpart under the models. Since the number of parameters is the same for all models, the likelihood values for the different models were also compared. The results are reported in Table 3.

One can observe that PBC AMH and PBC FGM perform very poorly compared to the other models. This was expected since the AMH and FGM families are not comprehensive, roughly meaning that they do not allow much dependence (see, e.g., [16]). The standard Gumbel copula performs poorly too, with one of the smallest log-likelihood values. One can also see that, since it has a single parameter, the dependence coefficients between the different pairs are

	$\rho_{S,A}$	$\rho_{A,L}$	$\rho_{S,L}$	$\tau_{S,A}$	$\tau_{A,L}$	$\tau_{S,L}$	log-likelihood
empirical data	0.70	0.30	0.13	0.5	0.21	0.08	
PBC AMH	0.25	0.25	0	0.17	0.17	0	7.05
PBC FGM	0.20	0.20	0	0.13	0.13	0	5.46
PBC Frank	0.44	0.30	0	0.30	0.21	0	9.09
PBC Gumbel	0.43	0.31	0	0.30	0.21	0	9.20
PBC Joe	0.41	0.27	0	0.29	0.18	0	8.05
Gumbel	0.39	0.39	0.39	0.27	0.27	0.27	6.16

Table 3: Optimized log-likelihood and pairwise dependence coefficients for the empirical data and the tested PBC copulas. The symbol ρ and τ stand for the Spearman's rho and Kendall's tau respectively. For instance, $\rho_{S,A}$ is the Spearman's rho coefficient between the variables S and A.

equal to each other. The PBC copulas with comprehensive families present a much better fit. The dependence coefficient with the smallest value, that of the pair (A,L), is very well approximated by the PBC Frank, PBC Gumbel, and PBC Joe. In particular, PBC Frank and PBC Gumbel both provide, for instance, a Kendall's tau of 0.21, which is the same as the empirical value. Also, these copulas possess the highest log-likelihood values, 9.20 and 9.09, a step above the third highest, 8.05. The dependence coefficient of the pair (S,A), which presents more dependence (0.7 for the Spearman's rho and 0.5 for the Kendall's tau) is underestimated. Although the theoretical upper bound for the Kendall's tau is 0.5 (see Table 1), the closest copulas are PBC Frank and PBC Gumbel with a Kendall's tau of 0.3 for both. Given that the third pair (S,L) presents low values for the Spearman's rho (0.13) and the Kendall's tau (0.08), its distribution might by approximated by the independence copula, as PBC models do. The test for independence proposed in [7], implemented in the `R` package `copula`, gave a p-value of 0.41. The Gumbel copula, instead, seems to overestimate the dependence in the third pair (S,L).

6 Discussion

In this paper, we have constructed a class of multivariate copulas, called PBC copulas, based on bivariate copulas. Therefore, this novel class benefits from the many bivariate families existing in the literature. No constraints on the parameters refrain the applicability of the PBC class and a natural graph structure helps to visualize the dependencies between the variables. Full joint multivariate inference can be performed, and shown to perform well, with the message-passing algorithm presented in the appendix. However, PBC copula models still suffer from weaknesses. First, the more there are edges in the graph, the more the bounds on the dependence coefficients are restrictive. Second, it was shown numerically that dependence coefficients of high magnitude were prone to be underestimated. In view of these remarks, it may be advisable to keep the number of neighbors in the graph associated to PBC models as low as possible, and to be careful with highly dependent data.

Acknowledgment. The authors thank "Banque HYDRO du Ministère de l'Écologie, du Développement durable et de l'Énergie" for providing the data and Benjamin Renard for fruitful dicussions about statistical issues in hydrological science.

Appendix

A Proofs

Proof of Theorem 2

From Theorem 1, it is straightforward to see that (6) is a copula. Let us now prove that (6) is the only copula arising from (5). Condition (i) implies that if $e \notin N(i)$ then $g_{ei} = 1$, $i = 1, \ldots, d$. Hence, the constraint over the functions reduces to $\prod_{e \in N(i)} g_{ei}(v) = v$, $v \in [0,1]$. In view of condition (ii), one has $g_{ei} = g_i$ for $e \in N(i)$, hence $(g_i(v))^{n_i} = v$. Therefore

$$g_{ei}(v) = \begin{cases} v^{1/n_i} & \text{if } e \in N(i), \\ 1 & \text{otherwise.} \end{cases}$$

To conclude it suffices to rewrite the product in (5) as

$$\prod_{e \in E} \tilde{C}_e(1, \ldots, 1, u_i^{1/n_i}, 1, \ldots, 1, u_j^{1/n_j}, 1, \ldots, 1) = \prod_{\{ij\} \in E} \tilde{C}_{ij}(u_i^{1/n_i}, u_j^{1/n_j})$$

which corresponds to (6).

Proof of Proposition 2

Let us first prove that (9) is the distribution function of (6). By (1) we have

$$\begin{aligned} F(x_1, \ldots, x_d) =& C(F_1(x_1), \ldots, F_d(x_d)) \\ =& \prod_{\{ij\} \in E} \tilde{C}_{ij}(F_i(x_i)^{1/n_i}, F_j(x_j)^{1/n_j}) \\ =:& \prod_{\{ij\} \in E} \Phi_{ij}(x_i, x_j). \end{aligned}$$

The first margin of Φ_{ij} is given by $\Phi_{ij,1}(x) = \Phi_{ij}(x, \infty) = F_i(x_i)^{1/n_i}$ which depends only on i. The same holds for the second margin $\Phi_{ij,2}$.

Let us prove that (6) is the copula associated to (9). Let $\Phi_{ij,k}$, $k = 1, 2$ be the k-th univariate marginal of $\Phi_{ij}, \{ij\} \in E$. The copula associated to F is given by

$$C_F(u_1, \ldots, u_d) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right) = \prod_{\{ij\} \in E} \Phi_{ij}\left(F_i^{-1}(u_i), F_j^{-1}(u_j)\right).$$

Let $\tilde{C}_{ij}$ be the copula associated to Φ_{ij}. We have

$$\Phi_{ij}(x_i, x_j) = \tilde{C}_{ij}\left(\Phi_{ij,1}(x_i), \Phi_{ij,2}(x_j)\right)$$

so that $\Phi_{ij}\left(F_i^{-1}(u_i), F_j^{-1}(u_j)\right) = \tilde{C}_{ij}\left(\Phi_{ij,1} \circ F_i^{-1}(u_i), \Phi_{ij,2} \circ F_j^{-1}(u_j)\right)$, and

$$C_F(u_1, \ldots, u_d) = \prod_{\{ij\} \in E} \tilde{C}_{ij}\left(\Phi_{ij,1} \circ F_i^{-1}(u_i), \Phi_{ij,2} \circ F_j^{-1}(u_j)\right). \qquad (15)$$

Moreover, since $\tilde{C}_F$ is a copula, it follows

$$\begin{aligned} u_k =& C_F(1,\dots,1,u_k,1,\dots,1) \\ =& \prod_{j>k:\{kj\}\in E} \tilde{C}_{kj}\left(\Phi_{kj,1}\circ F_k^{-1}(u_k),1\right) \prod_{j<k:\{jk\}\in E} \tilde{C}_{jk}\left(1,\Phi_{jk,2}\circ F_k^{-1}(u_k)\right) \\ =& \prod_{j:\{kj\}\in E} \Phi_{kj,1}\circ F_k^{-1}(u_k). \end{aligned}$$

Now by assumption $\Phi_{kj,1} = \Phi_{jk,2} = \Phi_k$ only depends on k and therefore $u_k^{1/n_k} = \Phi_k \circ F_k^{-1}(u_k)$ which implies $\Phi_k(z) = F_k(z)^{1/n_k}$, $z \in \mathbb{R}$. By pluging Φ_k into (15) the result follows.

Proof of Proposition 3

If $\{kl\} \in E$, then

$$\begin{aligned} C_{kl}(u_k,u_l) =& C(1,\dots,1,u_k,1,\dots,1,u_l,1,\dots,1) \\ =& \left(\prod_{e\in N(k)\setminus\{kl\}} \tilde{C}_e(u_k^{1/n_k},1)\right)\left(\prod_{e\in N(l)\setminus\{kl\}} \tilde{C}_e(u_l^{1/n_l},1)\right)\times \\ & \tilde{C}_{kl}(u_k^{1/n_k},u_l^{1/n_l}) \\ =& u_k^{(n_k-1)/n_k} u_l^{(n_l-1)/n_l} \tilde{C}_{kl}(u_k^{1/n_k},u_l^{1/n_l}). \end{aligned}$$

Otherwise,

$$\begin{aligned} C_{kl}(u_k,u_l) &= \left(\prod_{e\in N(k)} \tilde{C}_e(u_k^{1/n_k},1)\right)\left(\prod_{e\in N(l)} \tilde{C}_e(u_l^{1/n_l},1)\right) \\ &= u_k^{n_k/n_k} u_l^{n_l/n_l} \\ &= u_k u_l. \end{aligned}$$

Proof of Proposition 5

The Fréchet-Hoeffding bounds for copulas (see e.g. [16], p. 11) applied to $\tilde{C}_{kl}$ in (10) yield

$$W_{kl}(u_k,u_l) \le C_{kl}(u_k,u_l) \le M_{kl}(u_k,u_l), \tag{16}$$

where

$$\begin{aligned} W_{kl}(u_k,u_l) &= u_k^{1-1/n_k} u_l^{1-1/n_l} \max(u_k^{1/n_k} + u_l^{1/n_l} - 1, 0), \\ M_{kl}(u_k,u_l) &= u_k^{1-1/n_k} u_l^{1-1/n_l} \min(u_k^{1/n_k}, u_l^{1/n_l}). \end{aligned}$$

We have $M_{kl}(u,u)/u \to 0$ as $u \downarrow 0$. It is easily seen that $W_{kl}(u,u)/u \to 0$ as $u \downarrow 0$ which implies $C_{kl}(u,u)/u \to 0$. It is straightforward to see that $(1-2u+M_{kl}(u,u))/(1-u) \to 1/\max(n_k,n_l)$ as $u \uparrow 1$. To compute the lower and upper bounds for ρ_{kl} and τ_{kl}, it suffices to substitute W_{kl} and M_{kl} into (13) and (12). Lengthy but elementary computations lead to the results. Finally, letting n_k or n_l going to infinity in (16) yields that C_{kl} tends to independence.

B Algorithm to compute the full joint likelihood of PBC copulas

Denote the parameter vector by $\boldsymbol{\theta} = (\theta_{ij})_{\{ij\}\in E}$. Recall that the graph is assumed to be a tree, that is, there is no cycles in the graph (then $|E| = d-1$). Let $V = \{1, \dots, d\}$ and $u = (u_1, \dots, u_d)$ a vector in $[0,1]^d$. For a subset $A \subset V$, the notation $\partial_{u_A} C(u; \boldsymbol{\theta})$ stands for the derivative of C with respect to all the variables in A. For instance the density (hence the likelihood) writes

$$\frac{\partial^d C(u;\boldsymbol{\theta})}{\partial_{u_1} \dots \partial_{u_d}} = \partial_{u_V} C(u;\boldsymbol{\theta}) = c(u;\boldsymbol{\theta}), \tag{17}$$

and the gradient with respect to the parameter vector,

$$\left(\frac{\partial c(u;\boldsymbol{\theta})}{\theta_{ij}}\right)_{\{ij\in E\}}.$$

To keep the notation simple, the dependence on the parameter vector $\boldsymbol{\theta}$ is dropped in the remaining of the section. The purpose here is not to give the algorithm, but rather to provide an intuitive idea of it.

Let us write

$$C(u_1, \dots, u_d) = \prod_{\{ij\}\in E} \tilde{C}_{ij}(u_i^{1/n_i}, u_j^{1/n_j}) =: \prod_{\{ij\}\in E} \Phi_{ij}(u_i, u_j).$$

and let an arbitrary variable index i (the root) be given. Let τ_s^i denote the subtree rooted at the variable indexed by i and containing the edge indexed by e (see Figure 3). The idea is to note that, since the graph is a tree, the copula C can be decomposed over the subtrees rooted at i:

$$C(u) = \prod_{e\in E} \Phi_e(u) =: \prod_{e\in N(i)} T_{\tau_e^i}(u), \quad u = (u_1, \dots, u_d),$$

where $T_{\tau_e^i}(u)$ corresponds to the product of all edges located in the subtree τ_e^i. Since the $T_{\tau_e^i}(u)$'s do not share any variables (except the root), the derivative and the product operations commute, more precisely,

$$\begin{aligned}
\partial_{u_V} C(u) &= \partial_{u_i, u_{V\setminus i}} \left[\prod_{e\in N(i)} T_{\tau_e^i}(u)\right] \\
&= \partial_{u_i} \left[\prod_{e\in N(i)} \partial_{u_{\tau_e^i \setminus i}} T_{\tau_e^i}(u)\right] \\
&= \partial_{u_i} \left[\prod_{e\in N(i)} \mu_{e\to i}(u)\right].
\end{aligned} \tag{18}$$

The quantity $\mu_{e\to i}(u) := \partial_{u_{\tau_e^i \setminus i}} T_{\tau_e^i}(u)$ is called a message from the edge indexed by e to the variable indexed by i. Now consider $T_{\tau_e^i}(u)$ and let j be the neighbor variable index of e. One can go deeper in the tree, that is, we have

$$T_{\tau_e^i}(u) = \Phi_e(u_i, u_j) T_{\tau_j^e}(u)$$

where τ_j^e is the subtree rooted at the edge indexed by e and containing the variable indexed by j (see Figure 3). Hence,

$$\partial_{u_{\tau_e^i \setminus i}} T_{\tau_e^i}(u) = \partial_{u_j} \left[\phi_e(u_i, u_j) \partial_{u_{\tau_j^e \setminus j}} T_{\tau_j^e}(u) \right] = \partial_{u_j} \left[\phi_e(u_i, u_j) \mu_{j \to e}(u) \right].$$

A second type of message has been defined: $\mu_{j \to e}(u) := \partial_{u_{\tau_j^e \setminus j}} T_{\tau_j^e}(u)$ is called a message from the variable index j to the edge index e. Again,

$$T_{\tau_j^e}(u) = \prod_{e' \in N(j) \setminus e} T_{\tau_{e'}^j}(u),$$

hence,

$$\partial_{u_{\tau_j^e \setminus j}} T_{\tau_j^e}(u) = \prod_{e' \in N(j) \setminus e} \partial_{u_{\tau_{e'}^j \setminus j}} T_{\tau_{e'}^j}(u) = \prod_{e' \in N(j) \setminus e} \mu_{e' \to j}(u),$$

where the message $\mu_{e' \to j}(u)$ has been already defined in (18). To summarize, the calculation of $\mu_{e \to i}(u)$ requires the calculation of $\mu_{j \to e}(u)$, which, in turn, requires the calculation of $\mu_{e' \to j}(u)$, where $e = \{ij\}$ and e' is an edge index attached to j. The algorithm presented above allows to compute recursively all the messages from the leaves to the root. Once all the messages have been computed, the density is given by the derivative with respect to the root of the product of all the messages (18).

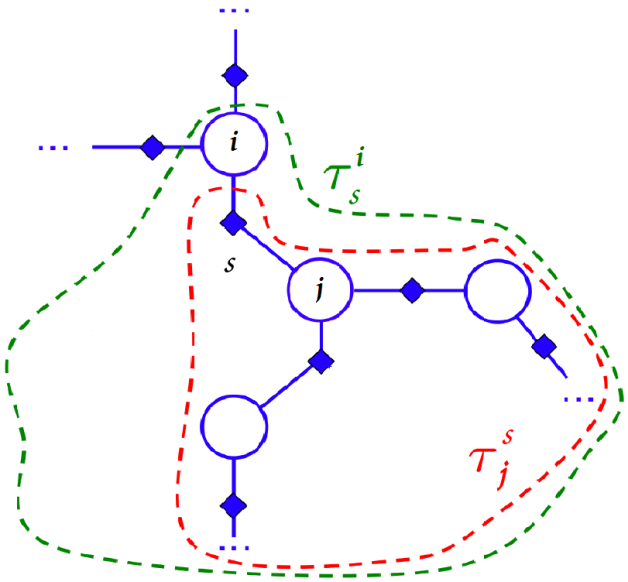

Figure 3: Examples of subtrees. This figure is partly drawn from [10].

References

[1] K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198, 2009.

[2] L.B.G Andersen and V.V. Piterbarg. *Interest Rate Modeling.* Atlantic Financial Press, 2010.

[3] S. Demarta and A. J. McNeil. The t copula and related copulas. *International Statistical Review*, 73(1):111–129, 2005.

[4] F. Durante and G. Salvadori. On the construction of multivariate extreme value models via copulas. *Environmetrics*, 21(2):143–161, 2010.

[5] G. Frahm, M. Junker, and A. Szimayer. Elliptical copulas: applicability and limitations. *Statistics & Probability Letters*, 63(3):275–286, 2003.

[6] C. Genest, K. Ghoudi, and L-P. Rivest. Understanding relationships using copulas. *North American Actuarial Journal*, 2(3):143–149, 1998.

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

[8] G. Gudendorf and J. Segers. Extreme-value copulas. In *Copula Theory and Its Applications*, page 127–145. Springer, 2010.

[9] M. Hofert, M. Mächler, and A. J. McNeil. Archimedean copulas in high dimensions: Estimators and numerical challenges motivated by financial applications. *Journal de la Société Française de Statistique*, 154(1):25–63, 2012.

[10] J. C. Huang. *Cumulative distribution networks: Inference, estimation and applications of graphical models for cumulative distribution functions.* PhD thesis, University of Toronto, 2009.

[11] J.C. Huang and N. Jojic. Maximum-likelihood learning of cumulative distribution functions on graphs. *Journal of Machine Learning Research W&CP Series*, 9:342–349, 2010.

[12] H. Joe. *Multivariate models and dependence concepts.* Chapman & Hall/CRC, 2001.

[13] D. Kim and J.M. Kim. Analysis of directional dependence using asymmetric copula-based regression models. *Journal of Statistical Computation and Simulation*, 84(9):1990–2010, 2014.

[14] E. Liebscher. Construction of asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 99(10):2234–2250, 2008.

[15] B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–39, 1988.

[16] R.B. Nelsen. *An introduction to copulas.* Springer, 2006.

[17] M. Sklar. *Fonctions de répartition à n dimensions et leurs marges.* Publications de l'Institut de Statistique de l'Université de Paris, 8:229-231, 1959.

[18] T. Van Pham and G. Mazo. *PBC: product of bivariate copulas.* http://cran.r-project.org, 2014. R package version 1.2.

[19] L. Zhang and V. P. Singh. Gumbel–Hougaard copula for trivariate rainfall frequency analysis. *Journal of Hydrologic Engineering*, 12(4):409–419, 2007.