Non-parametric Copula Estimation

Seminar: Copulas and their Applications

Viet Hoang

February 5, 2020

Institute of Stochastics
Ulm University

Preliminaries

Review: Basics on Copulas

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of a $d-$dimensional random vector $\boldsymbol{X}$ with distribution function H and continuous, univariate margins $F_1, \ldots, F_d$.

Theorem (Sklar's Theorem)

There exists a unique copula C on $[0,1]^d$ such that

$$H(\boldsymbol{x}) = C\left(F_1(x_1), \ldots, F_d(x_d)\right)$$

for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

Lemma (Stochastic analog of Sklar's theorem)

Let $\boldsymbol{X}$ be a $d-$dimensional random vector with continuous, univariate marginals $F_1, \ldots, F_d$. Then, $\boldsymbol{X}$ has copula C if and only if

$$\boldsymbol{U} = (F_1(X_1), \ldots, F_d(X_d)) \sim C.$$

Fully, semi-, and non-parametric estimation

Assumptions	**Estimation**
$C = C_{\boldsymbol{\theta_0}} \in \{C_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ $F_j = F_{j,\gamma_{0,j}} \in \{F_{j,\gamma_j} : \gamma_j \in \Gamma_j\}$	Fully parametric
$C = C_{\boldsymbol{\theta_0}} \in \{C_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$	semi-parametric
no assumptions on C ***or*** F_j	***non-parametric***

Non-parametric estimation

Non-parametric estimation: Margins

Definition

Define non-parametric estimators for the margins $F_1, \dots, F_d$ by

$$F_{n,j}(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\left(X_{ij} \leq x\right) \tag{1}$$

for $x \in \mathbb{R}$, $j = 1, \dots, d$ (i.e. empirical distribution function over the $j-$th elements of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$).

Definition

The ***pseudo observations*** of the copula C are defined by

$$\boldsymbol{U}_{i,n} = \left(F_{n,1}(X_{i1}), \ldots, F_{n,d}(X_{id})\right), \tag{2}$$

for $i = 1, \ldots, n$, where $F_{n,j}$ is the empirical margin defined in (1).

- Dividing by $(n+1)$ instead of n is asymptotically negligible but ensures that $\boldsymbol{U}_{i,n}$ lies in the interior $(0,1)^d$ which is important e.g. for maximum pseudo-likelihood estimation.
- Let R_{ij} be the rank of X_{ij} among $X_{1,j}, \ldots, X_{nj}$. Then, $F_{n,j}(X_{ij}) = R_{ij}/(n+1)$ and

$$\boldsymbol{U}_{i,n} = \frac{1}{n+1}\left(R_{i1}, \ldots, R_{id}\right)$$

(sample of multivariate scaled ranks).

What do you think is a natural choice for a non-parametric copula estimator?

The empirical copula

Definition

The ***empirical copula*** is defined by

$$C_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(\boldsymbol{U}_{i,n} \leq \boldsymbol{u}\right) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} \mathbb{1}\left(U_{ij,n} \leq u_j\right) \tag{3}$$

for $\boldsymbol{u} \in [0,1]^d$, where $\boldsymbol{U}_{i,n}$, $i = 1, \ldots, n$, are the pseudo observations defined in (2).

Properties of the empirical copula

- **Invariant under monotone increasing transformations of data**
  Only based on multivariate scaled ranks.
- **Asymptotic properties** can be derived from the ***empirical copula process***

$$\sqrt{n}\left(C_n(\boldsymbol{u}) - C(\boldsymbol{u})\right), \quad \boldsymbol{u} \in [0,1]^d.$$

- **Consistent estimator of** C (Deheuvels (1979))
- **Asymptotically centered Gaussian process**
  Assumption: Independence!

Smooth non-parametric estimators

Order statistics

- Recall: empirical copula

$$C_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\mathbb{1}\left(U_{ij,n} \leq u_j\right) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\mathbb{1}\left(\frac{R_{ij}}{n+1} \leq u_j\right).$$

Q: What causes „unsmoothness" of the empirical copula C_n?

- ***Idea:*** Replace indicator functions in empirical copula by the cumulative distribution function of the $R_{ij,n} = r - th$ order statistic of U_i.
- The $r-$th order statistic of a uniformly distributed sample is ***beta-distributed*** with parameters n and $n - r + 1$.

Empirical $\beta-$copula

Definition (Segers, Sibuya, Tsukahara (2017))

The ***empirical beta-copula*** is defined by

$$C_n^\beta(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} F_{n,R_{ij}}(u_j) \tag{4}$$

for $\boldsymbol{u} \in [0,1]^d$, where

- $F_{n,r}$ denotes the cdf of the beta-distribution with parameters r and $n+1-r$, $r = 1, \ldots, n$,
- R_{ij} is the rank of the observation X_{ij} among $X_{1j}, \ldots, X_{nj}$.

Facts on beta-copulas

- The empirical beta-copula is a genuine copula with has standard uniform univariate margins if the components $X_1, \ldots, X_n$ are independent.

  Margins: For $k \in 1, \ldots, d,\ u_k \in [0, 1]$

$$C_n^\beta(1, \ldots, 1, u_k, 1, \ldots, 1) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} F_{n,R_{ij}}(u_j) = \frac{1}{n} \sum_{i=1}^{n} F_{n,R_{ik}}(u_k)$$

$$= \frac{1}{n} \sum_{r=1}^{n} F_{n,r}(u_k) = \frac{1}{n} \sum_{r=1}^{n} \mathbb{E}\left[\mathbb{1}\left(U^{(r)} \leq u_j\right)\right]$$

$$= \mathbb{E}\left[\frac{1}{n} \sum_{r=1}^{n} \mathbb{1}\left(U^{(r)} \leq u_k\right)\right] = u_k.$$

Empirical Bernstein copulas

- The empirical beta-copula is a special case of ***empirical Bernstein copulas*** C_n^{Bern}.
- Under certain conditions Berstein polynomials form Copulas.
- What about weak convergence and asymptotics of

$$\sqrt{n}(C_n(\boldsymbol{u})^{Bern} - C(\boldsymbol{u})))?$$

Yes, ***but*** under many conditions!

Non-parametric estimation for extreme-value copulas

Motivation: Extrem-value theory

- Motivation from ***extreme-value theory*** (Resnick, Pickands) and transferred to the concept of copula theory by Gudendorf and Segers.
- Consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of $d-$dim. random vectors $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{i,d})$, and the vector of maxima

$$\boldsymbol{M}_n = (M_{n1}, \ldots, M_{nd}),$$

where $M_{nj} = \max_{i=1,\ldots,n} X_{ij}$.

Copula of max vector

Proposition

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ *have copula* C *and marginal distributions* $F_1, \ldots, F_d$. *Then, the copula of the max vector* $\boldsymbol{M}_n = (M_{n1}, \ldots, M_{nd})$ *is given by*

$$C_{M_n}(u_1, \ldots, u_d) = C\left(u_1^{1/n}, \ldots, u_d^{1/n}\right)^n$$

Extreme-value copulas

Definition

A d-dimensional copula C is an ***extreme-value copula*** if there exists a copula C^* such that, for any $\boldsymbol{u} \in [0,1]^d$,

$$\lim_{n\to\infty} C^*\left(u_1^{1/n}, \ldots, u_d^{1/n}\right)^n = C\left(u_1, \ldots, u_d\right).$$

The copula C^* is then said to be in the *maximum domain of attraction* of C.

Remark

There are useful characterizations of extreme-value copulas, e.g.

- *Characterization based on* ***max-stability***
- *Characterization based on* ***Pickands dependence formula***

Definition

A copula is called ***max-stable*** if it satisfies the relationship

$$C(u_1, \ldots, u_d) = C\left(u_1^{1/m}, \ldots, u_d^{1/m}\right)^m$$

for all integeres $m \geq 1$ and $\boldsymbol{u} = (u_1, \ldots, u_d) \in [0, 1]^d$.

Theorem

A copula is an extreme-value copula if and only if it is max-stable.

The Pickands dependence function

Proposition

A copula C is an extreme value copula if and only if there exists a function $A : \Delta_{d-1} \to \left[\frac{1}{d}, 1\right]$ such that for any $\boldsymbol{u} \in (0,1]^d \setminus \{(1, \dots, 1)\}$

$$C(\boldsymbol{u}) = \exp\left(\left(\sum_{j=1}^{d} \log u_j\right) A\left(\frac{\log u_1}{\sum_{j=1}^{d} \log u_j}, \dots, \frac{\log u_d}{\sum_{j=1}^{d} \log u_j}\right)\right). \quad (5)$$

The function A is called the ***Pickands dependence function*** *associated with C.*

Convexity and bounds

The Pickands dependence function is

(1) convex,

(2) bounded by

$$\max\left\{w_1, \ldots, w_{d-1}, 1 - \sum_{j=1}^{d-1} w_j\right\} \leq A(\boldsymbol{w}) \leq 1$$

for $\boldsymbol{w} = (w_1, \ldots, w_{d-1}) \in \Delta_{d-1}$.

Remark

(1) and (2) are not sufficient conditions to characterize the class of Pickands dependence functions ***unless*** $d = 2$.

Special case: $d = 2$

Example

In the case $d = 2$ an extrem-value copula is specified by the convex, one-dimensional Pickands dependence function $A : [0, 1] \to [1/2, 1]$ which satisfies

$$C(u, v) = (uv)^{A(\log(v)/\log(uv))}$$

and

$$\max\{t, 1-t\} \leq A(t) \leq 1, \quad t \in [0, 1].$$

Special case: $d = 2$ (Cont'd)

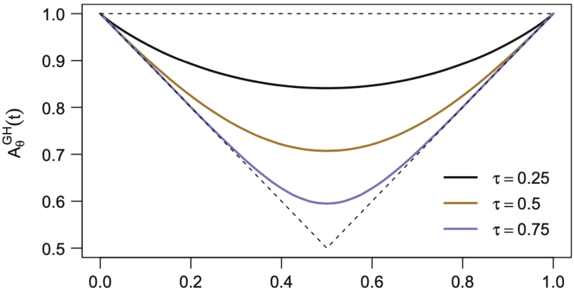
Example

Consider the ***bivariate Gumbel-Hougaard copula***. It is given by

$$C_\theta(u, v) = \exp\left(-\left[(-\log u)^\theta + (-\log v)^\theta\right]^{1/\theta}\right)$$

You can easily verify that the Pickands dependence function is given by

$$A_\theta^{GH}(t) = \left(t^\theta + (1-t)^\theta\right)^{1/\theta}, \quad t \in [0, 1].$$



Non-parametric estimation: Idea

Assumption: C ***is an extreme-value copula.***

Idea

Given a non-parametric estimator A_n of A computed from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, define the plug-in estimator by plugging A_n into the Pickands characterization of C (5), i.e.

$$C_n(\boldsymbol{u}) = \exp\left(\left(\sum_{j=1}^{d} \log u_j\right) A_n\left(\frac{\log u_1}{\sum_{j=1}^{d} \log u_j}, \ldots, \frac{\log u_d}{\sum_{j=1}^{d} \log u_j}\right)\right).$$

- Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a sample of the $d-$dim. random vector $\boldsymbol{X}$ with corresponding pseudo observations $\boldsymbol{U}_{i,n} = (U_{i1,n}, \ldots, U_{id,n})$, $i = 1, \ldots, n$.
- Define

$$\zeta_{i,n}(\boldsymbol{w}) = \min_{j \in \{1,\ldots,d\}} \left\{ \frac{-\log(U_{ij,n})}{w_j} \right\}$$

for $\boldsymbol{w} \in \Delta_{d-1}$, $i = 1, \ldots, n$.

Definition

The ***Pickands estimator*** is defined by

$$A_n^P(\boldsymbol{w}) = \left(\frac{1}{n} \sum_{i=1}^{n} \zeta_{i,n}(\boldsymbol{w}) \right)^{-1}$$

for $\boldsymbol{w} \in \Delta_{d-1}$.

Why?!

Pickands estimator for $d = 2$

- Let (X, Y) be a bivariate random vector with continuous, marginal distributions F and G, and unique copula C.
- Let (U, V), where $U = F(X)$ and $V = G(Y)$.
- The random variables $-\log(U)$ and $-\log(V)$ are ***exponentially distributed with mean*** 1.

- For all $t \in (0, 1)$ set

$$\zeta(t) = \min\left\{\frac{-\log(U)}{1-t}, \frac{-\log(V)}{t}\right\},$$

  and $\zeta(0) = -\log(U)$ as well as $\zeta(1) = -\log(V)$.

- Then ,

$$\begin{aligned}\mathbb{P}(\zeta(t) > x) &= \mathbb{P}\left(\frac{-\log(U)}{1-t} > x, \frac{-\log(V)}{t} > x\right) \\ &= \mathbb{P}\left(\log(U) < -x(1-t), \log(V) < -xt\right) \\ &= \mathbb{P}\left(U < e^{-x(1-t)}, V < e^{-xt}\right) \\ &= C\left(e^{-x(1-t)}, e^{-xt}\right)\end{aligned}$$

  (Last eq.: $(U, V) \sim C$.)

$$\mathbb{P}(\zeta(t) > x) = C\left(e^{-x(1-t)}, e^{-xt}\right)$$

- Recall $C(u, v) = (uv)^{A(\log(v)/\log(uv))}$, hence

$$\mathbb{P}(\zeta(t) > x) = \left(e^{-x(1-t)}e^{-xt}\right)^{A(-xt/(-x))} = e^{-xA(t)}$$

- Consequently $\zeta(t)$ is exponentially distributed with

$$\mathbb{E}\zeta(t) = \frac{1}{A(t)}$$

With $\mathbb{E}\zeta(t) = \frac{1}{A(t)}$ in mind

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ sample of bivariate r.v. (X, Y)
- $U_i = F_n(X_i)$ and $V_i = G_n(Y_i)$ are the pseudo observations (F_n, G_n emp. distr. fn.)
- Set

$$\zeta_{i,n}(t) = \min\left\{\frac{-\log(U_i)}{1-t}, \frac{-\log(V_i)}{t}\right\}$$

- Define

$$A_n^P(t) = \left(\frac{1}{n}\sum_{i=1}^{n}\zeta_{i,n}(t)\right)^{-1}$$

Exponential and Gumbel distribution

We make the following observations (d=2):

- $Z \sim \mathit{Exp}(1)$ then $Z/\lambda \sim \mathit{Exp}(\lambda)$
- Gumbel distribution: $W \sim \mathit{Gumbel}(\mu, \beta)$ with $\mathbb{E}W = \mu + \beta\gamma$, where $\gamma \approx 0.577$ is the Euler-Mascheroni constant.
- $-\log Z \sim \mathit{Gumbel}(0, 1)$, thus

$$\mathbb{E}\left[\log \zeta(t)\right] = -\mathbb{E}\left[-\log\left(\frac{Z}{A(t)}\right)\right] = -\mathbb{E}\left[-\log Z\right] - \log A(t)$$

$$= -\gamma - \log A(t)$$

$$\Leftrightarrow A(t) = \exp\left(-\gamma - \mathbb{E}\left[\log \zeta(t)\right]\right)$$

Capéraà-Fougères-Genest estimator

Definition

The ***Capéraà-Fougères-Genest estimator*** is defined by

$$A_n^{CFG}(\boldsymbol{w}) = \exp\left(-\gamma - \frac{1}{n}\sum_{i=1}^{n} \log \zeta_{i,n}(\boldsymbol{w})\right)$$

for $\boldsymbol{w} \in \Delta_{d-1}$, where $\gamma \approx 0.577$ is the Euler-Mascheroni constant.

References i

[1] P. Deheuvels.
***La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance*.**
Bulletins de l'Académie Royale de Belgique, 1979.

[2] J.-D. Fermanian, D. Radulovic, and M. Wegkamp.
***Weak convergence of empirical copula processes*.**
Bernoulli, 2004.

[3] C. Genest and J. Segers.
***Rank-based inference for bivariate extreme-value copulas*.**
The annals of statistics, 2009.

[4] G. Gudendorf and J. Segers.
***Extrem value copulas*.**
Springer, 2010.

[5] M. Hofert, I. Kojadinovic, M. Mächler, and J. Yan.
***Elements of Copula Modeling with R*.**
Springer, 2018.

[6] R. B. Nelson.
***An Introduction to Copulas*.**
Springer, 2006.

[7] J. Segers, M. Sibuya, and H. Tsukahara.
***The Empirical Beta Copula*.**
Journal of multivariate analysis, 2017.

[8] W. Stute.
***The oscillation behaviour of empirical processes*.**
The Annals of Probability, 1982.

[9] W. Stute.
The oscillation behaviour of empirical processes: The multivariate case.
The Annals of Probability, 1984.

Thank you!