Journal of Statistical Planning and Inference 160 (2015) 23–34

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Hybrid copula estimators

Johan Segers

Université catholique de Louvain, ISBA, Voie du Roman Pays 20, bte L1.04.01, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 8 May 2014
Received in revised form 22 November 2014
Accepted 25 November 2014
Available online 15 December 2014

Keywords:
Copula
Empirical process
Functional delta method
Influence function
Missing data

ABSTRACT

An extension of the empirical copula is considered by combining an estimator of a multivariate cumulative distribution function with estimators of the marginal cumulative distribution functions for marginal estimators that are not necessarily equal to the margins of the joint estimator. Such a hybrid estimator may be reasonable when there is additional information available for some margins in the form of additional data or stronger modelling assumptions. A functional central limit theorem is established and some examples are developed.

1. Introduction

Let H be a p-variate cumulative distribution function with continuous margins $F_1, \ldots, F_p$ and copula C (Sklar, 1959). We have

$$H(\boldsymbol{x}) = C\big(F_1(x_1), \ldots, F_p(x_p)\big), \qquad \boldsymbol{x} \in \mathbb{R}^p,$$
$$C(\boldsymbol{u}) = H\big(F_1^{\leftarrow}(u_1), \ldots, F_p^{\leftarrow}(u_p)\big), \qquad \boldsymbol{u} \in [0, 1]^p.$$

Here, $G^{\leftarrow}$ denotes the left-continuous inverse of a univariate cumulative distribution function G, i.e.,

$$G^{\leftarrow}(u) = \inf\{x \in \mathbb{R} : G(x) \geqslant u\}, \quad u \in [0, 1].$$

Throughout, standard conventions regarding infinities are employed: $\inf \varnothing = +\infty$, $G(-\infty) = 0$, and $G(+\infty) = 1$.

Let $\hat{H}_n$ and $\hat{F}_{n,j}$ be estimator sequences of H and F_j ($j = 1, \ldots, p$), respectively. Consider the copula estimator

$$\hat{C}_n(\boldsymbol{u}) = \hat{H}_n\big(\hat{F}_{n,j}^{\leftarrow}(u_1), \ldots, \hat{F}_{n,p}^{\leftarrow}(u_p)\big), \quad \boldsymbol{u} \in [0, 1]^p. \tag{1.1}$$

Note that $\hat{F}_{n,j}$ is not necessarily equal to the jth marginal distribution function, $\hat{H}_{n,j}$, of $\hat{H}_n$. We call $\hat{C}_n$ a *hybrid copula estimator*.

Given a rate $0 < r_n \to \infty$ (typically $r_n = \sqrt{n}$), the normalized estimation error of the hybrid copula estimator is

$$\mathbb{C}_n(\boldsymbol{u}) = r_n\big(\hat{C}_n(\boldsymbol{u}) - C(\boldsymbol{u})\big), \quad \boldsymbol{u} \in [0, 1]^p. \tag{1.2}$$

The aim is to establish weak convergence of $\mathbb{C}_n$ in the space $\ell^\infty([0, 1]^p)$ of bounded, real-valued functions on $[0, 1]^p$ equipped with the supremum norm.

E-mail address: johan.segers@uclouvain.be.

http://dx.doi.org/10.1016/j.jspi.2014.11.006
0378-3758/

If $\hat{H}_n$ and $\hat{F}_{n,j} = \hat{H}_{n,j}$ are the joint and marginal empirical distribution functions of a p-variate sample of size n, then $\hat{C}_n$ is just the Deheuvels–Rüschendorf empirical copula, see Examples 3.1 and 3.2. However, there may be good reasons not to estimate F_j by $\hat{H}_{n,j}$ but by a different estimator. It may be that there is information available on the jth margin which cannot directly be used by the joint estimator $\hat{H}_n$.

- A parametric model may be reasonable for some or all of the marginal distributions but not for the joint distribution (Example 3.4). This is the case for instance when the data are vectors of annual maxima. Asymptotic theory then suggests to model the vector of componentwise maxima by a multivariate max-stable distribution (de Haan and Resnick, 1977; Deheuvels, 1978; Galambos, 1978). The marginal distributions are univariate extreme-value distributions, whereas the copula belongs to the infinite-dimensional family of extreme-value copulas.
- Some entries in the $n \times p$ data matrix may be missing (Example 3.5). Then $\hat{H}_n$ may be defined as the empirical distribution function of all data rows which are complete, whereas $\hat{F}_{n,j}$ is the empirical distribution function of all observed entries in the jth column.
- Similarly, in a time series setting, the observation periods of the p univariate series could be different and overlap only partially. Again, one could estimate F_j by the complete series for that variable but estimate H only based on the time period where all series were recorded simultaneously. In the same spirit, there may be additional samples for some of the variables.

The structure of the paper is as follows. The main result, Theorem 2.3, is given in Section 2, stating weak convergence of the hybrid copula estimator process in (1.2) under high-level conditions on the estimators of the joint and marginal distribution functions. Special cases and examples are worked out in Section 3. All proofs and calculations are deferred to Section 4.

Throughout, the following notations are used. For an arbitrary set T, let $\ell^\infty(T)$ be the space of bounded, real-valued functions on T, the space being equipped with the supremum distance $\|f\|_\infty = \sup_{t \in T} |f(t)|$ for $f \in \ell^\infty(T)$. The indicator variable of a set E is denoted by $\mathbb{1}_E$, whereas the identity mapping on a set E is denoted by id_E. Weak convergence in the sense of J. Hoffmann-Jørgensen is denoted by the arrow '$\rightsquigarrow$'; see Part 1 in the monograph by van der Vaart and Wellner (1996).

2. Main result

Besides the continuity of the margins $F_1, \ldots, F_p$, two assumptions will be made. The first assumption imposes among others a bit of smoothness on the target copula C, without which there is little hope of establishing weak convergence of $\mathbb{C}_n$ in (1.2) with respect to the supremum norm on $\ell^\infty([0, 1]^p)$ (Segers, 2012). The second assumption is a high-level condition concerning the asymptotic distribution of the estimators $\hat{H}_n$ and $\hat{F}_{n,j}$ and is to be checked on a case-by-case basis. See Remarks 2.4 and 2.5 and see the examples in Section 3.

Condition 2.1. (a) *The p-variate distribution function H has continuous margins $F_1, \ldots, F_p$ and copula C.*
(b) *For all $j \in \{1, \ldots, p\}$, the first-order partial derivative $\dot{C}_j(\boldsymbol{u}) = \partial C(\boldsymbol{u})/\partial u_j$ exists and is continuous on the set $\{\boldsymbol{u} \in [0, 1]^p : 0 < u_j < 1\}$.*

For convenience, collect the marginal distribution and quantile functions into vector-valued functions $\boldsymbol{F}$ and $\boldsymbol{F}^{\leftarrow}$:

$$\boldsymbol{F}(\boldsymbol{x}) = \big(F_1(x_1), \ldots, F_p(x_p)\big), \qquad \boldsymbol{x} \in \mathbb{R}^p; \tag{2.1}$$

$$\boldsymbol{F}^{\leftarrow}(\boldsymbol{u}) = \big(F_1^{\leftarrow}(u_1), \ldots, F_p^{\leftarrow}(u_p)\big), \qquad \boldsymbol{u} \in [0, 1]^p. \tag{2.2}$$

Condition 2.2. *There exists $0 < r_n \to \infty$ such that in the space $\ell^\infty(\mathbb{R}^p) \otimes (\ell^\infty(\mathbb{R}) \otimes \cdots \otimes \ell^\infty(\mathbb{R}))$ equipped with the topology of uniform convergence, we have joint weak convergence*

$$\big(r_n(\hat{H}_n - H);\ r_n(\hat{F}_{n,1} - F_1), \ldots, r_n(\hat{F}_{n,p} - F_p)\big) \rightsquigarrow (\alpha \circ \boldsymbol{F};\ \beta_1 \circ F_1, \ldots, \beta_p \circ F_p), \quad n \to \infty. \tag{2.3}$$

The stochastic processes α and β_j take values in $\ell^\infty([0, 1]^p)$ and $\ell^\infty([0, 1])$, respectively, and are such that $\alpha \circ \boldsymbol{F}$ and $\beta_j \circ F_j$ have continuous trajectories on $[-\infty, \infty]^p$ and $[-\infty, \infty]$ almost surely.

Usually, $r_n = \sqrt{n}$, although Condition 2.2 allows for different convergence rates. Joint weak convergence in (2.3) can typically be established when the estimators $\hat{H}_n$ and $\hat{F}_{n,j}$ can be written as functionals of the same underlying empirical process. Because $\dot{C}_j(\boldsymbol{u})$ need not be defined if $u_j \in \{0, 1\}$, some care is needed in the formulation of the following theorem.

Theorem 2.3 (*Hybrid Copula Process*)**.** *If Conditions 2.1 and 2.2 hold, then, uniformly in $\boldsymbol{u} \in [0, 1]^p$,*

$$r_n\{\hat{C}_n(\boldsymbol{u}) - C(\boldsymbol{u})\} = r_n\{\hat{H}_n(\boldsymbol{F}^{\leftarrow}(\boldsymbol{u})) - C(\boldsymbol{u})\} - \sum_{j=1}^{p} \dot{C}_j(\boldsymbol{u})\, r_n\{\hat{F}_{n,j}(F_j^{\leftarrow}(u_j)) - u_j\}\, \mathbb{1}_{(0,1)}(u_j) + o_p(1), \tag{2.4}$$