Journal of Multivariate Analysis 159 (2017) 82–110

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Asymptotic behavior of the empirical multilinear copula process under broad conditions

Christian Genest [a,*], Johanna G. Nešlehová [a], Bruno Rémillard [b]

[a] Department of Mathematics and Statistics, McGill University, 805, rue Sherbrooke ouest, Montréal (Québec), Canada H3A 0B9

[b] Département de sciences de la décision, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec), Canada H3T 2A4

ARTICLE INFO

Article history:
Received 14 January 2017
Available online 19 April 2017

AMS 2010 subject classifications:
62E20
60F05
62G30
62H20

Keywords:
Checkerboard copula
Copula model
Discontinuous margins
Empirical copula process
Mixed data
Multiplier bootstrap
Rank-based inference
Ties

ABSTRACT

The empirical checkerboard copula is a multilinear extension of the empirical copula, which plays a key role for inference in copula models. Weak convergence of the corresponding empirical process based on a random sample from the underlying multivariate distribution is established here under broad conditions which allow for arbitrary univariate margins. It is only required that the underlying checkerboard copula has continuous first-order partial derivatives on an open subset of the unit hypercube. This assumption is very weak and always satisfied when the margins are discrete. When the margins are continuous, one recovers the limit of the classical empirical copula process under conditions which are comparable to the weakest ones currently available in the literature. A multiplier bootstrap method is also proposed to replicate the limiting process and its validity is established. The empirical checkerboard copula is further shown to be a more precise estimator of the checkerboard copula than the empirical copula based on jittered data. Finally, the weak convergence of the empirical checkerboard copula process is shown to be sufficiently strong to derive the asymptotic behavior of a broad class of functionals that are directly relevant for the development of rigorous statistical methodology for copula models with arbitrary margins.

1. Introduction

Copula-based models are now widespread. They have become indispensable tools for the analysis of multivariate data, e.g., in biostatistics, finance, insurance, and hydrology. Introductions to copula modeling are provided, e.g., by Genest and Favre [17], Kurowicka and Joe [30], Genest and Nešlehová [19], and Joe [28].

A copula model for a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is based on the decomposition of its joint distribution function H in terms of its univariate margins $F_1, \ldots, F_d$ and a copula C, i.e., a d-variate distribution function whose margins are standard uniform. As shown by Sklar [43], there always exists a copula C such that, for all $x_1, \ldots, x_d \in \mathbb{R}$,

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

A copula model for H is then obtained by assuming that $F_1, \ldots, F_d$ and C belong to suitable classes of univariate distributions and copulas, respectively. As the right-hand side of (1) leads to a *bona fide* distribution for any choice of margins and copula,

* Corresponding author.
E-mail addresses: christian.genest@mcgill.ca (C. Genest), johanna.neslehova@mcgill.ca (J.G. Nešlehová), bruno.remillard@hec.ca (B. Rémillard).

http://dx.doi.org/10.1016/j.jmva.2017.04.002

this modeling approach is extremely flexible, whence its success. The main challenge typically resides in the inference on the copula, through which dependence is introduced in the model.

Let $\mathbf{X}_1 = (X_{11}, \ldots, X_{1d}), \ldots, \mathbf{X}_n = (X_{n1}, \ldots, X_{nd})$ be a random sample from H, and let H_n be the corresponding empirical distribution function with univariate margins $F_{n1}, \ldots, F_{nd}$. A fundamental tool for inference on C is the empirical copula $\hat{C}_n$, defined as the empirical distribution of the pseudo-sample $(F_{n1}(X_{11}), \ldots, F_{nd}(X_{1d})), \ldots, (F_{n1}(X_{n1}), \ldots, F_{nd}(X_{nd}))$.

When $F_1, \ldots, F_d$ are continuous, it is indeed natural to rely on $\hat{C}_n$. In this case, the copula C in (1) is unique and equal to the distribution function of the unobservable vector $(F_1(X_1), \ldots, F_d(X_d))$. Furthermore, the empirical copula process $\hat{\mathbb{C}}_n = \sqrt{n}\,(\hat{C}_n - C)$ converges weakly, as $n \to \infty$, to a centered Gaussian process. Various refinements of this result were obtained under successively weaker smoothness assumptions on the underlying copula C; see, e.g., Rüschendorf [37], Deheuvels [9–11], Stute [44], Gänßler and Stute [16], van der Vaart and Wellner [46], Fermanian et al. [15], Genest and Rémillard [24], Ghoudi and Rémillard [26], Tsukahara [45], Segers [41], as well as Bücher and Volgushev [6] and Bücher et al. [5].

The purpose of this paper is to extend the study of the empirical copula process to the case of arbitrary margins $F_1, \ldots, F_d$. This extension allows for the presence of ties in the data, which is a frequent phenomenon in practice. The key difficulty in this endeavor is that relation (1) is satisfied by an infinite (in fact, uncountable) number of copulas and so the process $\hat{\mathbb{C}}_n$ is no longer well defined as soon as at least one of the margins is discontinuous.

To resolve this issue, we propose to work with a simple multilinear extension $\hat{C}_n^{\maltese}$ of $\hat{C}_n$. This so-called empirical checkerboard copula, which was introduced by Deheuvels [9] in the continuous case, is always a *bona fide* copula. When the marginal distributions are continuous, $\hat{C}_n^{\maltese}$ and $\hat{C}_n$ never differ point-wise by more than d/n almost surely and hence their limiting behavior is the same. When $F_1, \ldots, F_d$ are purely discrete, Genest et al. [21] showed that $\hat{C}_n^{\maltese}$ is a consistent estimator of the unique checkerboard copula $C^{\maltese}$ of H, which appears in the proof of Sklar's theorem; see, e.g., Moore and Spruill [31] or Deheuvels [8]. Genest and Nešlehová [18], who studied this construction in detail, argue that the dependence structure of H is best represented by $C^{\maltese}$. Additional evidence pointing to the importance of $C^{\maltese}$ is provided by Denuit and Lambert [12], Nešlehová [33], and Genest et al. [21], who show that when computed from $\hat{C}_n^{\maltese}$, dependence measures such as Kendall's tau or Spearman's rho coincide with the classical tie-corrected versions of these coefficients.

It will be proved here that the empirical checkerboard copula process $\hat{\mathbb{C}}_n^{\maltese} = \sqrt{n}\,(\hat{C}_n^{\maltese} - C^{\maltese})$ converges weakly to a centered Gaussian process for any choice of margins $F_1, \ldots, F_d$, thereby extending the findings of Genest et al. [21], who only considered the case of purely discrete margins. When the margins of H are arbitrary, the matter is far more complex and entirely new arguments are needed. As in the purely discrete case, the key issue is the choice of space on which the weak convergence can be established. As noted by Fermanian et al. [15], the classical empirical copula process does not converge weakly in the space $\ell^\infty([0, 1]^d)$ of bounded functions on $[0, 1]^d$ equipped with the sup norm, unless the underlying copula C has continuous first-order partial derivatives. A similar problem arises here.

To reach our goal, we will require that the first-order partial derivatives of $C^{\maltese}$ exist and are continuous on an open set $\mathcal{O} \subset [0, 1]^d$, and we will show that, for any compact set $K \subset \mathcal{O}$, the process $\hat{\mathbb{C}}_n^{\maltese}$ converges weakly in the space $\mathcal{C}(K)$ of continuous functions on K equipped with the sup norm. In the special case of continuous margins, our approach implies the convergence of the classical empirical copula process $\hat{\mathbb{C}}_n$ under regularity conditions that are equally weak as those of Bücher et al. [5], whose result was heretofore regarded as the most general available.

In the literature, the presence of ties is frequently resolved by breaking them at random. This strategy, called "data jittering", is considered in the copula modeling context, e.g., by Cameron et al. [7], Kojadinovic and Yan [29], Shi and Valdez [42], and Bücher and Kojadinovic [4]; see also Genest and Nešlehová [18] and Genest et al. [22] for discussion. When the support of the margins is known, such as when the variables $X_1, \ldots, X_d$ are counts, jittering amounts to adding a small noise to each observation at random. As will be shown, the empirical copula based on the jittered data, say $\hat{C}_n^{\mathrm{R}}$, is a consistent estimator of $C^{\maltese}$. However, it will also be seen that the variance of the limit of $\hat{\mathbb{C}}_n^{\mathrm{R}} = \sqrt{n}\,(\hat{C}_n^{\mathrm{R}} - C^{\maltese})$ is systematically larger than the variance of the limiting empirical multilinear copula process. When the additional variability caused by randomization is removed by taking the expectation of $\hat{C}_n^{\mathrm{R}}$ over the jitters, the resulting estimator $\hat{C}_n^{®}$ differs from the empirical checkerboard copula $\hat{C}_n^{\maltese}$ by at most d/n. This implies that $\hat{C}_n^{®}$ has the same limiting behavior as $\hat{C}_n^{\maltese}$, which has the additional advantage of being a *bona fide* copula.

After introducing the notations in Section 2, we give an analytic definition of $C^{\maltese}$ in Section 3, along with a convenient stochastic representation of the same. The empirical checkerboard copula $\hat{C}_n^{\maltese}$ is then defined in Section 4 and its asymptotic behavior is determined in Section 5. The proof is outlined in Section 6 and detailed in a series of Appendices. In Section 7, the validity of the bootstrap method is established as a means of approximating the limiting distribution of $\hat{\mathbb{C}}_n^{\maltese}$. The checkerboard approach is then compared to the data jittering strategy in Section 8. Finally, it is shown in Section 9 that the weak convergence of $\hat{\mathbb{C}}_n^{\maltese}$ derived herein is sufficient to determine the asymptotics of a broad range of statistics arising in the context of dependence modeling.

2. Notations

Let G be a univariate distribution function. Its left-continuous inverse is defined, for all $u \in [0, 1]$, by $G^{-1}(u) = \inf\{t \in \mathbb{R} : G(t) \geq u\}$. Let $\mathcal{R}_G = \{G(y) : y \in \mathbb{R}\} \subseteq [0, 1]$ be the range of G and $\bar{\mathcal{R}}_G$ its closure in $[0, 1]$. For each $y \in [-\infty, \infty]$, $G(y_-)$