A bayesian semiparametric archimedean copula

Luis E. Nieto Barajas

(joint with R. Hoyos)

Department of Statistics
ITAM

ISBA 2022 – July 1, 2022

Contents

Copulas

- Let $\varphi(\cdot) : [0,1] \to [0,\infty)$ be a continuous, strictly decreasing function s.t. $\varphi(1) = 0$.
- Let $\varphi^{-1}(\cdot)$ be the pseudo-inverse of φ, where $\varphi(t) = 0$ for $t > \varphi(0)$. If $\varphi(0) = \infty$ the generator is called strict.
- An archimedean copula $C : [0,1]^2 \to [0,1]$ with generator φ defined as

$$C(u,v) = \varphi^{-1}\left(\varphi(u) + \varphi(v)\right) \tag{1}$$

- Require that φ is convex so $C(u,v)$ is well defined
- Properties : symmetric, associative and $C(u,u) < u \; \forall u \in (0,1)$
- Kendall's tau :

$$\kappa_\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t^+)} \mathrm{d}t, \tag{2}$$

where $\varphi'(t^+)$ denotes right derivative of φ at t.

Survival analysis

- We realised that φ^{-1} is a decreasing function from $[0, \infty)$ to $[0, 1]$ so it behaves as a survival function
- Idea : propose a semi/nonparametric inverse generator φ^{-1} by using survival analysis ideas
- We recall some basic definitions :
  - $h(t)$ is a hazard rate, i.e., a nonnegative function with domain $[0, \infty)$
  - $H(t) = \int_0^t h(s)ds$ is a cumulative hazard rate s.t. $H(t) \to \infty$ as $t \to \infty$
  - $S(t) = \exp\{-H(t)\}$ is a survival function which decreases from $[0, \infty)$ to $[0, 1]$
- Therefore $S(t)$ behaves like an inverse generator $\varphi^{-1}(t)$
- Plus the convexity condition (not free)
- Simple case :
  Let $h(t) = \theta$ for all $t \quad \Rightarrow \quad S(t) = e^{-\theta t} = \varphi(t)^{-1} \quad \Rightarrow \quad \varphi(t) = -(\log t)/\theta$.
  Using (1) we get $C(u, v) = uv$ is the independent copula (and does not depend on θ).

ITAM

Main proposal

- Consider a partition of the positive real line $0 = \tau_0 < \tau_1 < \cdots < \tau_K = \infty$. Then the first derivative of the hazard rate is

$$h'(t) = \sum_{k=1}^{K} \theta_k I(\tau_{k-1} < t \leq \tau_k) \quad \text{with} \quad \theta_K = 0$$

- The hazard rate becomes

$$h(t) = \sum_{k=1}^{K} (A_k + \theta_k t)\, I(\tau_{k-1} < t \leq \tau_k) \quad \text{with} \quad \theta_0 = h(0) > 0$$

- The cumulative hazard function is

$$H(t) = \sum_{k=1}^{K} \left(B_k + A_k t + \frac{\theta_k t^2}{2} \right) I(\tau_{k-1} < t \leq \tau_k)$$

where $A_1 = \theta_0$, $A_k = \theta_0 + \sum_{j=1}^{k-1} (\theta_j - \theta_{j+1})\tau_j$,

$B_1 = 0$ and $B_k = \sum_{j=2}^{k} (\theta_j - \theta_{j-1}) \frac{\tau_{j-1}^2}{2}$ for $k = 2, \ldots, K$

Main proposal

Proposition

Consider the semiparametric inverse generator

$$\varphi^{-1}(t) = S(t) = \exp\{-H(t)\}$$

and assume that $\{\theta_k, k = 0, 1, \ldots, K\}$ *are such that* $\theta_0 > 0$, $\theta_K = 0$ *and satisfy conditions (C1) and (C2) given by*

(C1) $A_k + \theta_k t > 0$, *for* $t \in (\tau_{k-1}, \tau_k]$ *and for all* $k = 1, \ldots, K$.

(C2) $(A_k + \theta_k t)^2 > \theta_k$, *for* $t \in (\tau_{k-1}, \tau_k]$ *and for all* $k = 1, \ldots, K$.

Then,

(i) $\varphi^{-1}(t)$ *is a continuous and injective function of* t,

(ii) $\varphi^{-1}(t)$ *is a convex function,*

(iii) $\varphi^{-1}(t)$ *induces a strict generator.*

Main proposal

- By property (i) in previous Proposition, we can obtain

$$\varphi(t)=\sum_{k=1}^{K}\left(\left[\operatorname{sgn}(\theta_k)\left\{\frac{2}{\theta_k}\left(\frac{A_k^2}{2\theta_k}-B_k-\log(t)\right)\right\}^{1/2}-\frac{A_k}{\theta_k}\right]I(\theta_k\neq 0)\right.$$
$$\left.-\frac{B_k+\log(t)}{A_k}I(\theta_k=0)\right)I\left(\varphi^{-1}(\tau_k)\leq t<\varphi^{-1}(\tau_{k-1})\right).$$

- Kendall's tau satisfies $\kappa_\tau\in(-1,1)$ and has expression

$$\kappa_\tau=-1+2\sum_{k=1}^{K}A_k\int_{\tau_{k-1}}^{\tau_k}\exp\left(-2B_k-2A_kt-\theta_kt^2\right)\mathrm{d}t.$$

- If $K=1\Rightarrow$ independence copula
- For larger $K\Rightarrow$ semiparametric flexible copula

Simulated illustrations

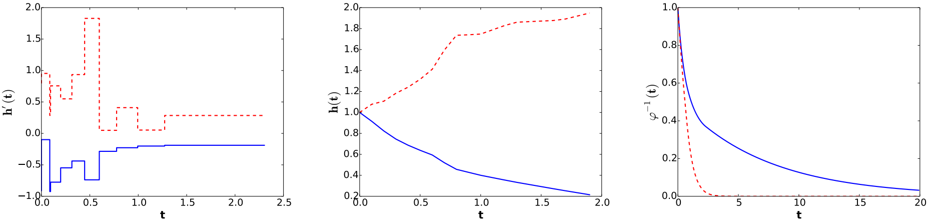


Figure – Functions $h'(t)$ (first panel), $h(t)$ (second panel) and $\varphi^{-1}(t)$ (third panel) for two scenarios of $\{\theta_k\}$. All negative values (solid line), and all positive values (dotted line).

Simulated illustrations

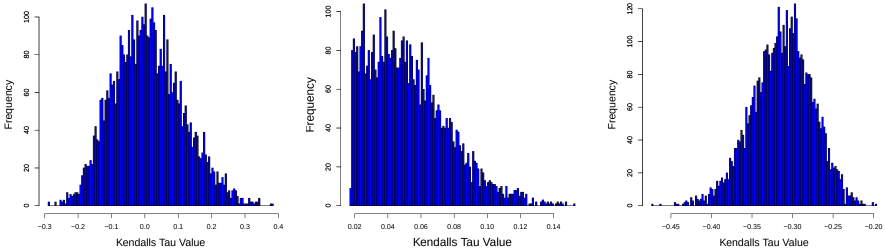


Figure – Kendall's tau prior distributions induced by our model, under three different scenarios.

Non uniqueness

- Nelsen (2006) : new generators can be defined by a scale transformation

$$\phi^{-1}(t) = \varphi^{-1}(\beta t) \quad \Longleftrightarrow \quad \phi(t) = \varphi(t)/\beta, \quad \text{for } \beta > 0$$

- Di Bernardino & Rullière (2013) : $\phi(t)$ and $\varphi(t)$ induce exactly the same copula
- Moreover, in terms of the hazard rate functions, $h_\phi(t)$ and $h_\varphi(t)$, induced by generators ϕ and φ, the relationship becomes

$$h_\phi(t) = \beta h_\varphi(\beta t).$$

- In order to make our semiparametric generator identifiable we impose the new constraint
  (C3) $\theta_0 = h(0) = 1$.

Likelihood

- The copula density $f_C(u,v)$ becomes

$$f_C(u,v) = \varphi^{-1('')}\left(\varphi(u)+\varphi(v)\right)\varphi^{(')}(u)\varphi^{(')}(v),$$

where

$$\varphi^{-1('')}(t) = \sum_{k=1}^{K}\left\{(A_k+\theta_k t)^2-\theta_k\right\}\exp\left\{-\left(B_k+A_k t+\frac{\theta_k}{2}t^2\right)\right\} I(\tau_{k-1}<t\le\tau_K)$$

and

$$\varphi^{(')}(t) = -\sum_{k=1}^{K}\frac{1}{t}\left(-2\theta_k B_k + A_k^2 - 2\theta_k\log(t)\right)^{-1/2} I\left(\varphi^{-1}(\tau_k)\le t<\varphi^{-1}(\tau_{k-1})\right).$$

- Let (U_i,V_i), $i=1,\ldots,n$ be a bivariate sample from $f_C(u,v)$.
Then the likelihood for $\boldsymbol{\theta}$ is just $\text{lik}(\boldsymbol{\theta}\mid\mathbf{u},\mathbf{v}) = \prod_{i=1}^{n} f_C(u_i,v_i\mid\boldsymbol{\theta})$

Prior distributions

- Recall that $\boldsymbol{\theta}$ must satisfy conditions : $(C1)$, $(C2)$, $(C3)$ and $\theta_K = 0$.
- Prior : Spike and slab

$$f(\boldsymbol{\theta}) \propto \prod_{k=1}^{K-1} \left\{ \pi_0 I(\theta_k = 0) + (1 - \pi_0) \mathrm{N}(\theta_k \mid \mu_0, \sigma_0^2) \right\} I(\boldsymbol{\theta} \in \Theta)$$

- We can thus define an independence test :
  H_0 : U and V independent ($\theta_1 = \cdots = \theta_{K-1} = 0$) vs.
  H_1 : U and V dependent ($\theta_k \neq 0$ for one k)
- We report $\mathrm{P}(H_1 \mid \text{data})$ as an evidence in favour of dependence

Posterior distributions

- Posterior conditionals : for $k = 1, \ldots, K-1$

$$f(\theta_k \mid \boldsymbol{\theta}_{-k}, \text{data}) \propto \text{lik}(\boldsymbol{\theta} \mid \mathbf{u}, \mathbf{v}) f(\boldsymbol{\theta})$$

- Suggest a MH step by sampling θ_k^* at iteration $(r+1)$ from a random walk proposal dist.

$$q(\theta_k \mid \boldsymbol{\theta}_{-k}, \theta_k^{(r)}) = \pi_1 I(\theta_k = 0) + (1-\pi_1)\text{Un}(\theta_k \mid \max\{a_k, \theta_k^{(r)} - \delta c_k\}, \min\{b_k, \theta_k^{(r)} + \delta c_k\})$$

where $b_k = \left(\theta_0 + \sum_{j=1}^{k-1}(\tau_j - \tau_{j-1})\theta_j\right)^2$ and
$a_k = \max_{k \le j \le K-1}\left\{\left(\sqrt{\theta_{j+1}} I(\theta_{j+1} \ge 0) - \theta_0 - \sum_{i=1, i \ne k}^{j}(\tau_i - \tau_{i-1})\theta_i\right) / (\tau_k - \tau_{k-1})\right\}$

- Parameters π_1 and δ are tuning parameters that control the acceptance rate
- We define $\{\tau_k\}$ via a Log-α partition defined by $\tau_k = -\alpha \log(1 - k/K)$ for $k = 0, \ldots, K-1$, with $\alpha > 0$

Simulation study

Clayton copula with $\theta = -0.8$, $n = 200$

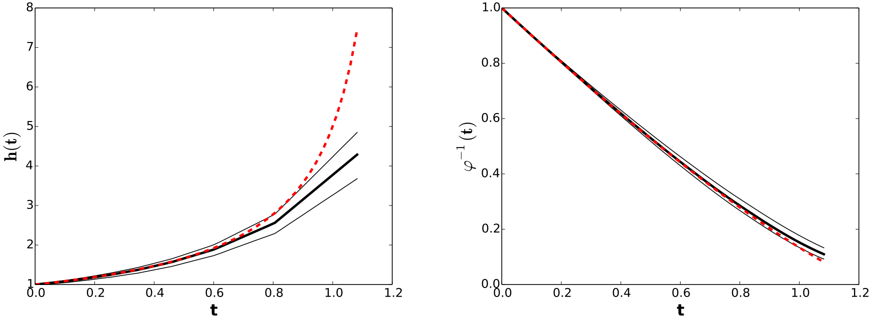


Figure – Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$, obtained with a Log-0.5 partition of size $K = 10$.

Simulation study

Clayton copula with $\theta = 1$, $n = 200$

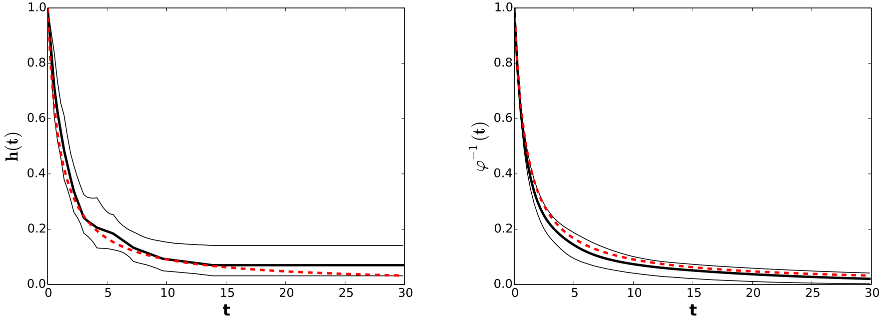
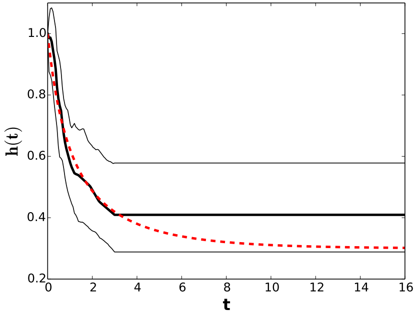


Figure – Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$, obtained with a Log-6 partition of size $K = 10$.

Simulation study

AMH copula with $\theta = 0.7$, $n = 200$



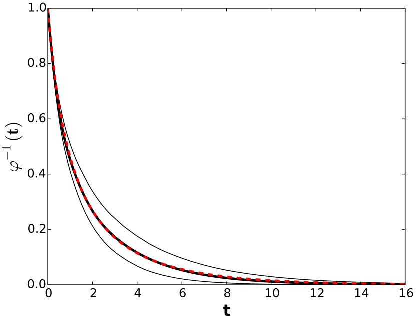


Figure – Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$, obtained with a Log-1 partition of size $K = 20$.

Simulation study

AMH copula with $\theta = -0.7$, $n = 200$

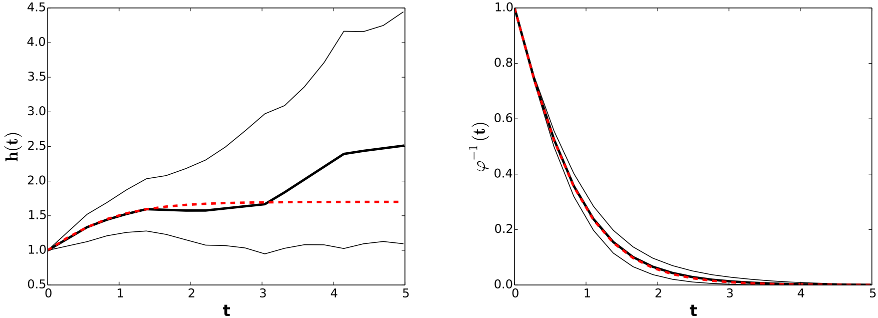


Figure – Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$, obtained with a Log-6 partition of size $K = 10$.

Simulation study

Gumbel copula with $\theta = 1.4$, $n = 200$

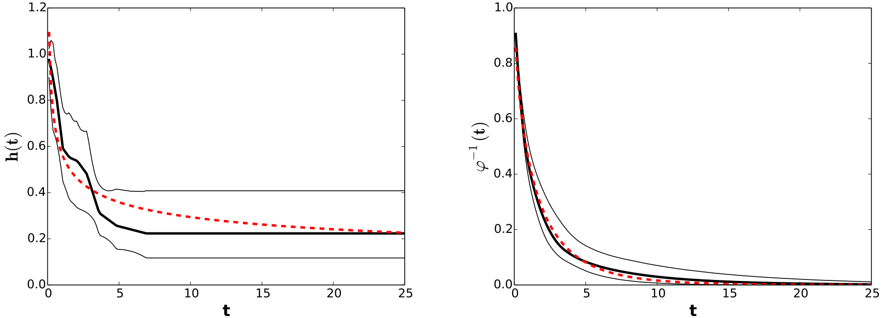


Figure – Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$, obtained with a Log-3 partition of size $K = 10$.

Real data analysis

- We study the dependence between the age of a mother (X) and the weight of her child (Y)
- Concentrated on mothers of 35 years old and above in the year 2017 in Mexico City ($n = 208$)
- Transform the original data, (X_i, Y_i), $i = 1, \ldots, n$, to the unit interval via a modified rank transformation : $U_i = \text{rank}(i, \mathbf{X})/n$ and $V_i = \text{rank}(i, \mathbf{Y})/n$ where

$$\text{rank}(i, \mathbf{X}) = k \quad \Longleftrightarrow \quad X_i = X_{(k)}$$

  for $i, k = 1, \ldots, n$
- To avoid problems due to ties we first include a perturbation by adding a $\text{Un}(-0.0, 0.01)$ to each coordinate
- The sample Kendall's tau value for the transformed data is $\tilde{\kappa}_\tau = -0.1162$.

Real data analysis

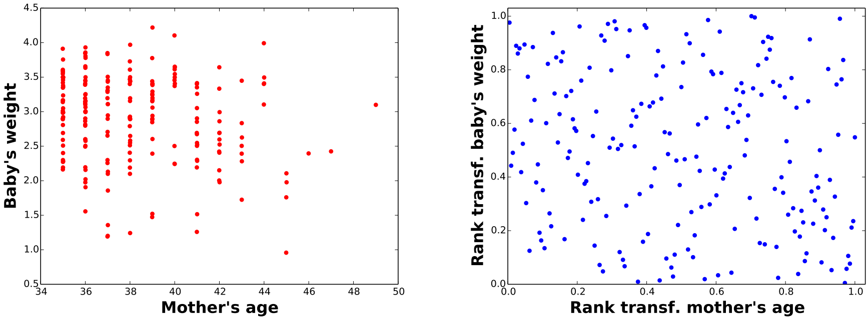


Figure – Scatter plots. Original data (left) and rank transformed data (right).

Real data analysis

- Define partitions as Log-α with $\alpha \in \{0.3, 0.5, 0.9, 1, 2, \ldots, 10\}$ and $K \in \{10, 20\}$.
- For the prior we took : $\pi_0 = 0$, $\mu_0 = -1$ and $\sigma_0^2 = 10$
- MCMC specifications :
  - For the proposal dist. $\pi_1 = 0$ and $\delta = 0.25 \Rightarrow$ an acceptance rate of around 30%.
  - Chains were ran for 20,000 iterations with a burn-in of 2,000 and keeping one of every 5^{th} iteration

Real data analysis

Table – GOF measures for the real dataset

Part.Type	K	$Q_{\kappa_\tau}^{(0.025)}$	$\hat{\kappa}_\tau$	$Q_{\kappa_\tau}^{(0.975)}$	Sample $\tilde{\kappa}_\tau$	LPML
Log-6	10	-0.212	-0.153	-0.092	-0.116	2.967
Log-8	10	-0.220	-0.153	-0.087	-0.116	2.294
Log-10	10	-0.208	-0.150	-0.094	-0.116	**3.702**
Log-6	20	-0.212	-0.153	-0.091	-0.116	0.461
Log-8	20	-0.217	-0.150	-0.097	-0.116	1.303
Log-10	20	-0.231	-0.158	-0.096	-0.116	2.046

Real data analysis

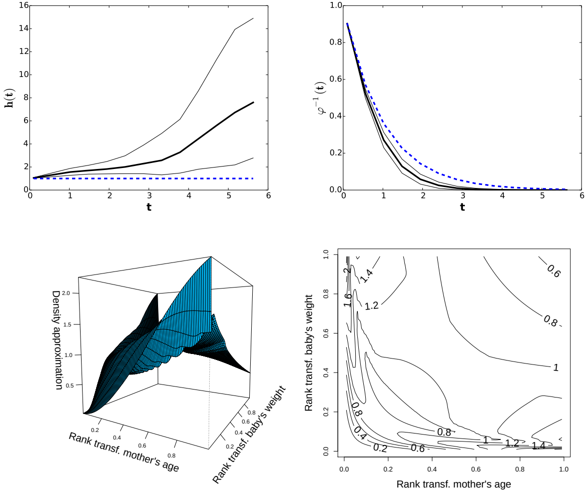


Figure – Posterior estimates obtained with a Log-10 partition of size $K = 10$.

References

- Di Bernardino, E. and Rullière, D. (2013). On certain transformation of Archimedean copulas : Application to non-parametric estimation of their generators. *Dependence Modeling* **1**, 1–36.
- Hoyos-Argüelles, R. and Nieto-Barajas, L.E. (2020). A Bayesian semiparametric Archimedean copula. *Journal of Statistical Planning and Inference* **206**, 298–311.
- Nelsen, R.B. (2006). *An introduction to copulas*. Springer, New York.