No. 2010-120

SUPEREFFICIENT ESTIMATION OF THE MARGINALS BY EXPLOITING KNOWLEDGE ON THE COPULA

By John H.J. Einmahl, Ramon van den Akker

November 2010

ISSN 0924-7815

Superefficient estimation of the marginals by exploiting knowledge on the copula

John H.J. Einmahl[a,*], Ramon van den Akker[a,**]

[a]*Department of Econometrics & OR and CentER, Tilburg University*

Abstract

We consider the problem of estimating the marginals in case there is knowledge on the copula. If the copula is smooth, it is known that it is possible to improve on the empirical distribution functions: optimal estimators still have rate of convergence $n^{-1/2}$, but a smaller asymptotic variance. In this paper we show that smoothness assumptions on the copula are necessary: we construct both a (non-smooth) copula and, exploiting the information our copula provides, estimators of the marginals with rate of convergence $\log n/n$.

KEYWORDS: copula, estimation of marginals, superefficient estimation

JEL CODES: C13, C14

1. Introduction

Suppose one observes a random sample from a bivariate distribution. By Sklar's theorem, see, e.g., Nelsen (1999), the distribution function is determined by its copula and the marginal distributions. In semiparametric copula models it is assumed that the copula depends on a Euclidean parameter and, apart from (absolute) continuity, no assumptions are imposed on the marginals. The study of efficient estimation for semiparametric copula models originated in Klaassen and Wellner (1997), which focused on efficient estimation of the copula parameter and noted that exploiting the knowledge on the copula may help to improve on the marginal empirical distribution functions. Following the setup in Klaassen and Wellner (1997), Chen, Fan, and Tsyrennikov (2006), Van den Akker (2007, Chapter 5), and Segers, Van den Akker, and Werker (2010) provide efficient estimators of the marginals, incorporating the information the copula provides, with the standard rate of convergence $n^{-1/2}$ and a limiting distribution that has less spread than the limiting distribution of the empirical distribution functions.

In those models smoothness assumptions on the copula are imposed. This paper shows that in absence of smoothness superefficient estimation of the marginals is possible. To

*Corresponding author, PO Box 90153, NL-5000 LE Tilburg, The Netherlands, Tel: +31134668208, Fax: +31134663280.

**PO Box 90153, NL-5000 LE Tilburg, The Netherlands, Tel: +31134663151, Fax: +31134663280.

Email addresses: j.h.j.einmahl@uvt.nl (John H.J. Einmahl), r.vdnakker@uvt.nl (Ramon van den Akker)

November 9, 2010

this end, we construct, in Section 2, a specific copula. In Section 3 we construct an estimator of the marginals that exploits the information our copula provides, and show that its rate of convergence is $\log n/n$. Our copula is a 'best copula' in the sense that $\log n/n$ is the best possible rate of convergence.

2. The copula

In this section we define our copula. To this end we introduce independent Bernoulli variables $(B_k)_{k\in\mathbb{N}}$ with success probability $1/2$, and define Bernoulli variables $(\tilde{B}_k)_{k\in\mathbb{N}}$ by $\tilde{B}_k = B_k$ for k odd and $\tilde{B}_k = 1 - B_k$ for k even. Using these Bernoulli sequences we introduce the random pair (U, V) by:

$$U = \sum_{k=1}^{\infty} \frac{B_k}{2^k}, \text{ and } V = \sum_{k=1}^{\infty} \frac{\tilde{B}_k}{2^k}.$$

Hence V is a one-to-one function of U and the inverse is the same function. Note that U and V are uniformly distributed on $[0, 1]$. The joint distribution of (U, V) thus defines a copula, which we will denote by C. This copula can be interpreted as an 'infinite shuffle of min' (see Mikusinski, Sherwood, and Taylor (1992) for shuffles of min).

We provide a second construction of C that might be more intuitive and allows us to introduce notation that is needed in the remainder of the paper. Define, for $k \in \mathbb{N}$ and $p, q = 1, \ldots, 2^k$, the sets $A^{(k)}_{p,q} = [(p-1)2^{-k}, p2^{-k}) \times [(q-1)2^{-k}, q2^{-k})$. Next, we define, for $k \in \mathbb{N}$ and $p = 1, \ldots, 2^k$, indices $q^{(k)}(p)$ as follows. For $k = 1$ we set $q^{(1)}(1) = 1$ and $q^{(1)}(2) = 2$. For $k \geq 2$ we set, for $p = 1, \ldots, 2^{k-1}$,

$$q^{(k)}(2p) = \begin{cases} 2q^{(k-1)}(p), & k \text{ odd;} \\ 2q^{(k-1)}(p) - 1, & k \text{ even,} \end{cases} \quad \text{and } q^{(k)}(2p-1) = \begin{cases} 2q^{(k-1)}(p) - 1, & k \text{ odd;} \\ 2q^{(k-1)}(p), & k \text{ even.} \end{cases}$$

Next we introduce, for $k \in \mathbb{N}$, $\mathcal{S}_k = \cup_{p=1}^{2^k} A^{(k)}_{p,q^{(k)}(p)}$; see Figure 1 for an illustration. Now

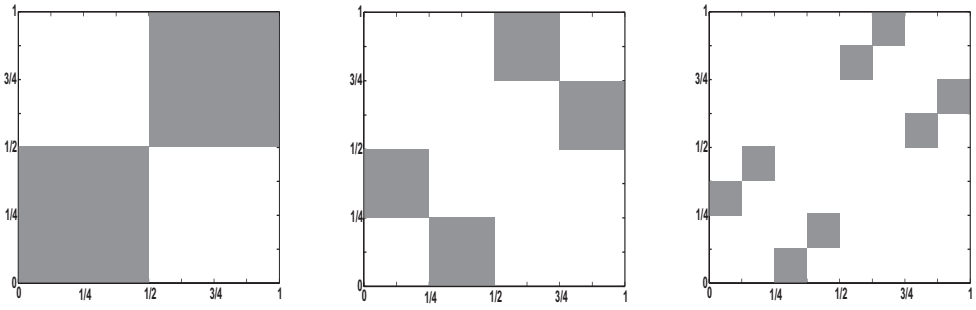

Figure 1: The support $\mathcal{S}_k$ of the copula C_k for $k = 1, 2, 3$.

we are able to introduce, for $k \in \mathbb{N}$, random variables $(U^{(k)}, V^{(k)})$ that are uniformly

distributed on $\mathcal{S}_k$ (the density equals 2^k). Note that $U^{(k)}$ and $V^{(k)}$ are uniformly distributed on $[0,1]$, so the law of $(U^{(k)}, V^{(k)})$ defines a copula C_k. It is easy to see that $C_k \xrightarrow{w} C$ as $k \to \infty$. In particular, we have, for all $k, m \in \mathbb{N}$ and all $p, q = 1, \ldots, 2^k$,

$$\mathbb{P}\left\{(U^{(k)}, V^{(k)}) \in A_{p,q}^{(k)}\right\} = \mathbb{P}\left\{(U^{(k+m)}, V^{(k+m)}) \in A_{p,q}^{(k)}\right\} = \mathbb{P}\left\{(U, V) \in A_{p,q}^{(k)}\right\},$$

and this probability equals 2^{-k} in case $q = q^{(k)}(p)$ and 0 in case $q \neq q^{(k)}(p)$.

3. The estimator and its limiting behavior

Available is a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a bivariate distribution function H which has C, as defined in Section 2, as copula. By Sklar's theorem we have, for all $(x, y) \in \mathbb{R}^2$, $H(x, y) = C(F(x), G(y))$, where F and G are the marginal distribution functions of X_1 and Y_1, respectively. The only assumption we impose on F and G is that they belong to $\mathcal{F}$, the set of continuous distribution functions on the real line.

We introduce our estimator of F via its quantile function. First, we define $Q_n(u)$ on the set $\{p2^{-k} | p = 0, \ldots, 2^k, k \geq 1\}$. Set $Q_n(0) = X_{1:n}$, $Q_n(1) = X_{n:n}$, and define $Q_n(p2^{-k})$ for $k \in \mathbb{N}$ and $p \in \{1, \ldots, 2^k - 1\}$ odd, recursively by (we adopt the usual convention $\max \emptyset = -\infty$):

$$Q_n\left(\frac{p}{2^k}\right) = \max\left\{Q_n\left(\frac{p-1}{2^k}\right), \tilde{Q}_n\left(\frac{p}{2^k}\right)\right\},$$

where

$$\tilde{Q}_n\left(\frac{p}{2^k}\right) = \max_{\substack{i:\, X_i \in \left(Q_n\left(\frac{p-1}{2^k}\right), Q_n\left(\frac{p+1}{2^k}\right)\right), \\ \exists X_j \in \left(X_i, Q_n\left(\frac{p+1}{2^k}\right)\right]}} \left\{ X_i \middle| \max_{j:\, X_j \in \left(Q_n\left(\frac{p-1}{2^k}\right), X_i\right]} Y_j < \min_{j:\, X_j \in \left(X_i, Q_n\left(\frac{p+1}{2^k}\right)\right]} Y_j \right\}$$

in case k is odd, and

$$\tilde{Q}_n\left(\frac{p}{2^k}\right) = \max_{\substack{i:\, X_i \in \left(Q_n\left(\frac{p-1}{2^k}\right), Q_n\left(\frac{p+1}{2^k}\right)\right), \\ \exists X_j \in \left(X_i, Q_n\left(\frac{p+1}{2^k}\right)\right]}} \left\{ X_i \middle| \min_{j:\, X_j \in \left(Q_n\left(\frac{p-1}{2^k}\right), X_i\right]} Y_j > \max_{j:\, X_j \in \left(X_i, Q_n\left(\frac{p+1}{2^k}\right)\right]} Y_j \right\}$$

in case k is even. Next, we extend the domain to $[0,1]$ by $Q_n(u) = \sup\{Q_n(p2^{-k}) | p2^{-k} \leq u\}$. As estimator of F we take the distribution function associated with Q_n. We denote this estimator by $\hat{F}_n$. Note that $\hat{F}_n$ can be written as $\hat{F}_n(x) = \sum_{i=1}^n p_i 1_{(-\infty, x]}(X_i)$, where the probability masses p_i only depend on the observations via the ranks (R_j^X, R_j^Y) of (X_j, Y_j), $j = 1, \ldots, n$.

The following theorem is the main result of this paper.

Theorem 3.1 *For $F, G \in \mathcal{F}$ we have ($\|\cdot\|_\infty$ denotes the sup-norm):*

$$\frac{1}{2} \leq \liminf_{n\to\infty} \frac{n}{\log n} \|\hat{F}_n - F\|_\infty \leq \limsup_{n\to\infty} \frac{n}{\log n} \|\hat{F}_n - F\|_\infty \leq 4 \text{ a.s.} \tag{1}$$

The theorem demonstrates that $\hat{F}_n$ is superefficient, i.e. the rate of convergence is $\log n/n$ instead of the usual rate $n^{-1/2}$.

Remark 1 *In the proof of Theorem 3.1 we exploit that any estimator $\tilde{F}_n$ of F that concentrates on $X_1, \ldots, X_n$ satisfies*

$$\liminf_{n\to\infty} \frac{n}{\log n} \|\tilde{F}_n - F\|_\infty \geq \frac{1}{2} \text{ a.s.} \tag{2}$$

This property implies that our estimator $\hat{F}_n$ achieves the best attainable rate of convergence $\log n/n$. As the bound (2) does not depend on the copula, our copula C can be interpreted as a 'best one' (in terms of rate of convergence).

Remark 2 *A natural question is whether $\mathbb{Z}_n = \left\{(\log n/n)(\hat{F}_n(x) - F(x)) \mid x \in \mathbb{R}\right\}$, seen as an element of $\ell^\infty(\mathbb{R})$, weakly converges (if so, the limit determines the limiting distribution of $(n/\log n)\|\hat{F}_n - F\|_\infty$ by an application of the continuous mapping theorem). The answer is negative. For $F = I$, where I denotes the distribution function of the Uniform$[0,1]$ distribution, the argument is as follows (the general case easily follows from the uniform case). Since $\hat{F}_n$ concentrates on the observations and, as we exploit in the proof of Theorem 3.1, the maximal spacing Δ_n of n i.i.d. draws from the Uniform$[0,1]$ distribution satisfies $(\log n/n)\Delta_n \to 1$ a.s. we have, for any $\eta \in (0,1)$, $\epsilon \in (0, 1/2)$ and any finite partition $\cup_{i=1}^k T_i$ of $[0,1]$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_i \sup_{u,u' \in T_i} \frac{n}{\log n} \left|\hat{F}_n(u) - \hat{F}_n(u') - (u - u')\right| > \epsilon\right) = 1 > \eta,$$

which shows that $\mathbb{Z}_n$ is not tight.

As an illustration, Figure 2 presents a realization of our estimator and the empirical distribution function F_n^{edf} for $n = 100$, $F = \Phi$, the standard normal distribution function, and $G \in \mathcal{F}$, and Figure 3 presents the centered versions of the estimates.

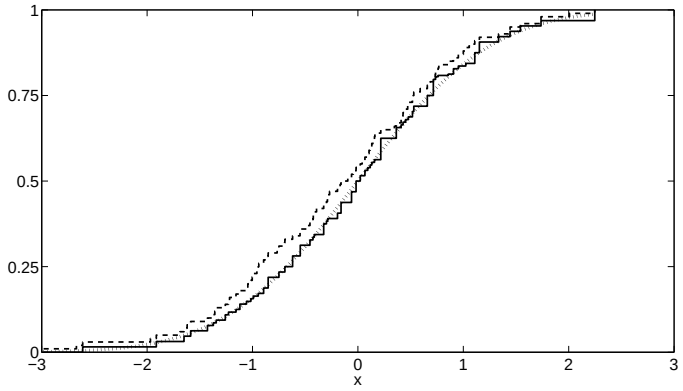

Figure 2: Realization of $\hat{F}_n$ (solid) and F_n^{edf} (dashed) for $n = 100$, $F = \Phi$ (dotted), and $G \in \mathcal{F}$.

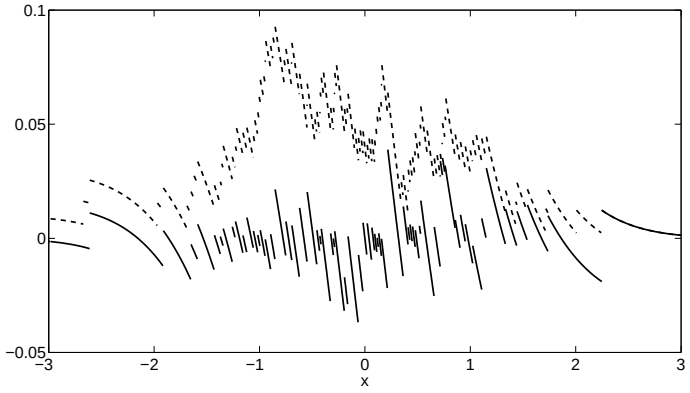

Figure 3: Realization of $\hat{F}_n - F$ (solid) and $F_n^{edf} - F$ (dashed) for $n = 100$, $F = \Phi$, and $G \in \mathcal{F}$.

PROOF OF THEOREM 3.1: Introduce $U_i = F(X_i)$ and $V_i = G(Y_i)$, and recall that monotone transformations of the marginals do not change the copula. Let $\hat{F}_n^U$ denote the distribution function resulting from computing $\hat{F}_n$ from $(U_i, V_i)_{i=1}^n$ instead of $(X_i, Y_i)_{i=1}^n$. As $\hat{F}_n(x) = \hat{F}_n^U(F(x))$ a.s. we have $\|\hat{F}_n - F\|_\infty = \|\hat{F}_n^U - I\|_\infty$ a.s., which shows that it suffices to prove (1) for $F = G = I$. To stress that we consider uniform marginals we denote the observations by (U_i, V_i) in the remainder of the proof.

As the probability of a tie in $(U_i)_{i=1}^\infty$ or $(V_i)_{i=1}^\infty$ equals zero, we throughout work on the event that there are no ties. Let $\Delta_n = \max_{i=1,\ldots,n+1} |U_{i:n} - U_{i-1:n}|$, with $U_{0:n} = 0$ and $U_{n+1:n} = 1$, denote the maximal spacing of $U_1, \ldots, U_n$. Observe that any estimator $\tilde{F}_n$ of I of the form $\tilde{F}_n(u) = \sum_{i=1}^n \tilde{p}_i 1_{[0,u]}(U_i)$ satisfies $\|\tilde{F}_n - I\|_\infty \geq \Delta_n/2$. Observe that $\|F_n - I\|_\infty = \|Q_n - I\|_\infty$. As it is well-known, see, e.g., Slud (1978), that $(n/\log n)\Delta_n \to 1$ a.s., we see that the theorem holds once we establish the bound $\|Q_n - I\|_\infty \leq 4\Delta_n$. As $|Q_n(0) - 0| \leq \Delta_n$ and $|Q_n(1) - 1| \leq \Delta_n$ we have to prove

$$|Q_n(u) - u| \leq 4\Delta_n, \quad \text{for all } u \in (0,1). \tag{3}$$

Denote $\mathcal{U}_n = \{U_1, \ldots, U_n\}$ and introduce the random variable

$$\tilde{K} = \max\left\{k \in \mathbb{N} \mid \forall p = 1, \ldots, 2^{k+1} : \left(\frac{p-1}{2^{k+1}}, \frac{p}{2^{k+1}}\right] \cap \mathcal{U}_n \neq \emptyset\right\}.$$

In case $\tilde{K} = -\infty$ we have $\Delta_n \geq 1/4$ and (3) trivially holds, so we only need to consider $\tilde{K} \geq 1$. We will prove, for $k = 1, \ldots, \tilde{K}$ and $p = 1, \ldots, 2^k - 1$ odd,

$$Q_n\left(\frac{p}{2^k}\right) = \max_{i=1,\ldots,n}\left\{U_i | U_i < \frac{p}{2^k}\right\}. \tag{4}$$

Before we prove (4) we show that (4) implies (3). From (4) it is immediate that (3) holds for $u \in \{p2^{-\tilde{K}} | p = 1, \ldots, 2^{\tilde{K}} - 1\}$; to be precise, we have, for $p = 1, \ldots, 2^{\tilde{K}} - 1$,

$$\left|Q_n\left(\frac{p}{2^{\tilde{K}}}\right) - \frac{p}{2^{\tilde{K}}}\right| \leq \Delta_n.$$

Let $K^\star = \tilde{K} + 1$ and note that the intervals $((p-1)2^{-K^\star}, p2^{-K^\star}]$ and $(p2^{-K^\star}, (p+1)2^{-K^\star}]$ both contain at least one observation. The definition of Q_n and (4) now yield $Q_n(p2^{-K^*}) \in [(p-1)2^{-K^*}, (p+1)2^{-K^*})$ and the definition of $\tilde{K}$ implies $\Delta_n \geq 2^{-(K^\star+1)}$. A combination of these observations immediately yields

$$\left| Q_n\left(\frac{p}{2^{K^*}}\right) - \frac{p}{2^{K^*}} \right| \leq 2\Delta_n,$$

which shows that (3) holds for all $u \in \{p2^{-K^\star} | p = 1, \ldots, 2^{K^\star} - 1\}$. Finally, we consider $u \in (0,1)$ with $u2^{K^\star} \notin \mathbb{N}$. Let p^* such that $u \in (p^*2^{-K^*}, (p^*+1)2^{-K^*})$. We easily obtain the bound

$$-4\Delta_n \leq Q_n\left(\frac{p^*}{2^{K^*}}\right) - \frac{p^*}{2^{K^*}} - \frac{1}{2^{K^*}} \leq Q_n(u) - u \leq Q_n\left(\frac{p^*+1}{2^{K^*}}\right) - \frac{p^*+1}{2^{K^*}} + \frac{1}{2^{K^*}} \leq 4\Delta_n.$$

We conclude that (3) indeed holds.

We conclude the proof by establishing (4). We start with $k = p = 1$. Since the squares $A^{(1)}_{1,1}$ and $A^{(1)}_{2,2}$ both contain at least two observations and $A^{(1)}_{2,2}$ is 'north' to $A^{(1)}_{1,1}$, it follows from the definition of $Q_n(1/2)$ that $Q_n(1/2) \geq \max_i\{U_i | U_i < 1/2\}$. As the square $A^{(2)}_{3,4}$ is 'north' to $A^{(2)}_{4,3}$ and both squares contain at least one observation it is also immediate that $Q_n(1/2) < \min_i\{U_i | U_i \geq 1/2\}$. Hence (4) indeed holds for $k = p = 1$. Suppose that we have shown (4) to hold for $k = 1, \ldots, K-1$, with $K \leq \tilde{K}$. We show that then (4) also holds for $k = K$. We have to discuss the cases K even and K odd separately. As the arguments are similar, we only discuss the case K odd. For p odd we obtain from the induction hypothesis that all observations that are relevant for $Q_n(p2^{-K})$, i.e. the observations U_i that belong to the interval $(Q_n((p-1)2^{-K}), Q_n((p+1)2^{-K})]$, correspond to to observations (U_i, V_i) that fall in the sets $A^{(K)}_{p,q^{(K)}(p)}$ and $A^{(K)}_{p+1,q^{(K)}(p+1)}$. As $K \leq \tilde{K}$ both squares contain at least one observation. As K is odd $A^{(K)}_{p+1,q^{(K)}(p+1)}$ is 'north' to $A^{(K)}_{p,q^{(K)}(p)}$. It follows that $Q_n(p2^{-K}) \geq \max_i\{U_i | U_i < p2^{-k}\}$. The mass that C assigns to the set $A^{(K)}_{p+1,q^{(K)}(p+1)}$ concentrates in the two subsets $A^{(K+1)}_{2p+1,q^{(K+1)}(2p+1)}$ and $A^{(K+1)}_{2(p+1),q^{(K+1)}(2(p+1))}$, and both sets contain at least one observation. As $K+1$ is even the set $A^{(K+1)}_{2(p+1),q^{(K+1)}(2(p+1))}$ is 'south' to $A^{(K+1)}_{2p+1,q^{(K+1)}(2p+1)}$. This easily yields $Q_n(p2^{-K}) < \min_i\{U_i | U_i \geq p2^{-k}\}$. We conclude that (4) holds for $k = K$ as well, which concludes the induction argument. □

References

Chen, X., Y. Fan, and V. Tsyrennikov (2006). Efficient estimation of semiparametric multivariate copula models, *Journal of the American Statistical Association* 101, 1228–1240.

Klaassen, C.A.J. and J.A. Wellner (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable, *Bernoulli* 3, 55–77.

Mikusinski, P., H. Sherwood, and M. Taylor (1992). Shuffles of min, *Stochastica* XIII, 61–74.

Nelsen, R. (1999). An introduction to copulas (1st ed.). New York: Springer- Verlag.

Segers, J.J.J., R. van den Akker, and B.J.M. Werker (2010). Improving upon the marginal empirical distribution functions when the copula is known, Working paper.

Slud, E. (1978). Entropy and maximal spacings for random partitions, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 41, 341–352.

Van den Akker, R. (2007). Integer-valued time series, PhD thesis, CentER dissertation series 197, Tilburg University, available at: http://arno.uvt.nl/show.cgi?did=306632.