Journal of Statistical Planning and Inference 166 (2015) 2–13

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Inference on Archimedean copulas using mixtures of Pólya trees

Simon Guillotte [a,*], Julien Perron [b]

[a] Département de mathématiques, Université du Québec à Montréal, 201 Ave du Président-Kennedy, H2X 3Y7, Canada

[b] Université du Québec à Montréal, Canada

ARTICLE INFO

Article history:
Received 25 November 2013
Received in revised form 10 January 2014
Accepted 10 January 2014
Available online 20 January 2014

Keywords:
Archimedean copulas
Kendall distribution
MCMC
Mixtures of Pólya trees

ABSTRACT

Assume that $X = (X_1, \ldots, X_d)$, $d \geqslant 2$ is a random vector having joint cumulative distribution function H with continuous marginal cumulative distribution functions $F_1, \ldots, F_d$ respectively. Sklar's decomposition yields a unique copula C such that $H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$ for all $(x_1, \ldots, x_d) \in \mathbb{R}^d$. Here $F_1, \ldots, F_d$ and C are the unknown parameters, the one of interest being the copula C. We assume C to belong to the Archimedean family, that is $C = C_\psi$, for some Archimedean generator ψ. We exploit the well known fact that such a generator is in one-to-one correspondence with the distribution function of a nonnegative random variable R with no atom at zero. In order to adopt a Bayesian approach for inference, a prior on the Archimedean family may be selected via a prior on the cumulative distribution function F of R. A mixture of Pólya trees is proposed for F, making the model very flexible, yet still manageable. The induced prior is concentrated on the space of absolutely continuous d-dimensional Archimedean copulas and explicit forms for the generator and its derivatives are available. To the best of our knowledge, others in the literature have not yet considered such an approach. An extensive simulation study is carried out to compare our estimator with a popular frequentist nonparametric estimator. The results clearly indicate that if intensive computing is available, our estimator is worth considering, especially for small samples.

1. Introduction

Let $\mathscr{H}$ be the space of multivariate cumulative distribution functions with continuous margins. Consider a random vector $X = (X_1, \ldots, X_d)$ with joint cumulative distribution function (cdf) $H \in \mathscr{H}$, and let $F_1, \ldots, F_d$ be the (continuous) marginal cdfs of $X_1, \ldots, X_d$ respectively. Now let $U_1 = F_1(X_1), \ldots, U_d = F_d(X_d)$, and consider the copula C given by

$$C(u_1, \ldots, u_d) = P(U_1 \leqslant u_1, \ldots, U_d \leqslant u_d), \quad (u_1, \ldots, u_d) \in [0, 1]^d.$$

In this case, Sklar's Theorem says that C is the unique copula for which H admits the representation $H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$, $(x_1, \ldots, x_d) \in \mathbb{R}^d$. The copula C holds the properties of the joint cdf H which are invariant with respect to strictly increasing transformations of the margins $X_1, \ldots, X_d$. A d-dimensional Archimedean copula has the form

$$C_\psi(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)), \quad u_1, \ldots, u_d \in [0, 1], \tag{1.1}$$

* Corresponding author.
E-mail addresses: guillotte.simon@uqam.ca (S. Guillotte), perrjul@gmail.com (J. Perron).

http://dx.doi.org/10.1016/j.jspi.2014.01.003
0378-3758/

where ψ is called an Archimedean generator, that is, a continuous nonincreasing function defined on $[0, \infty)$, with $\psi(0) = 1$, $\psi(\infty) = 0$, and $\psi^{-1}(0) = \inf\{x : \psi(x) = 0\}$, where $\inf \emptyset = +\infty$. In particular, ψ is decreasing on $[0, \psi^{-1}(0))$.

A characterization of Archimedean generators that yield d-dimensional copulas for some fixed dimension $d \geqslant 2$ was recently established by McNeil and Nešlehová (2009). In particular, the characterization openned the door to new inferential techniques for Archimedean copulas, and that, in the multivariate case. This has been the topic of the paper by Genest et al. (2011), where a rank-based approach is developed in a frequentist setting. This approach is well justified only in the cases $d = 2$ and $d = 3$. In short, their methodology is valid if an Archimedean copula C_ψ is in one-to-one correspondence with its Kendall distribution

$$K(w) = \mathrm{P}(W \leqslant w), \quad w \in [0, 1], \tag{1.2}$$

with $W = C_\psi(U_1, \ldots, U_d)$ and $(U_1, \ldots, U_d)$ distributed according to C_ψ. The estimator for C_ψ is constructed via the empirical estimator for K, but for this to make sense, there must be only one C_ψ associated to K. This can be shown to hold in dimensions $d = 2$ and 3, but it nevertheless remains a conjecture for $d > 3$. As discussed in Embrechts and Hofert (2011), Janssen and Duchateau (2011), Segers (2011), Tsukahara (2011) and Wang and Emura (2011), this problem is challenging and although there is strong evidence in its favor, according to Segers (2011), nothing should be taken for granted. In the bivariate case ($d = 2$), nonparametric Bayesian methodology based on this result appears in Lambert (2007), using a polynomial splines model for the Kendall distribution K. The latter approach is interesting, but seems a bit complicated and works only in dimension $d = 2$.

One objective here is to propose Bayesian nonparametric inference which does not rely on the above identifiability result, and valid in any dimension $d \geqslant 2$. The Bayesian approach requires a likelihood function and so we want a prior that is concentrated on Archimedean copulas with densities (with respect to d-dimensional Lebesgue measure). Note here in passing that the Genest et al. (2011) estimator does not have a density due to the fact that it is constructed from the empirical cdf of K. To compute the likelihood, we shall need to evaluate high-order derivatives of ψ. As pointed out by Genest et al. (2011), by Embrechts and Hofert (2011), and by Hering and Stadtmüller (2012), this is not immediate mainly because of the complexity of the resulting algebraic expressions, numerical accuracy losses or computational speed. Our main goal is to construct a flexible prior on the generator ψ for which both its representation and its derivatives can be made explicit (in closed form). This will make the methodology conceptually simple, easy to implement, and fast to run on a computer. Ultimately, the approach should also give a low mean integrated squared error.

In Section 2, we select a prior on the generator ψ by exploiting the result that the latter is in one-to-one correspondence with the distribution function of a nonnegative random variable R with no atom at zero, see McNeil and Nešlehová (2009). In fact, if S is a random vector uniformly distributed over the standard simplex, and independent of R, then C_ψ corresponds to the survival copula of the random vector $Y = RS$. The prior on ψ is therefore obtained by selecting a prior on the distribution of R. To do so, we proceed as in Hanson and Johnson (2002), Branscum and Hanson (2008) and Zhao et al. (2009), and consider a mixture of finite Pólya trees, see also Hanson (2006) and Christensen et al. (2008). Essentially, the base (or centering) distribution of the Pólya tree is assumed to belong to a parametric family, instead of being fixed, and a mixing distribution is selected. This has the advantage of removing the influence of a single centering distribution on the inference. In particular, it smooths out the effect of the partition induced by the centering distribution. The particular choice of the parametric family together with the mixing distribution play an important role in that respect. They are also critical for the evaluation of ψ and its derivatives. Next, in Section 3, we describe how to obtain a numerical approximation of the Bayes estimator and discuss a way to make predictions (this being a true practical advantage over a frequentist approach for such problems). Finally, the results of an extensive simulation study is carried out where comparisons are made with the nonparametric estimator proposed by Genest et al. (2011). Using essentially the same simulation setup as theirs, our estimator outperformed their estimator in every case (81 different situations in total), and the difference is quite remarkable especially for small sample sizes.

2. The model

As in Genest et al. (2011), our model is based on the following. Consider a d-dimensional, $d \geqslant 2$, Archimedean copula of the form (1.1). A characterization of Archimedean generators that yield d-dimensional copulas for some fixed dimension $d \geqslant 2$ is provided via the so-called Williamson d-transform. The Williamson d-transform of a nonnegative random variable R with cumulative distribution function F is given by

$$\mathfrak{W}_d F(x) = \mathrm{E}\left(1 - \frac{x}{R}\right)_+^{d-1} = \int_{(x,\infty)} \left(1 - \frac{x}{r}\right)^{d-1} F(dr), \quad \text{for all } x \geqslant 0, \tag{2.1}$$

and the correspondence between F and $\mathfrak{W}_d F$ is one to one. It turns out that an Archimedean generator ψ yields a d-dimensional copula C_ψ via (1.1) if and only if there exists a random variable R with cumulative distribution function F, such that $F(0) = 0$ and $\psi = \mathfrak{W}_d F$. Let Ψ_d denote the space of such Archimedean generators. Here F is called the radial cumulative distribution function associated with ψ. Important to us is the equivalence between absolute continity of F, absolute continuity of $\psi^{(d-1)}$ on $(0, \infty)$, and a density for the copula C_ψ. When the latter copula has a density, its expression