

ON A WAVELET-BASED METHOD FOR ESTIMATING THE COPULA FUNCTION

A. GANNOUN and N. HOSSEINIOUN

Copula models are becoming increasingly popular for modelling dependencies between random variables. The range of their recent applications include such fields as analysis of extremes in financial assets and returns, failure of paired organs in health science, and human mortality in insurance. The aim of this work is to establish an upper bound on L^p -losses ($2 \leq p < \infty$) of the linear wavelet-based estimator for copula function when the copula function is assumed to be bounded.

AMS 2010 Subject Classification: 62N02.

Key words: copulas, nonparametric estimation, rank statistics, wavelets.

1. INTRODUCTION

In the recent years, the *copula* models became an increasingly popular tool for modeling dependencies between random variables, especially in such fields as biostatistics, actuarial science, and finance. One of the advantages of copula models is their relative mathematical simplicity. Another advantage is the possibility to build a variety of dependence structures based on existing parametric or non-parametric models of the marginal distributions. The copulas model has been extensively studied in a parametrical framework for the distribution function C . Large classes of copulas, such as the elliptic family, which contains the Gaussian copula and the Student copula, and the Archimedian family, which contains the Gumbel copula, the Clayton copula and the Frank copulas, have been identified.

Mainly, people have worked in two directions. Firstly, an important activity has concerned the modelling in view to find new copulas and methodologies to simulate data coming from these new copulas. Secondly, usual statistical inference (estimation of the parameters, goodness-of-fit test, etc) has been developed using the copulas. As usual, the nonparametric point of view is useful when no a priori model of the phenomenon is specified. For the practitioners, the non-parametrical estimators could be seen as a benchmark allowing to specify the model, comparing with the available parametrical families. This explains the success of the nonparametric estimator of the copula.

Suppose that the relation between variables X and Y is of interest and assume for simplicity that both of them are real-valued with rank statistics R and S . Let $F(x) = Pr(X \leq x)$ and $G(y) = Pr(Y \leq y)$ be their cumulative distribution functions. Following Sklar [19], the joint distribution function of the pair (X, Y) on (x, y) may be expressed in the form

$$(1.1) \quad H(x, y) = C(F(x), G(y))$$

for some distribution function C whose margins are uniform on the interval $(0, 1)$. When F and G are continuous, C is unique and coincides with the distribution function of the pair $(U, V) = (F(X), G(Y))$. In practice, H is unknown. A copula model for the pair (X, Y) can then be constructed by assuming that F, G and C belong to specific classes of distributions. An advantage of this approach is that the copula C , which characterizes the dependence between X and Y , can be chosen separately from the marginal models.

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from the unknown distribution H . Denote by F_n and G_n the empirical distributions associated with F and G . A first step in selecting an appropriate class of copulas consists of plotting the pairs

$$(1.2) \quad \left(\frac{R_i}{n}, \frac{S_i}{n} \right) = (F_n(X_i), G_n(Y_i)), \quad i = \{1, \dots, n\},$$

here, R_i is the rank of X_i among X_1, \dots, X_n and S_i is the rank of Y_i among Y_1, \dots, Y_n . The motivation behind this graphical approach is that the pseudo observations $(\frac{R_i}{n}, \frac{S_i}{n})$ are close substitutes to the unobservable pairs $(U_i, V_i) = (F(X_i), G(Y_i))$, forming a random sample from C . We denote by c the density of C where

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}, \quad u, v \in (0, 1).$$

This density is assumed to exist and to be square-integrable in the sequel.

Nonparametric estimators of copula densities have been suggested by Gijbels and Mielniczuk [13] and Fermanian and Scaillet [11], who used kernel methods, Sancetta [16] and Sancetta and Satchell [17], who employed techniques based on Bernstein polynomials. Biau and Wegkamp [5] proposed estimating the copula density through a minimum distance criterion. Their estimator enjoys good properties but its computation entails non-trivial implementation issues that are left unaddressed. Estimation of a class of copula-based semiparametric stationary Markov models discussed by Chen and Fan [7], characterized by nonparametric marginal distributions and parametric copula functions, while the copulas capture all the scale-free temporal dependence of the processes. Dearden, Fitzsimons and Goodman [10] illustrated both the usefulness of copulas as a statistical technique for modelling dependence in earning across the life-cycle, as well as contrast it with more traditional approaches for

modelling earnings dynamics that appear in the literature. Pseudo-likelihood estimator for copula function based on delicate empirical process theory has been developed by Breaks and Keilegom [6] with the asymptotic normality of the proposed estimator. Morettin, Toloi, Chiann and Miranda [14] presented a brief review of the methods often used for copula estimation in the context of independent, identically distributed random variables and discussed their use for time series data. Finally, Genest, Masiello and Tribouley [12] proposed a wavelet-based estimator for a copula function and the estimation procedure is shown to be optimal in the minimax sense on a large functional class of regular copula densities. Their approach is illustrated with actuarial and financial data.

The rest of this paper is organized as follows. We provide a brief introduction to copulas and discuss the class of *wavelet*-based estimator for copula function, defined by Genest, Masiello and Tribouley [12] in Section 2. We also discuss the class of Besov spaces as functional spaces. Section 3 provides some Lemmas which will be used throughout our main results, while the L_p -losses for the proposed estimator is stated in Section 4.

2. WAVELET-BASED ESTIMATORS

First, for any univariate function $h(\cdot)$, we denote by $h_{jk}(\cdot)$ the function $2^{j/2}h(2^j \cdot - k)$ where $j \in \mathbb{N}$ and $k \in \mathbb{Z}$. Now, let ϕ and ψ be respectively a scaling function and an associated wavelet function. We assume that these functions are compactly supported on $[0, L]$ for some $L \geq 1$ and form an orthonormal basis, and we have the representation at scale j_0

$$h(t) = \sum_{k \in \mathbb{Z}} \alpha_{j_0 k} \phi_{j_0 k}(t) + \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}(t).$$

This construction can be easily extended to bivariate functions $h(\cdot, \cdot)$. We build a bivariate wavelet basis as follows

$$\begin{aligned} \phi_{j,k}(x, y) &= \phi_{jk_1}(x) \phi_{jk_2}(y), \\ \psi_{j,k}^\epsilon(x, y) &= \prod_{m=1}^2 \phi_{jk_m}^{1-\epsilon_m}(x) \psi_{jk_m}^{\epsilon_m}(y), \end{aligned}$$

for all $k = (k_1, k_2) \in \mathbb{Z}^2$ and $\epsilon = (\epsilon_1, \epsilon_2) \in S_2 = \{(0, 1), (1, 0), (1, 1)\}$. For any $j_0 \in \mathbb{N}$, the set $\{\phi_{j_0,k}, \psi_{j,k}^\epsilon \mid j \geq j_0, k \in \mathbb{Z}^2, \epsilon \in S_2\}$ is an orthonormal basis of $L_2(\mathbb{R}^2)$. The expansion of the analysed function on the wavelet basis splits into the trend at the level j_0 and the sum of the “details” for all the larger levels $j, j \geq j_0$, is given by

$$h(x, y) = h_{j_0}(x, y) + D_{j_0} h(x, y), \quad x, y \in \mathbb{R},$$

where

$$h_{j_0}(x, y) = \sum_{k \in \mathbb{Z}^2} \alpha_{j_0 k} \phi_{j_0 k}(x, y),$$

is a trend (or approximation) and $D_{j_0} h(x, y)$ is represented as follows

$$D_{j_0} h(x, y) = \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}^2} \sum_{\epsilon \in S_2} \beta_{j k}^\epsilon \psi_{j k}^\epsilon(x, y).$$

Assuming that the copula density c belongs to L_2 , we present wavelet procedures of its estimation. Motivated by the wavelet expansion, we first estimate the coefficients of the copula density on the wavelet basis. Given a bounded copula density c , one can then expand it in the form (2.1) with

$$\alpha_{j_0 k} = \int_{(0,1)^2} c(u, v) \phi_{j_0 k}(u, v) du dv, \quad k \in \mathbb{Z}^2.$$

If the marginal distributions F and G were known, a natural (moment-based) estimator of $\alpha_{j_0 k}$ would then be given by

$$(2.1) \quad \hat{\alpha}_{j_0 k} = \frac{1}{n} \sum_{i=1}^n \phi_{j_0 k}(F(X_i), G(Y_i)).$$

The wavelet-based estimator of c is given by

$$(2.2) \quad \hat{c}_{j_0}(u, v) = \sum_{k \in \mathbb{Z}^2} \hat{\alpha}_{j_0 k} \phi_{j_0 k}(u, v), \quad (u, v) \in (0, 1).$$

When F and G are unknown, a nonparametric analogue is obtained by Genest, Masiello and Tribouley [12], upon replacing F and G by their empirical counterparts, F_n and G_n . In view of relation (1.2), the estimator is thus rank-based, viz.

$$\tilde{\alpha}_{j_0 k} = \frac{1}{n} \sum_{i=1}^n \phi_{j_0 k}(F_n(X_i), G_n(Y_i)) = \frac{1}{n} \sum_{i=1}^n \phi_{j_0 k} \left(\frac{R_i}{n}, \frac{S_i}{n} \right).$$

The linear wavelet-based estimator of c is then given by

$$(2.3) \quad \tilde{c}_{j_0}(u, v) = \sum_{k \in \mathbb{Z}^2} \tilde{\alpha}_{j_0 k} \phi_{j_0 k}(u, v), \quad (u, v) \in (0, 1).$$

Note that \tilde{c}_{j_0} may sometimes be negative on parts of its domain and fail to integrate to 1. If in applications, an intrinsic copula density estimate is deemed necessary, it can be derived from \tilde{c}_{j_0} by truncation and normalization. For recent developments and applications on wavelets see Antoniadis [1].

The purpose of this section is to study the performance of \tilde{c}_{j_0} as an estimator of the underlying copula density c . First, let us give the definition of Besov spaces in terms of wavelet coefficients. This is convenient as it gives a description in terms of sequence spaces. Then, we state two Lemmas, needed

to establish the main results. From now on we denote K any constant that may change from one line to another, which does not depend on j , k and n , but depends on the wavelet basis and on $\|c\|_\infty = \sup_{(u,v) \in (0,1)} |c(u,v)|$ and $\|c\|_2 = \int c(u,v)^2 dudv$.

Since we deal with wavelet methods, it is very natural to consider Besov spaces as functional spaces because they are characterized in term of wavelet coefficients as follows. Besov spaces depend on three parameters $s > 0$, $1 < p < \infty$ and $1 < q < \infty$ and are denoted by $B_{s,p,q}$. Let $f \in L^2(\mathbb{R}^2)$ and let s be smaller than r (wavelet regularity). Using Tribouley [20], $f \in B_{s,p,q}$ if and only if

$$\|f\|_{s,p,q} = \|\alpha_0\|_p + \left(\sum_{j \geq 0} \left(2^{j(s+d/2+d/p)} \|\beta_j\|_p \right)^q \right)^{1/q} < \infty,$$

where

$$\|\beta_j\|_p = \left(\sum_{k \in \mathbb{Z}^2} \sum_{\epsilon \in S_2} |\beta_{j,k}^\epsilon|^p \right)^{1/p}.$$

We assume that the copula function c belongs to the Besov space. More precisely, it is assumed to be bounded. There are other definitions and characterizations of Besov spaces; see Peetre [15] and Bergh and Loftstorm [4] for more discussions.

3. INTERMEDIATE RESULTS

LEMMA 3.1 (Rosenthal's inequality). *Let $p \geq 2$ and $\{X_1, \dots, X_n\}$ be independent random variables such that $E(X_i) = 0$, $\|X_i\|_\infty < M$. Then there exists a constant $C(p)$, depends on p , such that*

$$E \left(\left| \sum_{i=1}^n X_i \right|^p \right) \leq C(p) \left\{ M^{p-2} \sum_{i=1}^n E|X_i|^p + \left(\sum_{i=1}^n E(X_i^2) \right)^{p/2} \right\}.$$

In the theory of probability, the Glivenko-Cantelli theorem determines the asymptotic behaviour of the empirical distribution function as the number of i.i.d. observations grows, strengthening this result by proving uniform convergence of F_n to F . This uniform convergence of more general empirical measures becomes an important property of the Glivenko-Cantelli classes of functions or sets.

LEMMA 3.2 (Glivenko-Cantelli's inequality). *Assume that X_1, \dots, X_n are i.i.d. random variables in R with common cumulative distribution function*

$F(x)$. The empirical distribution function is defined by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i),$$

where I_A is the indicator function, then

$$\|F_n - F\|_\infty = \sup_{x \in R} |F_n(x) - F(x)| \rightarrow 0 \quad a.s.$$

Proof. See for example Van der Vaart [21]. \square

THEOREM 3.1. *Let $p \geq 2$ and ϕ be a scaling function having m derivatives and for arbitrary resolution level $j_0 \in \mathbb{N}$, let \hat{c}_{j_0} and \tilde{c}_{j_0} be the estimators of a copula density c defined by (2.3) and (2.4). Then there exists a constant K such that for given $(u, v) \in (0, 1)^2$ and any level j_0 satisfying $2^{j_0} \leq n$,*

$$(3.1) \quad E\|\hat{c}_{j_0} - c_{j_0}\|_p \leq K \frac{2^{3j_0}}{n},$$

$$(3.2) \quad E\|\hat{c}_{j_0} - \tilde{c}_{j_0}\|_p = o_p(1).$$

Proof. The proof of equation (3.1) requires the evaluation of

$$\hat{\alpha}_{j_0 k} - \alpha_{j_0 k} = \frac{1}{n} \sum_{i=1}^n (\phi_{j_0 k}(F(x_i), G(y_i)) - \alpha_{j_0 k}).$$

Using the definitions (2.1) and (2.3), one may easily have

$$(3.3) \quad \begin{aligned} E\|\hat{c}_{j_0} - c_{j_0}\|_p &= E\left\| \sum_{k \in \mathbb{Z}^2} (\hat{\alpha}_{j_0 k} - \alpha_{j_0 k}) \phi_{j_0 k}(x, y) \right\|_p \\ &\leq \sum_{k \in \mathbb{Z}^2} E\|(\hat{\alpha}_{j_0 k} - \alpha_{j_0 k})\|_p \|\phi(u)\|_\infty \|\phi(v)\|_\infty. \end{aligned}$$

Using the definition (2.2), we obtain

$$\hat{\alpha}_{j_0 k} - \alpha_{j_0 k} = \frac{1}{n} \sum_{i=1}^n (\phi_{j_0 k}(F(x_i), G(y_i)) - \alpha_{j_0 k}).$$

Let us introduce the following notation

$$\xi_i = \phi_{j_0 k}(F(x_i), G(y_i)) - \alpha_{j_0 k}.$$

Since $E\xi_i = 0$ and $E\xi_i^2 \leq 2^{j_0} \|\phi(u)\phi(v)\|_\infty \|c\|_\infty$,

$$\|\xi_i\|_\infty \leq 2^{j_0} \|\phi(u)\phi(v)\|_\infty.$$

Recall the Rosenthal's inequality in Lemma (3.1), we easily get

$$(3.4) \quad \left(\frac{1}{n} \sum_{i=1}^n E\xi_i \right)^{\dot{p}} \leq K n^{-\dot{p}} \left\{ (2^{j_0})^{\dot{p}-2} n 2^{j_0} + (n 2^{j_0})^{\dot{p}/2} \right\}.$$

Since the support of the scaling function is compact, there are at most 2^{2j} terms in the sums over k appearing in the right-hand terms of (3.3), hence by substituting (3.4) in (3.3), we easily conclude

$$\begin{aligned} E\|\hat{c}_{j_0} - c_{j_0}\|_{\dot{p}} &\leq K_1 2^{3j_0} (n^{1-\dot{p}} 2^{j_0(\dot{p}-1)} + n^{-\dot{p}/2} 2^{j_0\dot{p}/2}) \\ &= K_1 2^{2j_0} \left(\left(\frac{2^{j_0}}{n} \right)^{1-2/\dot{p}} \frac{2^{j_0}}{n} + \frac{2^{j_0}}{n} \right). \end{aligned}$$

But since $n \geq 2^{j_0}$ and $1 - 2/\dot{p} \geq 0$ imply $(\frac{2^{j_0}}{n})^{1-2/\dot{p}} \leq 1$,

$$E\|\hat{c}_{j_0} - c_{j_0}\|_{\dot{p}} \leq K_2 \frac{2^{3j_0}}{n}.$$

Now for the next equation, using the definition of Autin, Le Pennec, Tribouley [3],

$$(3.5) \quad E\|\hat{c}_{j_0} - \tilde{c}_{j_0}\|_{\dot{p}} = E\left\| \sum_{m=1}^2 C_m^2 \lambda_{m,j} \right\|_{\dot{p}},$$

where

$$\begin{aligned} \lambda_{1,j} &= \psi_{j,k_1}^\epsilon(F(X_i))(\delta_j(X_i)), \\ \lambda_{2,j} &= \psi_{j,K_1}^\epsilon(F(X_i))\psi_{j,k_2}^\epsilon(F(Y_i))\delta_j(X_i)\delta_j(Y_i), \end{aligned}$$

with

$$\delta_j(\cdot) = \psi_{j,k_m}(\hat{F}_m(\cdot)) - \psi_{j,k_m}(F_m(\cdot)).$$

Following along the lines of Autin, Le Pennec, Tribouley [3], one may easily have

$$(3.6) \quad |\lambda_{m,j}| \leq \sum_{m'=0}^m 2^{j(m+\hat{m}/2)} (\max \Delta(x_i)^{m+\hat{m}}) \|\psi\|_\infty^{4-\hat{m}} 2^{j/2(2-\hat{m})}, \quad m = 1, 2$$

with $\Delta(\cdot) = \hat{F}_m(\cdot) - F_m(\cdot)$. Using (3.6) in (3.5) by using Lemma 3.2, since the support of scaling functions is compact, we conclude the second result. \square

4. MAIN RESULTS

Now we are in a position to provide an upper bound on $L_{\dot{p}}$ -losses for the mentioned estimator, similar to the one obtained in the case of curve estimation by Antoniadis [1] and Doosti et al. [9] for negatively dependent random variables. Suppose $c \in B_{s,p,q}$, with $s \geq 1/p$, $p \geq 1$.

THEOREM 4.1. *For $\dot{p} \geq \max(2, p)$, there exists a constant K , such that*

$$E\|\tilde{c}_{j_0} - c\|_{\dot{p}}^2 \leq K(2^{6j_0}n^{-2} + 2^{-2j_0(\dot{s}+2/\dot{p}+1)}),$$

where

$$\dot{s} = s + 1/\dot{p} - 1/p.$$

Proof. First we decompose $E\|\tilde{c}_{j_0} - c\|_{\dot{p}}^2$ as

$$E\|\tilde{c}_{j_0} - c\|_{\dot{p}}^2 \leq 2E\|\tilde{c}_{j_0} - c_{j_0}\|_{\dot{p}}^2 + 2\left\|\sum_{j \geq j_0} \sum_{k, \varepsilon} \beta_{j,k}^\varepsilon \psi_{j,k}^\varepsilon\right\|_{\dot{p}}^2 = 2(T_1 + T_2).$$

Now, we want to find upper bounds for T_1 and T_2 , separately. Note that

$$\sqrt{T_1} \leq E\|\tilde{c}_{j_0} - \hat{c}_{j_0}\|_{\dot{p}} + E\|\hat{c}_{j_0} - c_{j_0}\|_{\dot{p}} = T_{11} + T_{12}.$$

Using Theorem 3.1, one may easily bound the terms in T_1 . Next, we have

$$\begin{aligned} \sqrt{T_2} &\leq E\left(\sum_{j \geq j_0} \left\|\sum_{k, \varepsilon} \beta_{j,k}^\varepsilon \psi_{j,k}^\varepsilon\right\|_{\dot{p}}\right) = \\ &= E\left(\sum_{j \geq j_0} 2^{j(\dot{s}+2/\dot{p}+1)} \left\|\sum_{k, \varepsilon} \beta_{j,k}^\varepsilon \psi_{j,k}^\varepsilon\right\|_{\dot{p}} 2^{-j(\dot{s}+2/\dot{p}+1)}\right). \end{aligned}$$

Using Holder inequality, with $\frac{1}{q} + \frac{1}{\dot{q}} = 1$, the above equation implies

$$T_2 \leq K\|c\|_{\dot{s}, \dot{p}, q} \left\{ \sum_{j \geq j_0} 2^{-j(\dot{s}+2/\dot{p}+1)\dot{q}} \right\}^{2/\dot{q}}.$$

Now, using the continuous Sobolev injection in Donoho et al. [8], we conclude that $B_{p,q}^s \subset B_{\dot{p},q}^{\dot{s}}$, hence one gets

$$T_2 \leq \|c\|_{s,p,q} 2^{-2j_0(\dot{s}+2/\dot{p}+1)},$$

and in turn, we get

$$T_2 \leq K 2^{-2j_0(\dot{s}+2/\dot{p}+1)}. \quad \square$$

REFERENCES

- [1] A. Antoniadis and M.C. Gregoire, *Wavelet methods for curve estimation*. J. Amer. Statist. Assoc. **89** (1994), 1340–1353.
- [2] A. Antoniadis, *Wavelet methods in statistics: some recent developments and their applications*. Statist. Surv. **1** (2007), 16–55.
- [3] F. Autin, E. Le Pennec and K. Tribouley, *Thresholding method to estimate the copula density*. J. Multivariate Anal. **101** (2010), 200–222.
- [4] J. Bergh and J. Loftstorm, *Interpolation spaces: An introduction*. Springer, 1976.
- [5] G. Biau and M.H. Wegkamp, *A note on minimum distance estimation of copula densities*. Statist. Probab. Lett. **73** (2006), 105–114.
- [6] R. Breaks and I. Keilegom, *Flexible modelling based on copulas in nonparametric regression*. J. Multivariate Anal. **6** (2009), 1270–1281.
- [7] X. Chen and Y. Fan, *Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection*. Canad. J. Statist. **33** (2005), 389–414.
- [8] D.L. Donoho, I.M. Johnstone, G. Kerkycharian and D. Picard, *Wavelet shrinkage: asymptopia (with discussion)*. J. Roy. Statist. Soc. Ser. B **57(2)** (1995), 301–369.
- [9] H. Doosti and Y.P. Chaubey, *Wavelet linear density estimation for negatively dependent random variables*. Curr. Dev. Theory Appl. Wavelets **1** (2007), 57–64.
- [10] L. Dearden, E. Fitzsimons and A. Goodman, *Estimating Lifetime Earnings Distributions Using Copulas*. IFS Working paper, 2006.
- [11] J.D. Fermanian and O. Scaillet, *Some statistical pitfalls in copula modelling for financial applications*. In: E. Klein (Ed.), *Capital Formation, Governance and Banking*. Nova Science Publishing, New York, 2005.
- [12] C. Genest, E. Masiello and K. Tribouley, *Estimating copula densities through wavelets*. Math. Econom. **44** (2010), 170–181.
- [13] I. Gijbels and J. Mielniczuk, *Estimating the density of a copula function*. Comm. Statist. Theory Methods **19** (1990), 445–464.
- [14] A. Morettin, M.C. Toloi, C. Chiann and J. Miranda, *Wavelet smoothed empirical copula estimators*. Revista Brasileira de Finanças **8** (2010), 263–281.
- [15] J. Peetre, *New thoughts on Besov spaces*. Duke University, Mathematics Series 1, 1976.
- [16] A. Sancetta, *Nonparametric estimation of multivariate distributions with given marginals: L_2 theory*. Cambridge Working Papers in Economics No. 0320, 2003.
- [17] A. Sancetta and S. Satchell, *The Bernstein copula and its applications to modeling and approximations of multivariate distributions*. Econometric Theory **20** (2004), 535–562.
- [18] A. Sklar, *Fonctions de repartition a n dimensions et leurs marges*. Publ. Inst. Statist. Univ. Paris **8** (1959), 229–231.
- [19] K. Tribouley, *Practical estimation of multivariate densities using wavelet methods*. Stat. Neerl. **49** (1995), 41–62.
- [20] A.W. Van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.

Received 14 June 2012

Université Montpellier 2
13M Montpellier
34095 Montpellier Cedex 05, France

Payame Noor University
Statistics Department
19395-4697 Tehran, Iran