Journal of Multivariate Analysis 101 (2010) 1837–1845

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On the covariance of the asymptotic empirical copula process

Christian Genest [a,*], Johan Segers [b]

[a] Département de mathématiques et de statistique, Université Laval, 1045, avenue de la Médecine, Québec, Canada G1V 0A6

[b] Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université catholique de Louvain, Voie du Roman Pays 20, B–1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 21 August 2009
Available online 18 April 2010

AMS subject classifications:
primary 60F17
secondary 60E15
62H20

Keywords:
Asymptotic variance
Copula
Dependence parameter
Empirical process
Independence
Left-tail decreasing
Rank-based inference

ABSTRACT

Conditions are given under which the empirical copula process associated with a random sample from a bivariate continuous distribution has a smaller asymptotic covariance function than the standard empirical process based on observations from the copula. Illustrations are provided and consequences for inference are outlined.

1. Introduction

Consider a pair (X, Y) of continuous random variables whose joint and marginal cumulative distribution functions are defined for all $x, y \in \mathbb{R}$ by

$$H(x, y) = \Pr(X \leqslant x, Y \leqslant y), \qquad F(x) = \Pr(X \leqslant x), \qquad G(y) = \Pr(Y \leqslant y),$$

respectively. The transformed variables $U = F(X)$ and $V = G(Y)$ are then uniform on $[0, 1]$ and their joint distribution function, defined at every $u, v \in [0, 1]$ by

$$C(u, v) = \Pr(U \leqslant u, V \leqslant v),$$

is the unique copula C associated with H. The two functions are related through the equation $C(u, v) = H\{F^{-1}(u), G^{-1}(v)\}$, where $F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geqslant u\}$ and $G^{-1}(v) = \inf\{y \in \mathbb{R} : G(y) \geqslant v\}$ for all $u, v \in (0, 1)$. Inference on C is of interest, as it characterizes the dependence in the pair (X, Y); see, e.g., [1,2]. In particular, all margin-free concepts and measures of association such as Kendall's tau, Spearman's rho, Blomqvist's beta, Gini's gamma and Spearman's footrule depend only on C.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from H and write $U_i = F(X_i)$, $V_i = G(Y_i)$ for all $i \in \{1, \ldots, n\}$. When F and G are known, a natural estimate of C is then given by the empirical distribution function of the sample $(U_1, V_1), \ldots, (U_n, V_n)$, defined at every $u, v \in [0, 1]$ by

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(U_i \leqslant u, V_i \leqslant v).$$

* Corresponding author.
E-mail addresses: Christian.Genest@mat.ulaval.ca (C. Genest), Johan.Segers@uclouvain.be (J. Segers).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2010.03.018

In fact, standard results from the theory of empirical processes [3] imply that $\mathbb{C}_n = n^{1/2}(C_n - C)$ converges weakly, as $n \to \infty$, to a centered Gaussian process $\mathbb{C}$ on $[0, 1]^2$ with continuous trajectories and covariance function given by

$$\mathrm{cov}\{\mathbb{C}(u, v), \mathbb{C}(s, t)\} = C(u \wedge s, v \wedge t) - C(u, v)C(s, t),$$

for all $u, v, s, t \in [0, 1]$, with $a \wedge b = \min(a, b)$ for arbitrary $a, b \in \mathbb{R}$.

When the margins F and G are unknown, as is generally the case in practice, they can be estimated by their empirical counterparts, F_n and G_n. A surrogate sample from C is then given by the pairs $(\hat{U}_1, \hat{V}_1), \ldots, (\hat{U}_n, \hat{V}_n)$, where $\hat{U}_i = F_n(X_i)$ and $\hat{V}_i = G_n(Y_i)$ for all $i \in \{1, \ldots, n\}$. The corresponding empirical distribution function, defined at every $u, v \in [0, 1]$ by

$$\hat{C}_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\hat{U}_i \leqslant u, \hat{V}_i \leqslant v),$$

is traditionally called the empirical copula [4], although it is not a copula *stricto sensu*. The function $\hat{C}_n$ provides a rank-based, consistent estimate of C often used in practice for copula model selection and goodness-of-fit purposes; see, e.g., [5,6].

To be specific, let $\dot{C}_1(u, v) = \partial C(u, v)/\partial u$ and $\dot{C}_2(u, v) = \partial C(u, v)/\partial v$ denote the partial derivatives of an arbitrary copula C, known to exist almost everywhere [2, Chapter 2]. Now assume that they exist in fact everywhere and that they are continuous on $(0, 1)^2$. Under these mild regularity conditions, it is then well known [7,8] that the empirical copula process $\hat{\mathbb{C}}_n = n^{1/2}(\hat{C}_n - C)$ converges weakly, as $n \to \infty$, to a centered Gaussian process $\hat{\mathbb{C}}$ defined at every $u, v \in [0, 1]$ by

$$\hat{\mathbb{C}}(u, v) = \mathbb{C}(u, v) - \dot{C}_1(u, v)\,\mathbb{C}(u, 1) - \dot{C}_2(u, v)\,\mathbb{C}(1, v).$$

In fact, it is observed in [8, Theorem 4] that the assumption on the existence and continuity of the partial derivatives is necessary for the empirical copula process to converge on $[0, 1]^2$.

In the copula modeling literature, the difference between $\mathbb{C}$ and $\hat{\mathbb{C}}$ is often interpreted as "the price to pay for the fact that the margins are unknown". This suggests that if F and G were known, it would be preferable to base the inference on C_n rather than on $\hat{C}_n$. It is shown here, perhaps surprisingly, that the opposite is true under weak positive dependence conditions on C. When this happens, procedures based on $\hat{\mathbb{C}}$ are thus more efficient than the analogous procedures based on $\mathbb{C}$.

The key result is stated and illustrated in Section 2, along with a partial extension to the case of negative dependence. In Section 3, circumstances are delineated under which a dependence parameter, say $\theta = \mathscr{T}(C)$, can be estimated more efficiently by a rank-based estimate $\hat{\theta}_n = \mathscr{T}(\hat{C}_n)$ than by the analogous estimate $\theta_n = \mathscr{T}(C_n)$ which exploits the knowledge of the margins. Concluding remarks are given in Section 4, along with a partial multivariate extension of the main result. All technical arguments are collected in the Appendix.

2. Main result

Following [9], suppose that the two continuous random variables X and Y are such that the mappings $t \mapsto \Pr(X \leqslant x \mid Y \leqslant t)$ and $t \mapsto \Pr(Y \leqslant y \mid X \leqslant t)$ are both decreasing in t whatever $x, y \in \mathbb{R}$. These tail monotonicity conditions, jointly referred to as left-tail decreasingness (LTD), imply that the pair (X, Y) satisfies the concept of positive quadrant dependence (PQD). From [10], this means that for all $x, y \in \mathbb{R}$,

$$\Pr(X \leqslant x, Y \leqslant y) \geqslant \Pr(X \leqslant x)\Pr(Y \leqslant y).$$

Both PQD and LTD can be stated in terms of the underlying copula only. As shown, e.g., in [2, Chapter 5], PQD holds if and only if $C(u, v) \geqslant uv$ for all $u, v \in [0, 1]$, while the LTD property is verified if for almost all $u, v \in (0, 1)$,

$$\dot{C}_1(u, v) \leqslant \frac{C(u, v)}{u}, \qquad \dot{C}_2(u, v) \leqslant \frac{C(u, v)}{v}. \tag{1}$$

Many bivariate models with positive dependence meet Conditions (1), including the bivariate Normal, Beta, Gamma, Student and Fisher distributions. Other examples are provided by the Cook–Johnson bivariate Pareto, Burr and logistic distributions, Gumbel's bivariate exponential and logistic distributions, the Ali–Mikhail–Haq bivariate logistics, the Clayton, Frank, Plackett and Raftery families of copulas.

As it turns out, the LTD property implies a dominance relation between the asymptotic covariance functions of the empirical processes $\mathbb{C}$ and $\hat{\mathbb{C}}$. A formal statement of this fact is given below and proved in the Appendix.

Proposition 1. *Suppose that C is an LTD copula whose partial derivatives $\dot{C}_1$ and $\dot{C}_2$ exist everywhere and are continuous on $(0, 1)^2$. Then for all $u, v, s, t \in [0, 1]$,*

$$\mathrm{cov}\{\hat{\mathbb{C}}(u, v), \hat{\mathbb{C}}(s, t)\} \leqslant \mathrm{cov}\{\mathbb{C}(u, v), \mathbb{C}(s, t)\}. \tag{2}$$

Inequality (2) seems to have been intuited in [11] in the context of copula density estimation; a heuristic explanation was offered, but a formal result was neither stated nor proved. Proposition 4.2 in [12] is also a forerunner of Proposition 1 in the case of extreme-value copulas. As shown in [13], the latter are monotone regression dependent in the sense of [10]; this concept of dependence is stronger than the LTD property.