A Bayesian semiparametric Archimedean copula

RICARDO HOYOS & LUIS NIETO-BARAJAS

Department of Statistics, ITAM, Mexico

ricardo.hoyos@itam.mx and lnieto@itam.mx

Abstract

An Archimedean copula is characterised by its generator. This is a real function whose inverse behaves as a survival function. We propose a semiparametric generator based on a quadratic spline. This is achieved by modelling the first derivative of a hazard rate function, in a survival analysis context, as a piecewise constant function. Convexity of our semiparametric generator is obtained by imposing some simple constraints. The induced semiparametric Archimedean copula produces Kendall's tau association measure that covers the whole range $(-1, 1)$. Inference on the model is done under a Bayesian approach and for some prior specifications we are able to perform an independence test. Properties of the model are illustrated with a simulation study as well as with a real dataset.

Keywords: Archimedean copula, Bayes nonparametrics, piecewise constant, survival analysis, quadratic spline.

AMS Classification: 60E05 · 62G05 · 62N86.

1 Introduction

Let $\varphi(\cdot)$ be a continuous, strictly decreasing function from $[0, 1]$ to $[0, \infty)$ such that $\varphi(1) = 0$. Let $\varphi^{-1}(\cdot)$ be the inverse or the pseudo-inverse of φ, where the latter is defined as zero for $t > \varphi(0)$. If $\varphi(0) = \infty$ the generator is called strict. An Archimedean copula $C(u, v)$ with generator φ is a function from $[0, 1]^2$ to $[0, 1]$ defined as

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

A further requirement for (1) to be well defined is that φ must be convex (e.g. Nelsen, 2006).

There are many properties that characterize Archimedean copulas, for instance, they are symmetric, associative and their diagonal section $C(u,u)$ is always less than u for all $u \in (0,1)$. Generators $\varphi(\cdot)$ are usually parametric families defined by a single parameter. Most of them are summarised in Nelsen (2006, Table 4.1).

Association measures induced by Archimedean copulas are a function of the generator. For instance, Kendall's tau becomes

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

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

In this work we propose a Bayesian semiparametric generator defined through a quadratic spline. Within a survival analysis context, we model the first derivative of a hazard rate function with a piecewise constant function. The hazard rate and the cumulative hazard functions become linear and quadratic continue functions, respectively. The induced survival function is used as an inverse generator for an Archimedean copula. Convexity constrains are properly addressed and inference on the model is done under a Bayesian approach.

Other studies on semiparametric generators for Archimedean copulas can be found in Genest and Rivest (1993) where their model is based on an empirical Kendall's process. A new approach and extensions of this latter methodology can be found in Genest et al. (2011). In Guillote and Perron (2015) the model arises from the one-to-one correspondence between an Archimedean generator and a distribution function of a nonnegative random variable. In particular they use a mixture of Pólya trees as a prior for the corresponding distribution function under a Bayesian nonparametric approach. In a work more related to ours, Vanderhende and Lampert (2005) use the relationship between quantile functions and Archimedean generators to define a semiparametric generator by supplementing a parametric generator with $n+1$ dependence parameters. Differing to their work, our model is not based on any parametric generator and the Kendall's tau can take values on the whole interval $(-1, 1)$.

The contents of the rest of the paper is as follows. In Section 2 we present our proposal and characterise its properties. In Section 3 we provide details of how to make posterior inference under a Bayesian approach. In Section 4 we illustrate the performance of our model with a simulation study as well as with a real data set. We conclude with some remarks in Section 5.

2 Model

To define our proposal we realise that φ^{-1} is a decreasing function from $[0, \infty)$ to $[0, 1]$, so it behaves as a survival function, in a failure time data analysis context (e.g. Klein and Moeschberger, 2003). The idea is to propose a semi/non parametric form for the inverse generator φ^{-1} by using survival analysis ideas. For that we recall some basic definitions.

Let $h(t)$ be a nonnegative function with domain in $[0, \infty)$ such that $H(t) = \int_0^t h(s)\mathrm{d}s \to \infty$ as $t \to \infty$. Then $S(t) = \exp\{-H(t)\}$ is a decreasing function from $[0, \infty)$ to $[0, 1]$, so it behaves like an inverse generator $\varphi^{-1}(t)$. In a survival analysis context, functions $h(\cdot)$, $H(\cdot)$ and $S(\cdot)$ are the hazard rate, cumulative hazard and survival functions, respectively.

In particular, if $h(t) = \theta$, i.e. constant for all t, then $S(t) = e^{-\theta t}$. If we take $\varphi(t)^{-1} = e^{-\theta t}$, then $\varphi(t) = -(\log t)/\theta$. Using (1) we obtain that the resulting copula $C(u, v) = uv$ is the independence copula, and what is interesting, is that it does not depend on θ.

Using these ideas we construct a semiparametric generator in the following way. We first consider a partition of size K of the positive real line, with interval limits given by $0 = \tau_0 < \tau_1 < \cdots < \tau_K = \infty$. Then, we define the first derivative of the hazard rate, as a piecewise constant function of the form

$$h'(t) = \sum_{k=1}^{K} \theta_k I(\tau_{k-1} < t \leq \tau_k), \tag{3}$$

where $\theta_K \equiv 0$. We recover the hazard rate function as $h(t) = \int_0^t h'(s)\mathrm{d}s + \theta_0$, where $h(0) = \theta_0 > 0$ is an initial condition. Using (3), the hazard rate becomes a piecewise linear function of the form

$$h(t) = \sum_{k=1}^{K} (A_k + \theta_k t)\, I(\tau_{k-1} < t \leq \tau_k), \tag{4}$$

where $A_1 = \theta_0$ and $A_k = \theta_0 + \sum_{j=1}^{k-1}(\theta_j - \theta_{j+1})\tau_j$, for $k = 2, \ldots, K$.

Integrating now the hazard function (4), the cumulative hazard is a piecewise quadratic function given by

$$H(t) = \sum_{k=1}^{K} \left(B_k + A_k t + \frac{\theta_k}{2} t^2 \right) I(\tau_{k-1} < t \leq \tau_k), \tag{5}$$

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

We therefore define a semiparametric inverse generator as the induced survival function, which can be written as

$$\varphi^{-1}(t) = \exp\{-H(t)\}, \tag{6}$$

where $H(t)$ is given in (5).

After doing some algebra, we can invert this function to obtain an expression for the generator

$$\begin{aligned}\varphi(t) = \sum_{k=1}^{K} \Bigg(& \left[\mathrm{sgn}(\theta_k) \left\{ \frac{2}{\theta_k} \left(\frac{A_k^2}{2\theta_k} - B_k - \log(t) \right) \right\}^{1/2} - \frac{A_k}{\theta_k} \right] I(\theta_k \neq 0) \\ & - \frac{B_k + \log(t)}{A_k} I(\theta_k = 0) \Bigg)\, I\left(\varphi^{-1}(\tau_k) \leq t < \varphi^{-1}(\tau_{k-1}) \right).\end{aligned} \tag{7}$$

The value K controls the flexibility of the generator, and thus of the copula. If $K = 1$, the induced Archimedean copula is the independent copula, whereas for larger K, the generator, and the induced copula, become more nonparametric. Potentially K could be infinite.

We now discuss some properties of our semiparametric generator.

Proposition 1 *Consider the semiparametric inverse generator $\varphi^{-1}(t)$, given in (6), and the corresponding generator $\varphi(t)$, given in (7), and assume that $\{\theta_k\}$ are such that $\theta_0 > 0$, $\theta_K = 0$ and satisfy conditions (C1) and (C2) given by*

(C1) $A_k + \theta_k t \geq 0$, *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) $\varphi^{-1}(t)$ *and* $\varphi(t)$ *are continuous functions of* t,

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

(iii) $\varphi(t)$ *a strict generator.*

Proof For (i) we know that $h'(t)$, as in (3), is a piecewise constant discontinuous function, however, function $h(t)$, as in (4), is continuous. To see this, for each $k = 1, \ldots, K$, the limit from the left is $\lim_{t \to \tau_k^-} h(t) = \lim_{t \to \tau_k^-} A_k + \theta_k t = A_k + \theta_k \tau_k$, and the limit from the right becomes $\lim_{t \to \tau_k^+} h(t) = \lim_{t \to \tau_k^+} A_{k+1} + \theta_{k+1} t = A_{k+1} + \theta_{k+1} \tau_k$. Since $A_{k+1} = A_k + (\theta_k - \theta_{k+1})\tau_k$, then both limits coincide. For (ii) we take the second derivative of $\varphi^{-1}(t)$ which becomes $\varphi^{-1('')}(t) = \{h(t)\}^2 \exp\{-H(t)\} - h'(t) \exp\{-H(t)\}$, this is positive if and only if $\{h(t)\}^2 - h'(t) > 0$. For this to happen we require condition $(C2)$. For (iii), $\varphi^{-1}(t)$ must be a proper survival function, that is, $h(t)$ must be nonnegative, which is achieved by imposing condition $(C1)$. Furthermore, we need $\lim_{t \to \infty} \varphi^{-1}(t) = 0$, which is equivalent to prove that $\lim_{t \to \infty} H(t) = \lim_{t \to \infty} (B_K + A_K t + \theta_K t^2/2) = \infty$. This is true since B_K is a finite constant, $A_K > 0$ and $\theta_K = 0$, so the linear part goes to infinity when $t \to \infty$. $\diamond$

To see the kind of association induced by our proposal, we computed the Kendall's tau using expression (2) with generator (7). This is given in the following result.

Proposition 2 *The Kendall's tau obtained by the Archimedean copula with semiparametric generator* (7) *is given by*

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

Moreover, this $\kappa_\tau \in (-1, 1)$.

Proof Rewriting expression (2) in terms of the inversed generator we obtain $\kappa_\tau = 1 - 4\int_0^\infty t\{\varphi^{-1(\prime)}(t)\}^2 \mathrm{d}t$. Computing the derivative we get $\varphi^{-1(\prime)}(t) = -\sum_{k=1}^{K}(A_k + \theta_k t)\times \exp\{-(B_k + A_k t + \theta_k t/2)\}I(\tau_{k-1} < t \leq \tau_k)$. Doing the integral we obtain the expression. To obtain the range of possible values of κ_τ it is easier to re-write κ_τ in terms of $h(t)$ and $H(t)$. This becomes $\kappa_\tau = -2\int_0^\infty t h'(t)\exp\{-H(t)\}\,dt$. Here it is straightforward to see that the sign of κ_τ is determined by the sign of $h'(t)$, therefore $h'(t) > 0$ for all t implies $-1 < \kappa_\tau < 0$ and $h'(t) \leq 0$ implies $0 \leq \kappa_\tau < 1$. $\diamond$

The expression for κ_τ tells us that the concordance induced by our semiparametric copula is a function of both, the parameters $\{\theta_k\}$, as well as of the partition limits $\{\tau_k\}$. It depends on a definite integral and can be evaluated numerically. What is more important is that κ_τ covers the whole range from -1 to 1, showing that our proposal is very flexible.

To illustrate the flexibility of our model we define a partition of the positive real line of size $K = 10$, such that $\tau_k = -\log(1 - k/10)$ for $k = 0, 1, \ldots, 10$. We consider two scenarios for the values of the parameters $\{\theta_k\}$. The first scenario is defined by $\theta_k < 0$ for all k, whereas the second scenario contains $\theta_k > 0$ for all k. Conditions $(C1)$ and $(C2)$ were satisfied in both cases. Figure 1 contains functions $h'(t)$, $H(t)$ and $\varphi^{-1}(t)$ for two different scenarios, the solid (blue) line corresponds to the first scenario and the dotted (red) line to the second scenario. In the first case the corresponding hazard function (middle panel) is decreasing, whereas for the second case the hazard function is increasing. The induced concordance values are $\kappa_\tau = 0.368$ and $\kappa_\tau = -0.202$, respectively.

As a second example, we consider a partition of size $K = 50$, such that $\tau_k = -\log(1 -$

$k/50)$ for $k = 0, 1, \ldots, 50$. We consider three different scenarios for the parameters $\{\theta_k^{(i)}\}$ with $i = 1, 2, 3$, respectively. In the first scenario we assume $\theta_1^{(1)} \sim \text{Un}(-1, 1)$, in the second $\theta_1^{(2)} \sim \text{Un}(-50, 0)$ and in the third $\theta_1^{(3)} \sim \text{Un}(0, 1)$. Posteriorly, we define sequentially $\theta_k^{(i)} \sim \text{Un}(a_k^{(i)}, b_k^{(i)})$ with $a_k^{(i)}$ and $b_k^{(i)}$ constants such that constrains $(C1)$ and $(C2)$ are satisfied, for $k = 2, \ldots, K-1$ and $i = 1, 2, 3$. We repeated sampling from these distributions a total of 5,000 times, and for each repetition we computed κ_τ. The induced histogram densities for the three scenarios are presented in Figure 2. For the first scenario, the values of κ_τ range from -0.3 to 0.4, showing that our model can capture both negative and positive concordance measures. For the second scenario, the values of κ_τ are all positive and the distribution is right skewed, and for the third scenario the values of κ_τ are all negative showing a left skewed distribution.

According to (Nelsen, 2006), new generators can be defined if we apply a scale transformation of the form $\phi^{-1}(t) = \varphi^{-1}(\alpha t)$ iff $\phi(t) = \varphi(t)/\alpha$, for $\alpha > 0$, where $\phi(t)$ becomes a new Archimedean copula generator. More recently, (Di Bernardino and Rullière, 2013) realised that the new generator $\phi(t)$ induces exactly the same copula (1) as that obtained with $\varphi(t)$. To see this we have that $C_\phi(u, v) = \phi^{-1}(\phi(u) + \phi(v)) = \varphi^{-1}\left(\alpha\left\{\frac{1}{\alpha}\varphi(u) + \frac{1}{\alpha}\varphi(v)\right\}\right) = C_\varphi(u, v)$. In other words, an Archimedean copula generator is not unique.

Moreover, in terms of the hazard rate functions, $h_\phi(t)$ and $h_\varphi(t)$, induced by generators ϕ and φ, respectively, the relationship becomes $h_\phi(t) = \alpha h_\varphi(\alpha t)$. In order to make our semi-parametric generator identifiable, without loss of generality, we impose the new constraint

(C3) $\theta_0 = 1$.

This constraint is equivalent to impose $h(0) = 1$ in definition (4).

3 Posterior inference

The copula density $f_C(u,v)$, of an Archimedean copula, can be obtained by taking the second crossed derivatives with respect to u and v in expression (1). In terms of the generator and its inverse this density becomes

$$f_C(u,v) = \varphi^{-1('')}\left(\varphi(u)+\varphi(v)\right)\varphi^{(')}(u)\varphi^{(')}(v), \tag{8}$$

where the single and double primes denote first and second derivatives, respectively, and are given by

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

and

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

Let (U_i,V_i), $i=1,\ldots,n$ be a bivariate sample of size n from $f_C(u,v)$ defined in (8). With this we can construct the likelihood for $\boldsymbol{\theta}=(\theta_0,\theta_1,\ldots,\theta_K)$ as $\text{lik}(\boldsymbol{\theta}\mid\mathbf{u},\mathbf{v})=\prod_{i=1}^{n}f_C(u_i,v_i\mid\boldsymbol{\theta})$, where we have made explicit the dependence on $\boldsymbol{\theta}$ in the notation of the copula density. Recall that the parameters must satisfy several conditions, $(C1)$ and $(C2)$ given in Proposition 1, $(C3)$ to make our generator unique, and $\theta_K=0$.

We assume a prior distribution for the θ_k's of the form

$$f(\theta_k) = \pi_0 I(\theta_k=0)+(1-\pi_0)\text{N}(\theta_k\mid\mu_0,\sigma_0^2), \tag{9}$$

independently for $k=1,\ldots,K-1$.

Note that we explicitly allow the θ_k's, for $k=1,\ldots,K-1$ to be zero with positive probability π_0. This prior choice is useful to define an independence test. Specifically, the hypothesis $H_0: U$ and V independent is equivalent to $H_0:\theta_1=\cdots=\theta_{K-1}=0$. To perform the test we can compute the posterior probability of H_0 and its complement and make the decision, say via Bayes factors (Kass and Raftery, 1995).

The posterior distribution of $\boldsymbol{\theta}$ is simply given by the product of expressions (8) and (9), up to a proportionality constant. It is somehow easier to characterize the posterior distribution by implementing a Gibbs sampler (Smith and Roberts, 1993) and sampling from the conditional posterior distributions

$$f(\theta_k \mid \boldsymbol{\theta}_{-k}, \text{data}) \propto \text{lik}(\boldsymbol{\theta} \mid \mathbf{u}, \mathbf{v}) f(\theta_k), \tag{10}$$

for $k = 1, \ldots, K-1$. However, sampling from conditional distributions (10) is not trivial, we therefore propose a Metropolis-Hastings step (Tierney, 1994) by sampling θ_k^* at iteration $(r+1)$ from a random walk proposal distribution

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

where the interval (a_k, b_k) represents the conditional support of θ_k, $c_k = b_k - a_k$ is its length, with $a_k = \max_{k \le j \le K-1} \left\{ \left(\sqrt{\theta_{j+1}} I(\theta_{j+1} \ge 0) - \theta_0 - \sum_{i=1, i\neq k}^{j} (\tau_i - \tau_{i-1})\theta_i \right) / (\tau_k - \tau_{k-1}) \right\}$, for $k = 1, \ldots, K-1$, $b_k = \left(\theta_0 + \sum_{j=1}^{k-1} (\tau_j - \tau_{j-1})\theta_j \right)^2$, for $k = 2, \ldots, K-1$, and $b_1 = 1$. The justification of these bounds obeys the inclusion of constraints $(C1)$ and $(C2)$ and their derivations are given in Appendix A. The parameters π_1 and δ are tuning parameters that control de acceptance rate.

Therefore, at iteration $r+1$ we accept θ_k^* with probability

$$\alpha\left(\theta_k^*, \theta_k^{(r)}\right) = \min\left\{1\,,\ \frac{f(\theta_k^* \mid \boldsymbol{\theta}_{-k}, \text{data})\, q(\theta_k^{(r)} \mid \boldsymbol{\theta}_{-k}, \theta_k^*)}{f(\theta_k^{(r)} \mid \boldsymbol{\theta}_{-k}, \text{data})\, q(\theta_k^* \mid \boldsymbol{\theta}_{-k}, \theta_k^{(r)})}\right\}.$$

4 Numerical studies

We illustrate the performance of our model in two ways, through a simulation study, and with a real data set.

To define the partition $\{\tau_k\}$ of the positive real line we consider a Log-α partition defined by $\tau_k = -\alpha \log(1 - k/K)$ for $k = 0, \ldots, K-1$, with $\alpha > 0$. This partition is the result of transforming a uniform partition in the interval $[0,1]$ via a convex function. In particular we inspired ourselves in the generator of the product copula. Larger values of α increase the spread of the partition along the positive real line.

4.1 Simulation study

We generated simulated data from four parametric Archimedean copulas, namely the product, Clayton, Ali-Mikhail-Haq (AMH) and Gumbel copulas. Their features are summarised in Table 1, where we include the parameter space, the generator, the inverse generator, an indicator whether the copula is strict or not and the induced $h(t)$ function obtained through inversion of relationship (6).

For each parametric copula we took a sample of size $n = 200$. To specify the copulas we took $\theta \in \{-0.8, 1\}$ for the Clayton copula, $\theta \in \{-0.7, 0.7\}$ for the AMH copula, and $\theta = 1.4$ for the Gumbel copula. For the partition size we compared $K \in \{10, 20\}$ and tried values $\alpha \in \{0.3, 0.5, 0.9, 1, 2, \ldots, 10\}$.

For the prior distributions (9) we took $\pi_0 = 0, \mu_0 = -1$ and $\sigma_0^2 = 10$. We implemented a MH step with-in the Gibbs sampler where the proposal distributions were specified by $\pi_1 = 0$ and $\delta = 0.25$. The acceptance rate attained with these specifications are around 30%, which according to (Robert and Casella, 2010) are optimal for random walks. Finally, the chains were ran for 20,000 iterations with a burn-in of 2,000 and keeping one of every 5^{th} iteration to produce posterior estimates.

To assess goodness of fit (GOF) we computed several statistics. The logarithm of the pseudo marginal likelihood (LPML), originally suggested by (Geisser and Eddy, 1979), to assess the fitting of the model to the data. The supremum norm, defined by $\sup_t |\varphi^{-1}(t) - \widehat{\varphi}^{-1}(t)|$ to assess the discrepancy between our posterior estimate (posterior mean) $\widehat{\varphi}^{-1}(t)$ from the true inverse generator $\varphi^{-1}(t)$. We also computed the Kendall's tau coefficient and compare the posterior point and 95% interval estimates with the true value. These values are shown in Tables 2 to 7. Although we fitted our model with all values of α mentioned above, we only show results for some of them in the tables.

Note that, due to the nonunicity of an Archimedean generator, an equivalent constraint to $(C3)$ has to be imposed to the parametric generators that we are comparing to. That is we set $h(0) = 1$ for the product, Clayton and AMH copulas, and $h(\epsilon) = 1$ for the Gumbel copula, for say $\epsilon = 0.01$. The difference in the latter case is because, for a Gumbel copula, $h(t) \to \infty$ when $t \to 0$. These conditions are already included in the parametrisation used in Table 1.

For the product copula the GOF measures are presented in Table 2. With exception of the partition Log-3 for $K = 30$, for all settings considered, the true κ_τ lies inside the 95% credible intervals. The LPML chooses the model with Log-1 partitions of size $K = 10$, and corresponds to the second smallest value of the supremum norm. Posterior estimates of functions $h(t)$ and $\varphi^{-1}(t)$ are shown in Figure 3. In both cases the true function lies inside the 95% credible intervals.

For the Clayton copula we have two choices of θ, -0.8 and 1. The first choice, $\theta = -0.8$, corresponds to a generator that is not strict, that is, $\varphi^{-1}(t) > 0$ for $t \in [0, 5/4]$, and $\varphi^{-1}(t) = 0$ for $t > 5/4$. This is an interesting challenge because our model defined only strict generators. The settings with smallest supremum norm, Log-0.5 with $K = 10$, produces the 95% credible interval for κ_τ closest to the true value, however it does not achieve the largest LPML. The inconsistency of the GOF measures might be due to the non strictness feature of the true generator. Moreover, if we look at the graphs of the posterior estimates of $h(t)$ and $\varphi^{-1}(t)$ (Figure 4), for larger values of t the true functions lie outside of our posterior estimates. For $\theta = 1$, the best model is obtained with a Log-6 partition of size 10. In this case, posterior estimates of functions $h(t)$ and $\varphi^{-1}(t)$ with the best fitting (Figure 5), contain the true functions.

For the AMH copula we have two values of θ, -0.7 and 0.7. The best fitting consistently chosen by the three GOF criteria is obtained with a Log-6 and Log-1 partitions of size $K = 10$, respectively for the two values of θ. Posterior estimates of functions $h(t)$ and $\varphi^{-1}(t)$ with the best fitting are shown in Figures 7 and 6, respectively. In all cases the true functions lie within the 95% credible intervals.

For the Gumbel copula with $\theta = 1.4$ we have an interesting behaviour. The true $h(t)$ function has the feature that $h(0) = \infty$. This represents a challenge for our model since we have imposed the constrain $(C3)$ which is equivalent to $h(0) = 1$. The highest LPML value is obtained with a Log-7 partition of size $K = 10$, however the posterior 95% credible interval for κ_τ does not contain the true value. On the other hand, the second best value of LPML is obtained with an Log-3 partition of size 10, and in this case the 95% credible interval for κ_τ does contain the true value. We select this latter as the best fitting. Posterior estimates of functions $h(t)$ and $\varphi^{-1}(t)$ are shown in Figure 8. Recalling that the true hazard function goes asymptotically to infinity when $t \to 0$, therefore, for values close to zero the true $h(t)$ lies outside our posterior credible intervals, something similar happens in the estimates of the inverse generator. Apart from this, our posterior estimates are very good for $t > \epsilon$.

An important learning from the previous examples is that increasing the partition size doe not necessarily implies a better fitting.

4.2 Real data analysis

In public health it is important to study the factors that determine the birth weight of a child. Low birth weight is associated with high perinatal mortality and morbility (e.g. Stevens-Simon and Orleans, 2001). We study the dependence structure between the age of a mother (X) and the weight of her child (Y), and concentrated on mothers of 35 years old and above. The dataset was obtained from the General Hospital of Mexico through the opendata platform that can be accessed at https://datos.gob.mx/busca/dataset/perfiles-metabolicos-neonatales/resource/4ab603eb-b73a-498f-8c56-0dc6d21930e8. It contains $n = 208$ records of the neonatal metabolic profile of male babies registered in the year 2017 in Mexico City.

The marginal distributions for variables U and V, induced by copula (1), are uniform. In practice, copulas are used to model the dependence for any pair of random variables regardless of their marginal distributions. Let X and Y be two random variables with marginal cumulative distributions $F(x)$ and $G(y)$ respectively. Then the joint cumulative distribution function for (X,Y) is obtained as (Sklar, 1959), $H(x,y)=C(F^{-1}(x),G^{-1}(y))$, where C is given in (1).

Since we are just interested in modelling the dependence between X and Y, it is common in practice to transform the original data, (X_i,Y_i), $i=1,\ldots,n$, to the unit interval via a modified rank transformation (Deheuvels, 1979) in the following way. Let $\mathbf{X}'=(X_1,\ldots,X_n)$ and $\mathbf{Y}'=(Y_1,\ldots,Y_n)$ then $U_i=\text{rank}(i,\mathbf{X})/n$ and $V_i=\text{rank}(i,\mathbf{Y})/n$ are the transformed data, where $\text{rank}(i,\mathbf{X})=k$ iff $X_i=X_{(k)}$ for $i,k=1,\ldots,n$. This is based on the probability integral transform using the empirical cumulative distribution function of each coordinate.

In Figure 9 we show a dispersion diagram of the original data (left panel) and the rank transformed data (right panel). To avoid problems due to ties in the original data, we fist include a perturbation to the data by adding a uniform random variable $\text{Un}(0,0.01)$ to each coordinate. The sample Kendall's tau value for the transformed data is $\tilde{\kappa}_\tau=-0.1122$.

We fitted our model to the transformed data with the following specifications. To define the partitions we took values $\alpha\in\{0.3,0.5,0.9,1,2,\ldots,10\}$ with sizes $K\in\{10,20\}$. For the prior we took $\pi_0=0$, $\mu_0=-1$ and $\sigma_0^2=10$. The MCMC specifications were the same as those used for the simulated data.

The GOF measures computed were the LPML and the posterior estimates (point and 95% credible interval) of κ_τ. The results are reported in Table 8. The best fitting model according to LPML is that obtained with a partition of size $K=10$ and Log-10. The sample concordance $\tilde{\kappa}_\tau$ is included in our posterior 95% credible interval estimate $\kappa_\tau\in(-0.213,-0.098)$.

The estimated hazard rate function $h(t)$ and the inverse generator $\varphi^{-1}(t)$, with the best fitting model, are included in Figure 9. The solid thick line corresponds to the point estimates and the solid thin lines to the 95% credible intervals. For a visual comparison, the blue dotted line corresponds to the function of the independence (product) copula. This confirms that there is a negative (weak) dependence between the age of the mother and the birth weight of the child. The older the mother, the less weight of the child. This finding could potentially help the policy makers to focus campaigns to help the awareness of future mothers.

As mentioned in Section 2, we can use our model to undertake an independence test. For that we choose the prior distribution for the θ_k's, as in (9), such that the prior probability of $H_0 : \theta_1 = \cdots = \theta_{K-1} = 0$ is around 1/2, in other words, we want $\mathrm{P}(H_0) = \pi_0^{K-1} = 1/2$. Particularly, for a partition of size $K = 10$ we need to specify $\pi_0 = 0.9258$. In order to get a point of mass proposal in the Metropolis-Hasting step we consider $\pi_1 = 0.3$. We re-ran our model using these values with the other specifications left unchanged. The posterior probability of H_0 becomes $\mathrm{P}(H_0 \mid \text{data}) = 0.53$, which leads to an inconclusive test. This might be explained by the fact that although the association is negative, it is very close to zero. To calibrate our independence test, we consider the simulated data from previous section of two copulas, the product copula and the Clayton copula with $\theta = -0.8$. In these cases the best fitting was obtained with a partition of size $K = 10$, so we chose the same prior values as for the real data to perform an independence test. Posterior probabilities of H_0 are 0.81 and 0, respectively, showing that for independent data the posterior probability is a lot larger that 0.5, whereas for clearly dependent data the posterior probability of independence is zero.

5 Concluding remarks

We have proposed a semiparametric Archimedean copula that is flexible enough to capture the behaviour of several families of parametric arquimedean copulas. Our model is capable of modelling positive and negative dependence. The number of parameters in the model to produce a good estimation of the dependence in the data should not be extremely high. For most of the examples considered here 10 parameters are enough.

Defining an appropriate partition to analyse real data sets is not trivial. We suggest to try different values of α in a wide range and compare using a GOF criteria like the LPML we used here.

In the exposition and in examples considered here, we concentrated on bivariate copulas, however extensions to more than two dimensions is also possible, say $C(u_1, \ldots, u_m) = \varphi^{-1}\left(\varphi(u_1) + \cdots + \varphi(u_m)\right)$. Performance of our semiparametric copula in this multivariate setting is worth studying.

Our model is motivated by semiparametric proposals for survival analysis functions (Nieto-Barajas and Walker, 2002) and appropriately modified to satisfy the properties of an Archimedean generator. The semiparametric generator presented here turned out to be based on quadratic splines, however, alternative proposals are possible.

Instead of starting with a piecewise function for the derivative of a hazard rate, we could start by defining a piecewise constant function for the hazard rate itself. That is $h(t) = \sum_{k=1}^{K} \theta_k I(\tau_{k-1} < t \leq \tau_k)$ with $\theta_k > 0$, and $\{\tau_k\}$ a partition of the positive real line. In this case the cumulative hazard function becomes $H(t) = \sum_{j=1}^{k} \theta_j \Delta_j + \theta_k(t - \tau_{k-1})$, for $t \in (\tau_{k-1}, \tau_k]$, with $\Delta_j = \tau_j - \tau_{j-1}$. The inverse generator is then a linear spline of the form

$$\varphi^{-1}(t) = \exp\left\{-\sum_{k=1}^{K} \theta_k w_k(t)\right\},$$

with

$$w_k(t) = \begin{cases} \Delta_k, & t > \tau_k \\ t - \tau_{k-1} & t \in (\tau_{k-1}, \tau_k] \\ 0 & \text{otherwise} \end{cases}$$

and the corresponding Archimedean generator has the form

$$\varphi(t) = \sum_{k=1}^{K} \left\{\tau_{k-1} - \frac{1}{\theta_k}(\log t + \vartheta_{k-1})\right\} I(\vartheta_{k-1} < -\log t \leq \vartheta_k),$$

with $\vartheta_k = \sum_{j=1}^k \theta_j \Delta_j$. To ensure convexity of the generator we further require $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_K$. Furthermore, the Kendall's tau has a simpler expression

$$\kappa_\tau = 1 + \sum_{k=1}^{K} \left\{ e^{-2\vartheta_k}(1 + 2\theta_k \tau_k) - e^{-2\vartheta_{k-1}}(1 + 2\theta_k \tau_{k-1}) \right\}.$$

However, it can be shown that this expression for the Kendall's tau only allows positive values, constraining the possible associations captured by the model.

Acknowledgements

This research was supported by *Asociación Mexicana de Cultura, A.C.*

A Derivation of posterior conditional support of θ_k

In order to satisfy constraint $(C1)$, we consider first the case $\theta_k \leq 0$. Therefore $\min_{t \in (\tau_{k-1}, \tau_k]} A_k + \theta_k t = A_k + \theta_k \tau_k$. This implies the following constraint for θ_k,

$$\theta_k \geq \max_{k \leq j \leq K-1} \left\{ - \left(\theta_0 + \sum_{i=1, i \neq k}^{j} (\tau_i - \tau_{i-1}) \theta_i \right) / (\tau_k - \tau_{k-1}) \right\},$$

for $k = 1, \ldots, K-1$, where we define the empty sum as zero.

On the order hand, if $\theta_k > 0$ we have $\min_{t \in (\tau_{k-1}, \tau_k]} A_k + \theta_k t = A_k + \theta_k \tau_{k-1}$, and we get, from condition $(C2)$, the following restriction

$$\theta_k < \left(\theta_0 + \sum_{i=1}^{k-1} (\tau_i - \tau_{i-1}) \theta_i \right)^2.$$

This defines the upper bound b_k, for $k = 2, \ldots, K-1$, and $b_1 = 1$.

Because the term θ_k appears on the right side of the previous inequality for $j = k+1, \ldots, K-1$, we need to consider the following restriction

$$\theta_k > \left(\sqrt{\theta_j} - \theta_0 - \sum_{i=1, i \neq k}^{j-1} (\tau_i - \tau_{i-1}) \theta_i \right) \Big/ (\tau_k - \tau_{k-1})$$

if $\theta_j \geq 0$. Combining this with the constraint when $\theta_k \leq 0$ above, we get the lower bound a_k for $k = 1, \ldots, K-1$.

References

Deheuvels, P. (1979). La fonction de dépendance empirique et ses proprietés. Un test non paramétrique d'indépendance. *Roy. Belg. Bull. Cl. Sci.*, **65**, (5), 274–292.

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.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

Genest, C., Nešlehová J. and Ziegel J. (2011). Inference in multivariate Archimedean copula models. *TEST* **20**: 223. https://doi.org/10.1007/s11749-011-0250-6.

Genest, C. and Rivest, L. (1993). Statistical inference procedures for bivariate Archimedean copulas . *Journal of the American Statistical Association*, **88**, 1034-1043.

Guillote, S. and Perron, J. (2015). Inference on Archimedean copulas using mixtures of Pólya trees. *Journal of Statistical Planning and Inference*, **166**, 2-13.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association* **90**, 773–795.

Klein, J.P. and Moeschberger, M.L. (2003). *Survival analysis*. Springer, New York.

Nelsen, R.B. (2006). *An introduction to copulas*. Springer, New York.

Neuhaus, G. (1971). On weak convergence of stochastic processes with multidimensional time parameter. *The Annals of Mathematical Statistics* , **42** (4), 1285-1295.

Nieto-Barajas, L.E. and Walker, S.G. (2002). Markov beta and gamma processes for modeling hazard rates. *Scandinavian Journal of Statistics* **29**, 413–424.

Robert, C. P. and Casella, G. (2010). *Introducing Monte Carlo methods with R.* Springer, New York.

Sklar, M. (1959). Fonctions de répartition á n dimensions et leurs marges. Université Paris 8.

Smith, A. and Roberts, G. (1993). Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **55**, 3-23.

Stevens-Simon, C. and Orleans, M. (2001). Low-birthweight prevention programs: The enigma of failure. *Birth* **26**, 184–191.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics* **22**, 1701-1722.

Vanderhende, F. and Lampert, P. (2005). Local dependence estimation using semiparametric Archimedean copulas. *The Canadian Journal of Statistics* **33**, 377–388.

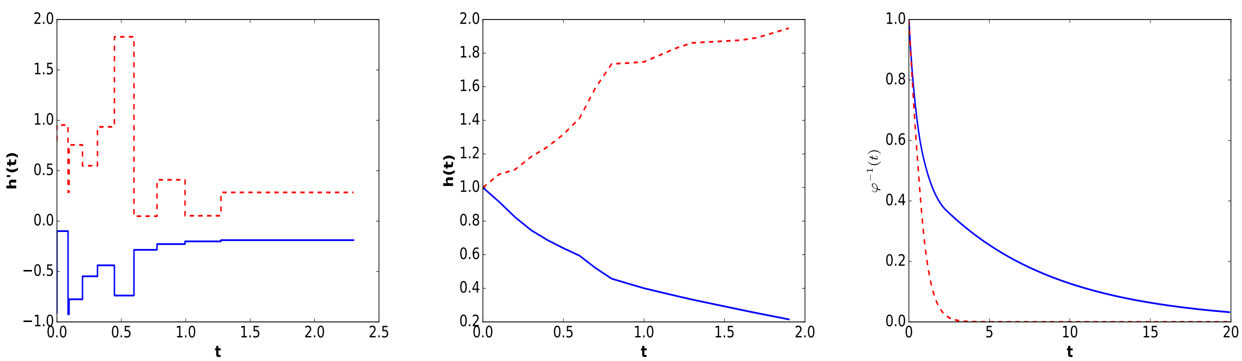

Figure 1: Functions $h'(t)$ (first panel), $h(t)$ (second panel) and $\varphi^{-1}(t)$ (third panel) for two scenarios of $\{\theta_k\}$. All negative values (solid line), and all positive values (dotted line).

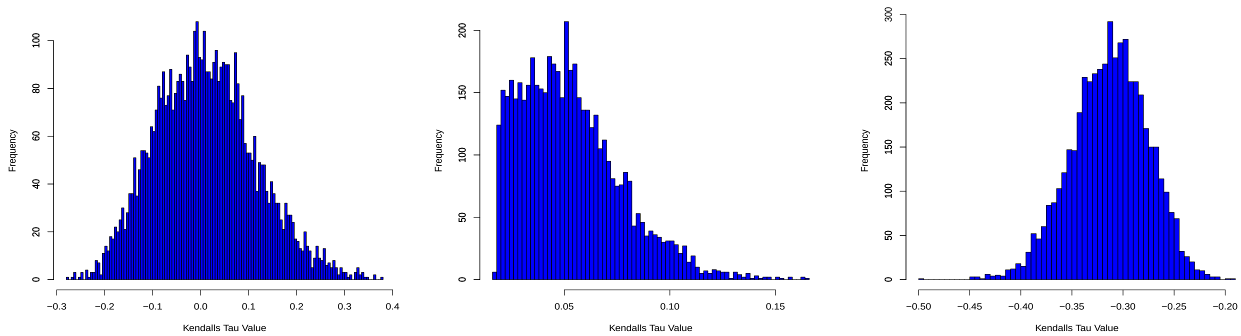

Figure 2: Prior distributions of Kendall's tau under three different scenarios.

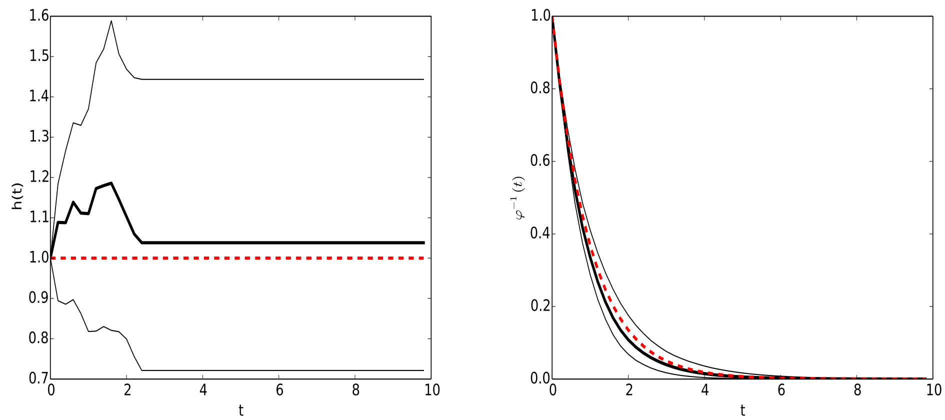

Figure 3: Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$ for simulated data from the product copula with Log-1 partition of size $K = 10$.

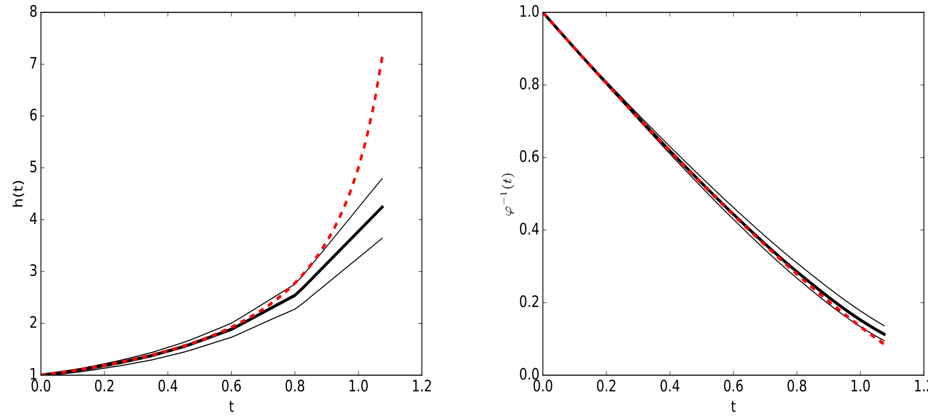

Figure 4: Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$ for simulated data from the Clayton copula with $\theta = -0.8$ and Log-0.5 partition of size $K = 10$.

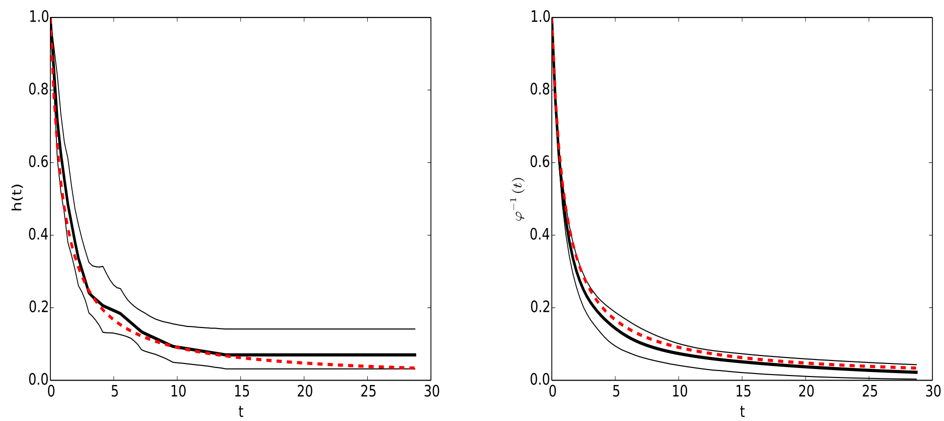

Figure 5: Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$ for simulated data from the Clayton copula with $\theta = 1$ and Log-6 partition of size $K = 10$.

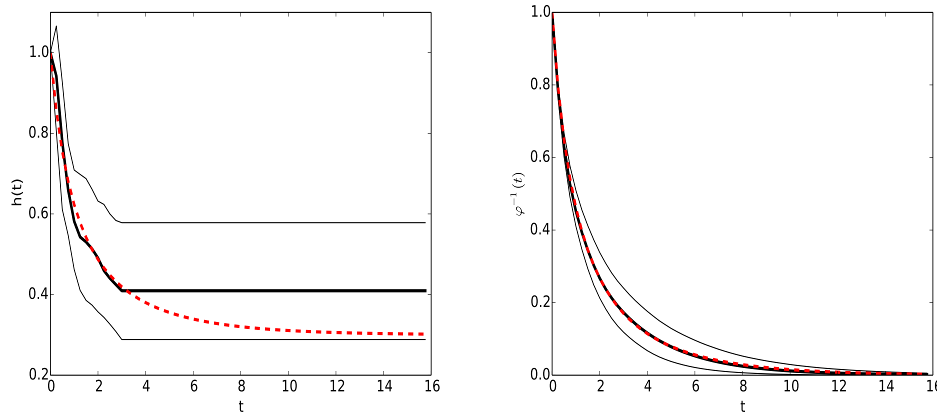

Figure 6: Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$ for simulated data from the AMH copula with $\theta = 0.7$ and Log-1 partition of size $K = 20$.

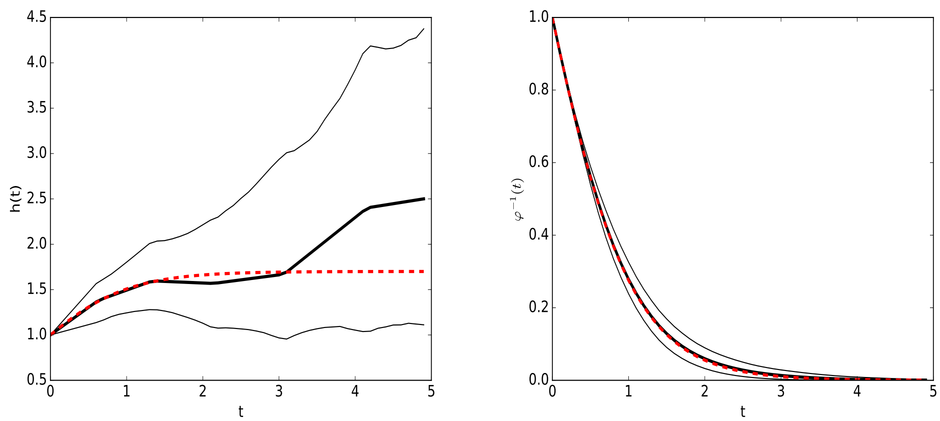

Figure 7: Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$ for simulated data from the AMH copula with $\theta = -0.7$ and Log-6 partition of size $K = 10$.

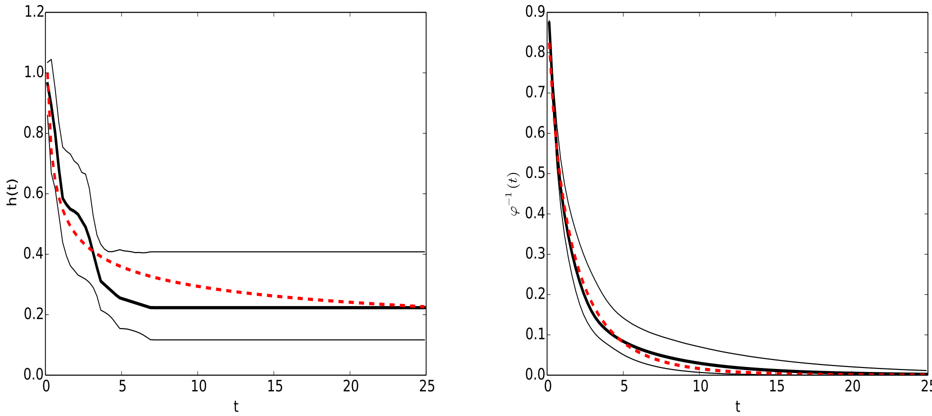

Figure 8: Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$ for simulated data from the Gumbel copula with $\theta = 1.4$ and Log-3 partition of size $K = 10$.

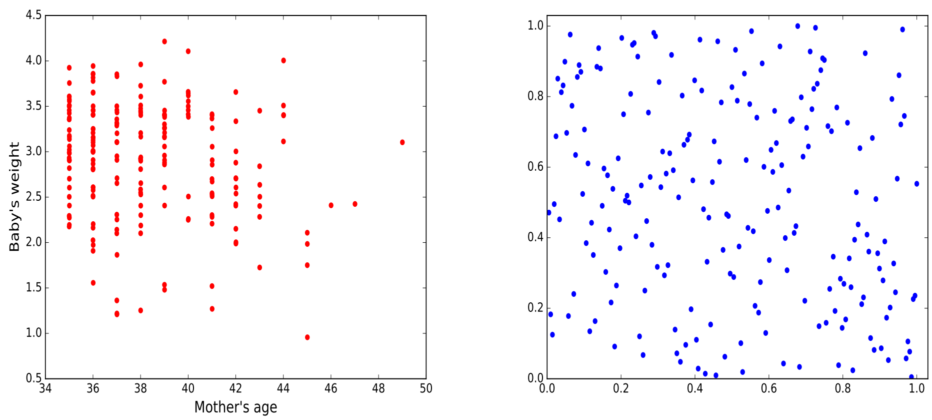

Figure 9: Dispersion plots. Original data (left) and rank transformed data (right).

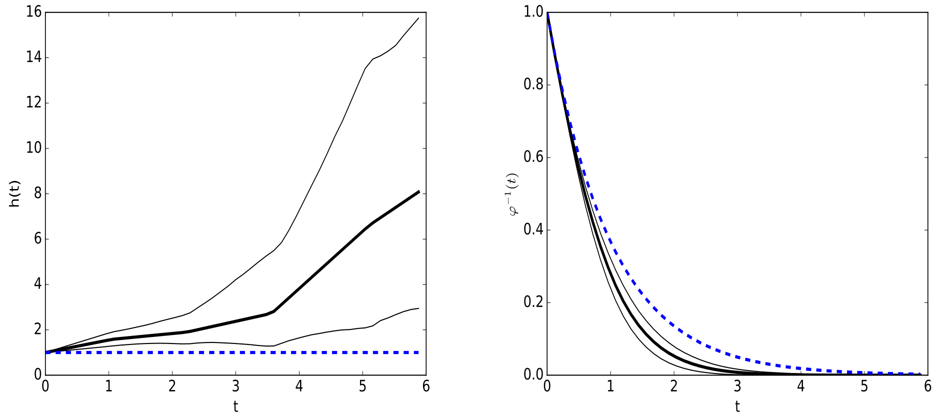

Figure 10: Posterior estimates of $h(t)$ and $\varphi^{-1}(t)$ for the real data (solid lines) with Log-10 partition of size $K = 10$. Corresponding functions from the product copula (dotted lines).

Table 1: Summary of some parametric Archimedean copulas parametrised such that $h(0) = 1$, for the first three copulas, and $h(\epsilon) = 1$, for the Gumbel copula.

Copula	Θ	$\varphi(t)$	$\varphi^{-1}(t)$	Estrict?	h(t)
Product	-	$-\log(t)$	e^{-t}	Yes	1
Clayton	$[-1, \infty)$	$\frac{1}{\theta}(t^{-\theta} - 1)$	$(1 + \theta t)^{-1/\theta}$	If $\theta \geq 0$	$\frac{1}{1+\theta t}$
AMH	$[-1, 1)$	$\frac{1}{1-\theta} \log\left(\frac{1-\theta+\theta t}{t}\right)$	$\frac{1-\theta}{e^{(1-\theta)t}-\theta}$	Yes	$\frac{(1-\theta)e^{(1-\theta)t}}{e^{(1-\theta)t}-\theta}$
Gumbel	$[1, \infty)$	$\epsilon\left(\frac{-\log(t)}{\epsilon\theta}\right)^{\theta}$	$\exp\left\{-\epsilon\theta\left(\frac{t}{\epsilon}\right)^{1/\theta}\right\}$	Yes	$\left(\frac{\epsilon}{t}\right)^{1-1/\theta}$

Table 2: GOF measures for simulated data from the product copula.

Part.Type	K	$Q_{\kappa_\tau}^{(0.025)}$	$\hat{\kappa}_\tau$	$Q_{\kappa_\tau}^{(0.975)}$	κ_τ	Sup.Norm	LPML
Log-1	10	-0.102	-0.029	0.056	0	0.045	-2.071
Log-2	10	-0.122	-0.054	0.013	0	0.061	-2.210
Log-3	10	-0.157	-0.072	0.007	0	0.061	-4.461
Log-1	20	-0.083	-0.005	0.066	0	0.036	-4.895
Log-2	20	-0.138	-0.064	0.002	0	0.065	-2.744
Log-3	20	-0.141	-0.073	-0.007	0	0.064	-3.051

Table 3: GOF measures for simulated data from the Clayton copula with $\theta = -0.8$.

Part.Type	K	$Q_{\kappa_\tau}^{(0.025)}$	$\hat{\kappa}_\tau$	$Q_{\kappa_\tau}^{(0.975)}$	κ_τ	Sup.Norm	LPML
Log-0.3	10	-0.262	-0.207	-0.142	-0.667	0.132	29.122
Log-0.5	10	-0.510	-0.478	-0.436	-0.667	0.028	112.249
Log-0.9	10	-0.487	-0.451	-0.409	-0.667	0.119	136.807
Log-0.3	20	-0.359	-0.305	-0.246	-0.667	0.104	49.245
Log-0.5	20	-0.516	-0.476	-0.429	-0.667	0.140	141.271
Log-0.9	20	-0.507	-0.461	-0.423	-0.667	0.122	152.163

Table 4: GOF measures for simulated data from the Clayton copula with $\theta = 1$.

Part.Type	K	$Q_{\kappa_\tau}^{(0.025)}$	$\hat{\kappa}_\tau$	$Q_{\kappa_\tau}^{(0.975)}$	κ_τ	Sup.Norm	LPML
Log-4	10	0.198	0.265	0.316	0.333	0.065	32.983
Log-6	10	0.222	0.293	0.350	0.333	0.045	34.744
Log-8	10	0.174	0.240	0.295	0.333	0.082	31.584
Log-4	20	0.085	0.247	0.323	0.333	0.092	27.594
Log-6	20	0.043	0.141	0.215	0.333	0.144	15.339
Log-8	20	0.015	0.078	0.139	0.333	0.167	5.326

Table 5: GOF measures for simulated data from the AMH copula with $\theta = -0.7$.

Part.Type	K	$Q_{\kappa_\tau}^{(0.025)}$	$\hat{\kappa}_\tau$	$Q_{\kappa_\tau}^{(0.975)}$	κ_τ	Sup.Norm	LPML
Log-4	10	-0.197	-0.133	-0.050	-0.134	0.022	1.733
Log-6	10	-0.200	-0.132	-0.070	-0.134	0.021	2.869
Log-8	10	-0.205	-0.147	-0.079	-0.134	0.022	2.695
Log-4	20	-0.208	-0.139	-0.067	-0.134	0.022	-0.865
Log-6	20	-0.205	-0.138	-0.067	-0.134	0.021	0.779
Log-8	20	-0.207	-0.137	-0.074	-0.134	0.019	1.828

Table 6: GOF measures for simulated data from the AMH copula with $\theta = 0.7$.

Part.Type	K	$Q_{\kappa_\tau}^{(0.025)}$	$\hat{\kappa}_\tau$	$Q_{\kappa_\tau}^{(0.975)}$	κ_τ	Sup.Norm	LPML
Log-1	10	0.127	0.193	0.249	0.195	0.030	6.721
Log-3	10	0.047	0.123	0.205	0.195	0.068	4.221
Log-7	10	0.006	0.088	0.169	0.195	0.071	0.847
Log-1	20	0.139	0.203	0.264	0.195	0.033	4.524
Log-3	20	-0.007	0.088	0.188	0.195	0.094	-2.083
Log-7	20	-0.037	0.018	0.085	0.195	0.130	-5.715

Table 7: GOF measures for simulated data from the Gumbel copula with $\theta = 1.4$.

Part.Type	K	$Q_{\kappa_\tau}^{(0.025)}$	$\hat{\kappa}_\tau$	$Q_{\kappa_\tau}^{(0.975)}$	κ_τ	Sup.Norm	LPML
Log-1	10	0.165	0.239	0.294	0.286	0.083	14.025
Log-3	10	0.129	0.197	0.290	0.286	0.055	18.013
Log-7	10	0.096	0.135	0.153	0.286	0.064	21.383
Log-1	20	0.073	0.160	0.222	0.286	0.064	11.991
Log-3	20	0.034	0.105	0.249	0.286	0.091	16.908
Log-7	20	-0.009	0.063	0.131	0.286	0.112	11.134

Table 8: GOF measures for the real data.

Part.Type	K	$Q_{\kappa_\tau}^{(0.025)}$	$\hat{\kappa}_\tau$	$Q_{\kappa_\tau}^{(0.975)}$	Sample $\tilde{\kappa}_\tau$	LPML
Log-6	10	-0.220	-0.165	-0.093	-0.112	-1.076
Log-8	10	-0.223	-0.165	-0.080	-0.112	-0.787
Log-10	10	-0.213	-0.152	-0.098	-0.112	1.002
Log-6	20	-0.229	-0.167	-0.102	-0.112	-3.377
Log-8	20	-0.232	-0.171	-0.112	-0.112	-2.944
Log-10	20	-0.233	-0.171	-0.101	-0.112	-2.018