Nonparametric Estimation of Multivariate Extreme Value Copulas with Known and Unknown Marginal Distributions

Samia AYARI¹,²,³,* and Mohamed BOUTAHAR²

¹Institut des Hautes Etudes de Tunis, Université de Tunis, Tunisie
²Institut de Mathématiques de Marseille, Université d’Aix-Marseille, France
³BESTMOD, Institut Supérieur de Gestion de Tunis, Université de Tunis, Tunisie

*Corresponding author email: samia.ihec@hotmail.fr

Abstract. The purpose of this paper is estimating the dependence function of multivariate extreme values copulas. Different nonparametric estimators are developed in the literature assuming that marginal distributions are known. However, this assumption is unrealistic in practice. To overcome the drawbacks of these estimators, we substituted the extreme value marginal distribution by the empirical distribution function. Monte Carlo experiments are carried out to compare the performance of the Pickands, Deheuvels, Hall-Tajvidi, Zhang and Gudendorf-Segers estimators. Empirical results showed that the empirical distribution function improved the estimators’ performance for different sample sizes.

Keywords: Multivariate extreme value copulas, dependence function, nonparametric estimation, empirical distribution function, dependence function, Monte-Carlo simulation.

1. Introduction

The last decades have seen an increase in statistical modeling research of extreme values and rare events. The extreme value theory was developed more than nine decades. It’s a branch of statistics that consists on analyzing the stochastic behavior of extreme values in order to estimate the probability of occurrence of extreme events. The cornerstone of extreme value theory is the Fisher-Tippet's theorem [7].

The extreme value analysis began with univariate data. Univariate extreme value theory assumes unrealistic hypotheses such as the independence of observations. In addition, most real extreme value applications are inherently multivariate by nature [16, 1]. This explains why recent researches in extreme value theory are more focusing on multivariate extreme value analysis [2, 6, 15]. Particularly, the copula theory has played a relevant role in modeling multivariate extreme value distributions [8, 17]. Copula is a mathematical model which can decompose a multivariate distribution into two components: the marginal distribution functions, on one hand, and a mathematical function that links them by modeling their dependence, on the other hand. Copula is the joint distribution of multivariate data with uniform marginal distribution functions. Extreme-value copulas are considered a relevant type of copulas [11]. Nonparametric estimation of multivariate copulas consists on the dependence function estimation. This line of research began with the works of Pickands [12] who developed an estimator of the dependence function for bivariate extreme value copulas. Several studies have criticized the work of Pickands because his nonparametric estimator doesn't satisfy any property of a dependence function. To circumvent the drawbacks of the Pickands estimator, several estimators were proposed in the literature namely the Deheuvels estimator [4], the CFG estimator [3] and the Hall-Tajvidi estimator [10].
As shown above, the majority of researches on multivariate extreme value theory didn't use more than two random variables. In the frequent case, when we attempt to model extreme values behavior, more than two random variables should be considered. That's why Falk and Reiss [5] generalized the bivariate extreme value copulas results to \( n \) random variables.

Multivariate extensions of the nonparametric dependence function estimators were carried out by Zhang et al. [18]. An ordinary least squares estimator of the dependence function of multivariate extreme value distributions was developed by Gudendorf and Segers [9].

The previous dependence function estimators, usually, assume that marginal distribution functions are generalized extreme value distributions. This assumption is unrealistic because marginal distributions are rarely known in practice. In this paper, we aim at estimating the dependence structure when the marginal distributions are unknown. We will propose the Pickands, Hall-Tajvidi, Deheuvels, Zhang and Gudendorf-Segers estimators in a new way which consists on replacing the Gumbel extreme value distribution by the empirical distribution function.

The paper is structured as follows. In section 2, we will deal with nonparametric estimators of the dependence function for multivariate extreme value copulas with known margins. In section 3, an overview of the empirical distribution function is described. A Monte-Carlo simulation is conducted, in section 4, to check the estimators' performance of multivariate extreme values copulas with known and unknown marginal distributions.

2. Nonparametric Estimation of Multivariate Extreme Value Distributions

Given a sample \((X_{i1},...,X_{ip})\), \(i = 1,...,n\), of independent and identically distributed (i.i.d) random vectors with marginal distributions \(F_j(x)\) of \(X_{ij}\), \(j \in \{1,...,p\}\) and a multivariate extreme value copula \(C\) presented as follows:

\[
C(U_1 ..., U_p) = C (F_1(X_1),...,F_p(X_p)) = \exp \left\{ \sum_{j=1}^{p} \lg(u_j) A \left( \frac{\lg(u_1)}{\sum_{j=1}^{p} \lg(u_j)}, \ldots, \frac{\lg(u_{p-1})}{\sum_{j=1}^{p} \lg(u_j)} \right) \right\}
\](1)

The dependence function \(A(s_1, s_2, \ldots, s_{p-1})\) is defined in the \((p-1)\)-dimensional unit simplex:

\[
S_{p-1} = \{ (s_1, s_2, \ldots, s_{p-1}) : s_j \geq 0; \sum_{j=1}^{p-1} s_j \leq 1 \text{ for } j = 1, \ldots, p-1 \}.
\]

It's a convex function such as \( \max(s_1, s_2, \ldots, s_{p-1}) \leq A(s_1, s_2, \ldots, s_{p-1}) \leq 1 \).

Zhang et al. [18] generalized the Pickands, Hall-tajvidi and Deheuvels estimators to the multivariate case; he also developed a new estimator of the dependence function which is the generalization of the CFG estimator. The estimators are presented as follows:

The Pickands estimator:

\[
\hat{A}^P(s_1, s_2, \ldots, s_{p-1}) = \frac{n}{\sum_{i=1}^{n} \min_{j=1}^{p} \frac{Y_{ij}}{S_j}}, s \in S_p
\]

The Hall-Tajvidi estimator:

\[
\hat{A}^HT(s_1, s_2, \ldots, s_{p-1}) = \frac{n}{\sum_{i=1}^{n} \frac{Y_{ij}}{S_j}}, s \in S_p
\]

The Deheuvels estimator:

\[
\hat{A}^D(s_1, s_2, \ldots, s_{p-1}) = \frac{n}{\sum_{i=1}^{n} \min_{j=1}^{p} \frac{Y_{ij}}{S_j} n \sum_{j=1}^{n} \frac{Y_{ij} s_j + n}{S_j}}, s \in S_p
\]

Where \( \hat{y}_j = \frac{\sum_{i=1}^{n} Y_{ij}}{n}, j = 1, \ldots, p \).

The Zhang estimator:

\[
\hat{A}^Z(s_1, s_2, \ldots, s_{p-1}) = \frac{\prod_{j=1}^{p} \left( \sum_{i=1}^{n} Y_{ij} \hat{y}_j \right)^{f_j}}{\prod_{j=1}^{p} \sum_{i=1}^{n} \min_{j=1}^{p} \frac{Y_{ij}}{S_j}}, s \in S_p
\]

(5)
where, $Y_{ij} = -\log(F_j(X_{ij})), Y_j = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}, 1 \leq j \leq p, \sum_{j=1}^{p} s_j = 1$.

Gudendorf and Segers [9] modeled the relationship between the dependent variable $(-\log \xi_i(s) - \gamma; \ i = 1, ..., n; \ s \in S_p)$ and the $p$ regressors $(-\log \xi_i(e_j) - \gamma; \ i = 1, ..., n; j = 1, ..., p; \ s \in S_p)$ in the form of linear regression:

$$(-\log \xi_i(s) - \gamma) = U_{n,0}(s) + \sum_{j=1}^{p} U_{n,j}(s) (-\log \xi_i(e_j) - \gamma) + \epsilon_i(s), i = 1, ..., n, s \in S_p,$$ (6)

where

- $\xi_i(s) = \sum_{j=1}^{p} \frac{Y_{ij}}{s_j}$ whatever $i = 1, ..., n$ and $s \in S_p$,
- $\gamma$ is the Euler–Mascheroni constant,
- $e_j = \{0, ..., 1, 0, ..., 0\} \in \mathbb{R}^p$ are standard unit vectors,
- $\epsilon_i(s)$ is the error term.

The ordinary least squares method consists on estimating the $(p + 1)$ regression coefficients $\beta = (\beta_{n,0}(s), \beta_{n,1}(s), ..., \beta_{n,p}(s))$ that minimize the sum of squared residuals:

$$\hat{\beta} = \arg \min \sum_{i=1}^{n} \epsilon_i^2(s)$$

$$= \arg \min \sum_{i=1}^{n} ((-\log \xi_i(s) - \gamma) - U_{n,0}(s) + \sum_{j=1}^{p} U_{n,j}(s) (-\log \xi_i(e_j) - \gamma))^2, \ s \in S_p$$ (7)

So, the ordinary least square estimator of the dependence function $\log \hat{A}_{LS}^0(s)$ is the estimated intercept $\hat{U}_{n,0}(s)$:

$$\hat{U}_{n,0}(s) = (-\log \xi_i(s) - \gamma) - \sum_{j=1}^{p} \hat{U}_{n,j}(s) (-\log \xi_i(e_j) - \gamma)$$

$$= \log \hat{A}(s) - \sum_{j=1}^{p} \hat{U}_{n,j}(s) \log \hat{A}(e_j)$$

$$= \log \hat{A}_{LS}^0(s)$$ (8)

3. Overview of the Empirical Cumulative Distribution Function

Statistics aim to understand and to model the stochastic behavior of random variables in order to predict accurately their potential occurrences in the future. There is two branches of statistics: the parametric statistics and the nonparametric statistics. Parametric statistics make some assumptions about the data and all the future inferences are based on these assumptions. If the assumptions are true, then, the resulting inferences are correct. Whereas, if they are false, the statistical analysis will be wrong. In practice, it is often difficult to find a suitable parametric model for real data. That’s why researches are more oriented towards nonparametric statistics. Nonparametric approach makes few or no assumptions about the data. Using the empirical distribution function represents an alternative to overcome the limitations of parametric statistics.

The cumulative distribution function $F(x)$ defines the probability that a random variable $X$ takes values less than or equal to $x$. While, the empirical distribution is the proportion of observations of a random sample which are less than or equal to a value $x$; it’s a step function that jumps by $\frac{1}{n}$ at each data point $x_i$.

Suppose $X_1, X_2, ..., X_n$, are i.i.d random variables, the empirical distribution function is defined by :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$$ (9)

Where $\mathbb{I}$ is the indicator function.

The empirical distribution function has the following properties:

- It is an increasing and right-continuous function.
- $\lim_{n \to +\infty} F_n(x) = 0$ and $\lim_{n \to +\infty} F_n(x) = 1$. 
A strong proof of convergence is given by the strong law of large numbers, for any $x$ convergence of

$$E[F_n(x)] = F(x) \quad \text{and} \quad \text{var}[F_n(x)] = \frac{1}{n} F(x) (1 - F(x)).$$

According to the Chebyshev’s inequality, for any real number $\varepsilon > 0$, we have:

$$\lim_{n \to \infty} P[F_n(x) - F(x) \geq \varepsilon] = 0$$

(10)

This fact implies that $F_n(x) \xrightarrow{prob} F(x)$ as $n \to \infty$ for any fixed value of $x$.

A strong proof of convergence is given by the strong law of large numbers, for any $x \in \mathbb{R}$:

$$F_n(x) \xrightarrow{a.s} F(x)$$

(11)

The strongest convergence result is given by the Glivenko--Cantelli theorem, it proves the uniform convergence of $F_n(x)$ to $(x)$, for any $x \in \mathbb{R}$:

$$\sup_x |F_n(x) - F(x)| \to 0$$

(12)

Finally, the Dvoretzky–Kiefer–Wolfowitz inequality gives another uniform property of $F_n(x)$ such that:

$$P(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}, \forall n \in \mathbb{N}, \varepsilon > 0$$

(13)

All the properties presented above show that the empirical distribution function is a reliable tool in nonparametric statistics.

In the multivariate case, the empirical distribution function is presented as follows:

Suppose $(X_1, \ldots, X_p)$, $i = 1, \ldots, n$, are i.i.d random vectors, the multivariate empirical distribution function is defined by:

$$F_{i,i_1, \ldots, i_p}(x_1, \ldots, x_p) = \frac{1}{n} \sum_{i=1}^{n} I(X_{i_1} \leq x_1, \ldots, X_{i_p} \leq x_p).$$

(14)

4. Monte Carlo Simulation Study

In this study, we focus on comparing some nonparametric dependence function estimators for multivariate extreme value copulas in the case of known and unknown marginal distributions. Firstly, we assume that margins are Gumbel extreme value distributions. Secondly, we substitute the extreme value marginal distributions by the empirical distribution functions.

In the first step, we simulate trivariate extreme value distributions having logistic dependence functions which is described as follows [14]:

$$A(s_1, s_2) = (\alpha r s_1^2 + \beta r s_2^2, \beta r s_1^2 + \beta r s_2^2, \alpha r s_1^2 + \beta r s_2^2)^{1/2} + (\alpha r s_1^2 + \beta r s_2^2)^{1/2} + (\alpha r s_1^2 + \beta r s_2^2)^{1/2}$$

(15)

where $s_1 = 1 - s_1 - s_2$.

The symmetric logistic dependence function is considered in our study with $r = 3$, $\alpha = 0$ and $\mu = 1$.

The simulation is repeated using different sample sizes and the Pickands, Deheuvels, Hall-tajvidi, Zhang and Gudendorf-Segers estimators are compared. The mean integrated square errors MISE is used to assess the estimators’ performance.

$$\text{MISE} = \int_{(s_1, s_2) \in S_2} E \left[ \hat{A}(s_1, s_2) - A(s_1, s_2) \right]^2 ds_1 ds_2$$

(16)

where $\hat{A}(s_1, s_2)$ is the estimated dependence function and $A(s_1, s_2)$ is the observed dependence function given by the R evd package. In this study, the MISE is approximated by the Monte Carlo approach:

$$\text{MISE (i)} \equiv \frac{1}{N^2} \sum_{j=1}^{N_1} \sum_{k=1}^{N_j} \left[ \hat{A}(s_{1,j}, s_{2,k}) - A(s_{1,j}, s_{2,k}) \right]^2$$

(17)

with $N_i = N_1 + 1$, $N_j = N_1 - j + 1$ and $N$ is the simplex parameter. The MISE results are presented in table 1.
Table 1. MISE for different sample sizes in the case of Gumbel marginal distributions.

|                | n=25    | n=50    | n=100   | n=200   |
|----------------|---------|---------|---------|---------|
| Pickands       | 1164.02 | 555.92  | 262.07  | 130.36  |
| Deheuvels      | 152.6   | 74.46   | 37.83   | 18.87   |
| Hall-Tajvidi   | 26.37   | 13.38   | 6.76    | 3.41    |
| Zhang          | 43.88   | 22.04   | 10.9    | 5.46    |
| Gudendorf-Segers | 18      | 8.26    | 3.93    | 1.91    |

As claimed in the bivariate case by Hall and Tajvidi [10], the Hall-Tajvidi estimator has better performance than the Pickands estimator in terms of MISE for different sample sizes. However, unlike the bivariate studies, the generalization of the Hall-Tajvidi estimator to the multivariate case allowed him to be a competitor estimator that has better performance than the CFG estimator generalized to the multivariate case.

Moreover, empirical results show that the Pickands estimator has lower performance than the Deheuvels and the Zhang estimators for different sample sizes. It’s clear that Gudendorf-Segers estimator has the best performance over the other estimators for different sample sizes.

The MISE results of the estimators when substituting the Gumbel extreme value distribution by the empirical distribution function are presented in table 2.

Table 2. MISE for different sample sizes in the case of empirical marginal distributions.

|                | n=25    | n=50    | n=100   | n=200   |
|----------------|---------|---------|---------|---------|
| Pickands       | 141.16  | 53.11   | 20.69   | 8.06    |
| Deheuvels      | 40.25   | 18.84   | 9.29    | 4.5     |
| Hall-Tajvidi   | 32.73   | 16.87   | 8.63    | 4.42    |
| Zhang          | 32.76   | 16.7    | 8.5     | 4.41    |
| Gudendorf-Segers | 28.18  | 13.81   | 7.02    | 3.39    |

As in the bivariate case [13], the Deheuvels and the Hall-Tajvidi estimators have higher performance than the Pickands estimator in modeling multivariate dependence structure. Furthermore, the Pickands estimator is more efficient in the multivariate case than the bivariate case. The robustness of the Hall-Tajvidi estimator is confirmed for estimating both multivariate and bivariate extreme value copulas. In addition, empirical results show that the Hall-Tajvidi and the Zhang estimators have similar results in terms of MISE for different sample sizes. Concerning the Gudendorf-Segers estimator, it has better performance than the other estimators for different sample sizes.

Our second interest is comparing the MISE results of table 2 with those obtained by the parametric approach which assumes that marginal distributions are known (table 1). Replacing the Gumbel extreme value distributions by the empirical distribution functions, the Pickands estimator performance is improved significantly for different sample sizes. Moreover, the Deheuvels estimator is more efficient with empirical margins than Gumbel marginal distributions. Regarding the Zhang estimator, it’s ameliorated slightly when Gumbel distributions are replaced by the empirical distribution functions. Lastly, unlike the other estimators, we note that Gudendorf-Segers and Hall-Tajvidi estimators have better performance with known distributions than unknown ones for different sample sizes.

5. Conclusion

In this study, we developed some nonparametric estimators of the dependence function for trivariate extreme value copulas in two cases. Firstly, we assume Gumbel extreme value margins. Secondly, we substitute the marginal distributions by the empirical distribution functions. Simulation study shows that the nonparametric approach improved the estimators’ performance for different sample sizes. This result was predicted because classical estimators assume unrealistic assumption consisting on Gumbel extreme value margins, so the estimator performance will depend on the estimation accuracy of the Gumbel extreme value distribution parameters (scale, shape and location). We can conclude that the rank-based estimators can be used as a reliable technique for modeling the dependence function of multivariate extreme value copulas in different fields such as stock markets, air pollution and hydrology.
References

[1] Ayari S and Boutahar M, 2017, multivariate extreme value theory and application to environment, International Journal of Management and Applied Science, vol 3 pp 11-13.

[2] Beck N, Genest, C Jalbert J and Mailhot M, 2020, Predicting extreme surges from sparse data using a copula - based hierarchical Bayesian spatial model Environmetrics, vol 31, p e2616.

[3] Capéraa P, Fougères A L, Genest C, 1997, A nonparametric estimation procedure for bivariate extreme value copulas, Biometrica, vol 84, pp 567-577.

[4] Deheuvels P, 1991, On the limiting behavior of the Pickands estimator for bivariate extreme-value distributions, Statistics & Probability Letters, vol 12, pp 429-439.

[5] Falk M and Reiss R D, 2005, On Pickands coordinates in arbitrary dimensions, Journal of Multivariate Analysis, vol 92, pp 426-453.

[6] Ferreira H, 2011, Dependence between two multivariate extremes, Statistics and Probability Letters, vol 81, pp 586-591

[7] Fisher R A and Tippett L H C, 1928, Limiting forms of the frequency distribution of the largest and smallest member of a sample, Proc. Cambridge Philosophical Society, vol 24, pp 180–190.

[8] Frees E W and Valdez E A, 1998, Understanding relationships using copulas, North American Actuarial Journal, vol 2, pp 1-25.

[9] Gudendorf G and Segers J, 2012, Nonparametric estimation of multivariate extreme-value copulas, Journal of Statistical Planning and Inference, vol 142, pp 3073-3085.

[10] Hall P and Tajvidi N 2000, Distribution and dependence-function estimation for bivariate extreme value distributions, Bernouilli, Vol 6, pp 835-844.

[11] Nelsen R B, 2007, An introduction to copulas Springer Science & Business Media.

[12] Pickands J, 1981, Multivariate extreme value distributions, Proc. of the 43rd Session of the International Statistical Institute Buenos Aires, vol 49, pp 859-878.

[13] Segers J, 2007, Nonparametric inference for bivariate extreme value copulas in Topics in extreme values, Nova Science Publishers New York, pp 181-203.

[14] Stephenson A, 2003, Simulating multivariate extreme value distributions of logistic Type, Extremes vol 6, pp 49-59.

[15] Stuart C G J and Tawn A, 1991, Modeling extreme multivariate events, Journal of the Royal Statistical Society Series B (Methodological), vol 53, pp 377-39

[16] Tawn J A, 1994, Applications of Multivariate Extremes in Extreme Value Theory and its Applications, Kluwer, pp 249-268.

[17] Tawn J A, 1990, Modelling multivariate extreme value distributions, Biometrica, vol 77, pp 245–253.

[18] Zhang D, Wells M T and Peng L, 2008, Nonparametric estimation of the dependence function for a multivariate extreme value distribution, Journal of Multivariate Analysis, vol 99, pp 577-588.