Dependent censoring based on parametric copulas

By Kaushik Raj V Nadar

208160499

Survival Time (T)

- **Definition:** Survival time is the duration of time from a defined starting point (like the beginning of a study, diagnosis, or treatment) until the occurrence of a particular event of interest (like death, relapse, failure of a device, etc.).
- **Example:** If you're studying how long patients live after being diagnosed with a certain disease, the survival time would be the number of days, months, or years from the diagnosis until the patient passes away.

Censoring Time (C)

- **Definition:** Censoring time is the point in time at which the observation of a subject ends, without the event of interest having occurred. This can happen for various reasons:
    1. **Study ends:** The study ends before the event occurs.
    2. **Loss of follow-up:** The subject is lost to follow-up (they move away, drop out of the study, etc.).
    3. **Event doesn’t occur:** The subject doesn’t experience the event during the observation period.
- For example, in a long-term study, different patients might drop out at different times, or the study might conclude after a certain number of years, which is not the same for every patient.

Dependence of T and C

- Often, an assumption in survival analysis is that T and C are independent, meaning that the time at which a person might be censored (e.g., leaves the study) doesn’t influence their actual survival time.
- However, this isn’t always true; sometimes T and C can be dependent. For example, if sicker patients are more likely to drop out of a study early, T and C might be related.

Objective

- The authors prove that if the marginal distributions of T and C and the copula function are all modelled parametrically, then under certain conditions the joint model is identifiable.
- In particular, the association parameter of the copula function is identifiable, which represents an important step forward in the use of copulas in survival analysis.

The Model

- Because of random right censoring we observe $Y = \min(T, C)$ and $\boldsymbol{\Delta} = I(T \leq C)$
- marginal distributions F_T and F_C of T and C are continuous, and belong to parametric families:

$F_T \in \{F_{T,\theta T} : \theta_T \in \Theta_T\}$, $F_C \in \{F_{C,\theta C} : \theta_C \in \Theta_C\}$ for certain parameter spaces Θ_T and Θ_C

- Similarly, we denote their densities by f_T and f_C , or by $f_{T,\theta T}$ and $f_{C,\theta C}$ in the parametric families.

- A copula is a bivariate distribution function $\mathsf{C} : [0, 1] \times [0, 1] \to [0, 1]$ with uniform margins.
- From Sklar's Theorem, we know that there is a unique copula C for which

$$F_{T,C}(t,c) = \mathsf{C}\{F_T(t), F_C(c)\} \text{ holds for any } t,\, c \geq 0$$

- For our approach we need to express conditional distribution functions in terms of their associated copulas and so define

$$h_{C|T,\theta}(v \mid u) = \frac{\partial \mathsf{C}_\theta(u,v)}{\partial u}, \qquad h_{T|C,\theta}(u \mid v) = \frac{\partial \mathsf{C}_\theta(u,v)}{\partial v}.$$

- let $F_{Y,(y,\delta)} = \mathrm{pr}(Y \leq y, = \delta)$ and $f_{Y,(y,\delta)} = (\mathrm{d}/\mathrm{d}y) F_{Y,\boldsymbol{\Delta}}(y,\delta)$ for $\delta = 0, 1$.
- Finally, let $F_Y(y) = \sum_{\delta=0}^{1} F_{Y,\boldsymbol{\Delta}}(y, \delta)$ and $f_Y(y) = \sum_{\delta=0}^{1} f_{Y,\boldsymbol{\Delta}}(y, \delta)$ be the distribution and density of the observable random variable Y, respectively.

- $F_{T|C}(t \mid c) = h_{T|C}\{F_T(t) \mid F_C(c)\}$,
- $F_{C|T}(c \mid t) = h_{C|T}\{F_C(c) \mid F_T(t)\}$,
- $F_Y(y) = F_T(y) + F_C(y) - C\{F_T(y), F_C(y)\}$,
- $f_{Y,\boldsymbol{\Delta}}(y,1) = f_T(y)[1 - h_{C|T}\{F_C(y) \mid F_T(y)\}]$,
- $f_{Y,\boldsymbol{\Delta}}(y, 0) = f_C(y)[1 - h_{T|C}\{F_T(y) \mid F_C(y)\}]$.
- Parametric form: (copula parameters θ and marginal parameters θ_T and θ_C)

For $\alpha = (\theta, \theta_T, \theta_C)^{\mathrm{T}}$, we write $f_{Y,\boldsymbol{\Delta},\alpha}(y, 1) = f_{T,\theta T}(y)[1 - h_{C|T,\theta}\{F_{C,\theta C}(y) \mid F_{T,\theta T}(y)\}]$.

Identifiability

Identifiability means that

$((\theta, \theta_T, \theta_C) \in \Theta \times \Theta_T \times \Theta_C)$ $uniquely\ determine\ the\ density\ of\ the\ observable\ variables$ $((Y, \Delta))$.

$Specifically, if$ $(f_{\{Y,\Delta;\,\alpha_1\}} \equiv f_{\{Y,\Delta;\,\alpha_2\}})$, $then$ $(\alpha_1 = \alpha_2)$, $where$ $(\alpha_j = \left(\theta_j, \theta_{\{Tj\}}, \theta_{\{Cj\}}\right)^{T})$ for $(j = 1, 2)$

Theorems in following slides will help us establish identifiability in different models.

Identifiability

THEOREM 1. *Suppose that the following two conditions hold.*

(i) *For $\theta_{T1}, \theta_{T2} \in \Theta_T$ and $\theta_{C1}, \theta_{C2} \in \Theta_C$ we have the four equivalences*

$$\lim_{t\to 0} \frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)} = 1 \iff \theta_{T1} = \theta_{T2}, \qquad \lim_{t\to\infty} \frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)} = 1 \iff \theta_{T1} = \theta_{T2},$$

$$\lim_{t\to 0} \frac{f_{C,\theta_{C1}}(t)}{f_{C,\theta_{C2}}(t)} = 1 \iff \theta_{C1} = \theta_{C2}, \qquad \lim_{t\to\infty} \frac{f_{C,\theta_{C1}}(t)}{f_{C,\theta_{C2}}(t)} = 1 \iff \theta_{C1} = \theta_{C2}.$$

(ii) *The parameter space $\Theta \times \Theta_T \times \Theta_C$ is such that*

$$\lim_{t\to 0} h_{T|C,\theta}(u_t \mid v_t) = 0 \quad \forall\, (\theta, \theta_T, \theta_C) \in \Theta \times \Theta_T \times \Theta_C$$

or

$$\lim_{t\to\infty} h_{T|C,\theta}(u_t \mid v_t) = 0 \quad \forall\, (\theta, \theta_T, \theta_C) \in \Theta \times \Theta_T \times \Theta_C,$$

and similarly for $h_{C|T,\theta}(v_t \mid u_t)$, where $u_t = F_{T,\theta_T}(t)$ and $v_t = F_{C,\theta_C}(t)$.

Then the model specified in (1)–(3) is identified.

Identifiability

Theorem 2: Condition (i) of Theorem 1 is satisfied for the families of lognormal, log-Student-t,Weibull and log-logistic densities.

LEMMA 1. *Suppose the generator ψ is differentiable on* $(0,1)$. *If* $\lim_{v\to 1}\psi'(v)\in(-\infty,0)$, *then* $\lim_{t\to\infty} h_{T|C,\theta}\{F_{T,\theta_T}(t)\mid F_{C,\theta_C}(t)\}=1$.

Archimedean Copula

$$C(u,v)=\psi^{[-1]}\{\psi(u)+\psi(v)\} \tag{4}$$

where ψ is a generator, that is, $\psi:[0,1]\to[0,\infty)$ is a continuous, strictly decreasing and convex function such that $\psi(1)=0$. Here, $\psi^{[-1]}$ is the pseudo-inverse of ψ, i.e., $\psi^{[-1]}(t)=\psi^{-1}(t)$ if $0\le t\le 1$ and $\psi^{[-1]}(t)=0$ if $t\ge\psi(0)$.

Gaussian Copula

$$C_\theta(u,v)=\Phi_\theta\{\Phi^{-1}(u),\Phi^{-1}(v)\}$$

where,

Φ: CDF of Standard Normal distribution

Φ_θ: CDF of Bivariate Standard Normal distribution with correlation θ

Identifiability

THEOREM 3 (FRANK, GUMBEL AND GAUSSIAN COPULAS). *Condition* (ii) *of Theorem* 1 *is satisfied by the following:*

(i) *the Frank copula, independently of the marginal distributions and the parameter space;*

(ii) *the Gumbel copula if* $\lim_{t\to 0} \log F_{T,\theta_T}(t)/\log F_{C,\theta_C}(t) \in (0,\infty)$ *for all* $(\theta_T,\theta_C) \in \Theta_T \times \Theta_C$*;*

(iii) *the Gaussian copula if*

$$\lim_{t\to 0} A_{\theta,F_{T,\theta_T},F_{C,\theta_C}}(t) = -\infty \quad \forall\,(\theta,\theta_T,\theta_C) \in \Theta \times \Theta_T \times \Theta_C$$

or

$$\lim_{t\to \infty} A_{\theta,F_{T,\theta_T},F_{C,\theta_C}}(t) = -\infty \quad \forall\,(\theta,\theta_T,\theta_C) \in \Theta \times \Theta_T \times \Theta_C,$$

and similarly for $A_{\theta,F_{C,\theta_C},F_{T,\theta_T}}$*, where* $A_{\theta,F_1,F_2}(t) = \Phi^{-1}\{F_1(t)\} - \theta\Phi^{-1}\{F_2(t)\}$.

Theorem 4

Assume that condition (i) of Theorem 1 is met, and that the parameter spaces $\Theta_T \times \Theta_C$ satisfy $\lim_{t\to 0} \frac{F_{T,\theta_T}(t)}{F_{C,\theta_C}(t)}$ is either 0 or $+\infty$ for all $\theta_T \in \Theta_T$ and $\theta_C \in \Theta_C$. Additionally, suppose the copula C_θ is a Clayton copula with $\theta > 0$. Then, the model defined in equations (1)–(3) is identifiable.

Examples of Identifiable Models

I have proved for following combinations of marginals and copulas:

- From Theorem 3, it is clear that we can create an identifiable model with the Frank Copula and any of the marginal distribution satisfying condition (i) of Theorem 1.
- Lognormal + Gumbel
- Lognormal + Clayton
- Weibull + Clayton

Estimation

- Assume that we have an independent and identically distributed sample D = $\{(y_i, \delta_i), i = 1, ..., n\}$ available. Then the joint loglikelihood for the parameter vector $\alpha=(\theta,\theta_T,\theta_C)^T$ is

$$l(\alpha; D) = \sum_{i=1}^{n} \log\{fY_{,\Delta,\alpha}(yi, \delta_i)\}$$

$$= \sum_{\delta_i=1} \log(fT_{,\theta T}(yi)[1 - h_{C|T,\theta}\{F_{C,\theta C}(y_i) \mid F_{T,\theta T}(y_i)\}])$$

$$+ \sum_{\delta_i=0} \log(fC_{,\theta c}(yi)[1 - hT_{|C,\theta}\{F_{T,\theta T}(y_i) \mid F_{C,\theta C}(y_i)\}])$$

- We follow a maximum likelihood approach by maximizing the log likelihood specified, i.e., we define parameter estimators

$$\hat{\alpha} = \left(\hat{\theta}, \hat{\theta}_T, \hat{\theta}_C\right)^T = argmax_{\alpha\in A}\, l(\alpha; D)$$

where $A = \Theta \times \Theta_T \times \Theta_C$

Asymptotic Normality

Let $\alpha=(\theta,\theta_T,\theta_C)^{\mathrm{T}}$ be the parameter vector that minimizes the Kullback–Leibler information criterion $E\{\log(f_{Y,\boldsymbol{\Delta}}(Y,\boldsymbol{\Delta})) - \log(f_{Y,\boldsymbol{\Delta},\alpha}(Y,\boldsymbol{\Delta}))\}$, and let $d = \dim(\Theta) + \dim(\Theta_T) + \dim(\Theta_C)$

(i) *Under the regularity conditions* (A1)–(A3) *in White (1982),*

$$(\hat{\theta},\hat{\theta}_T,\hat{\theta}_C)^T \rightarrow (\theta^*,\theta^*_T,\theta^*_C) \text{ in probability as } n \rightarrow \infty.$$

(ii) *Under the regularity conditions* (A1)–(A6) *in White (1982),*

$$n^{1/2}\,(\hat{\theta},\hat{\theta}_T,\hat{\theta}_C)^T - (\theta^*,\theta^*_T,\theta^*_C) \rightarrow N(0,V) \text{ in distribution as } n \rightarrow \infty,$$

where $V = A(\alpha^*)^{-1}B(\alpha^*)A(\alpha^*)^{-1}$, *with*

$$A(\alpha) = \left[E\left\{\frac{\partial^2}{\partial\alpha_j\partial\alpha_k}\log f_{Y,\Delta,\alpha}(Y,\Delta)\right\}\right]_{j,k=1}^{d}$$

$$B(\alpha) = \left[E\left\{\frac{\partial}{\partial\alpha_j}\log f_{Y,\Delta,\alpha}(Y,\Delta)\frac{\partial}{\partial\alpha_k}\log f_{Y,\Delta,\alpha}(Y,\Delta)\right\}\right]_{j,k=1}^{d}$$

- If the model is correctly specified, V equals $A(\alpha)^{-1}$, the inverse Fisher matrix.

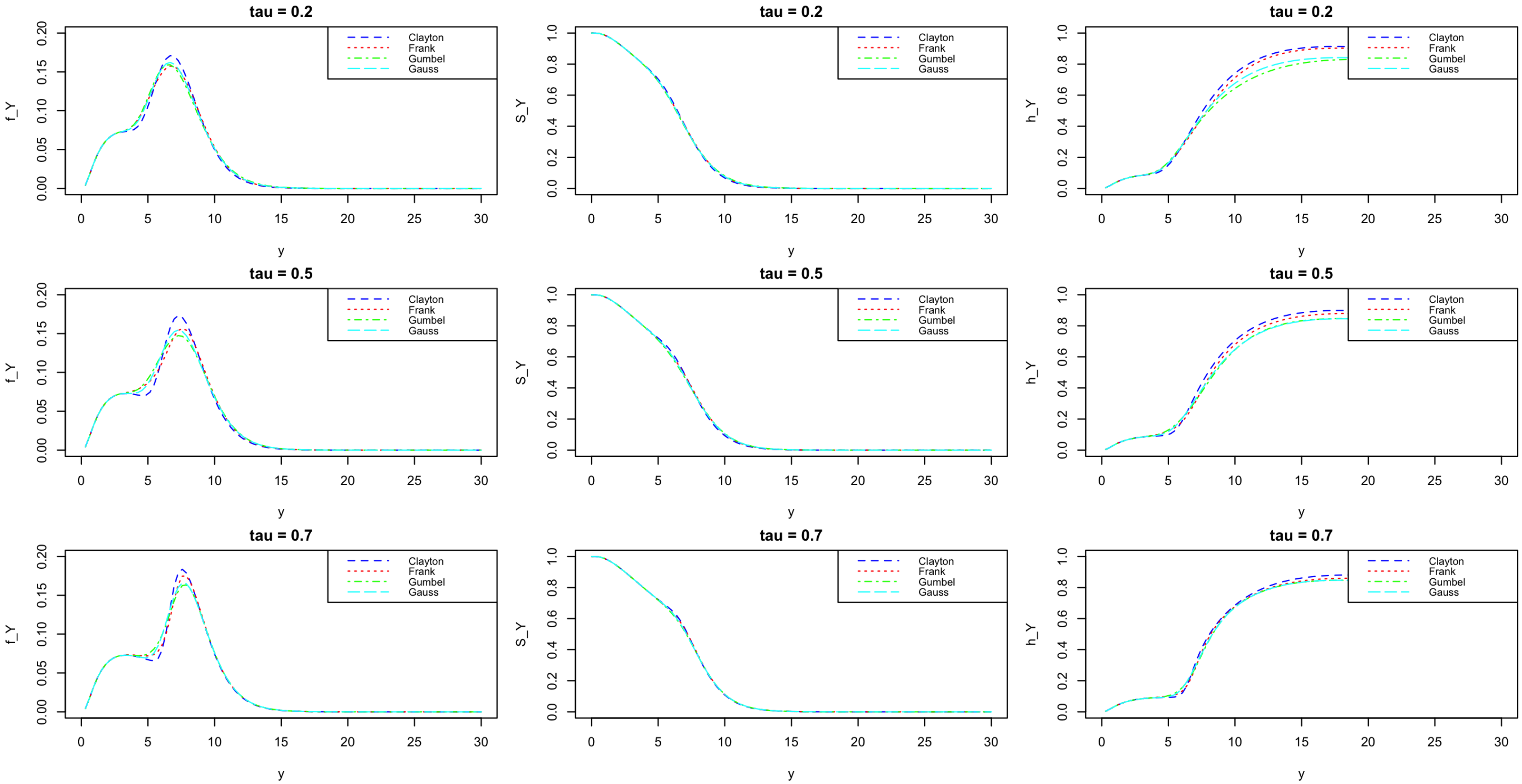
Simulation

- Log Normal Margins
- Took approx. 12 Hrs to complete 100 replications of Scenario 1
- For sample size n=200 and n=500
- Calculated Standard Error, Aymptotic Standard error and RMSE
- Maximised Log Likelihood using L-BFGS Optimizer

Scenario	μ_T	σ_T	μ_C	σ_C	τ	θ_{Frank}	θ_{Clayton}	θ_{Gumbel}	θ_{Gauss}
1	2.2	1.0	2.0	0.25	0.2	1.86	0.50	1.25	0.31
					0.5	5.74	2.00	2.00	0.71
					0.7	11.74	4.67	3.33	0.89
2	2.5	1.0	2.0	0.50	0.2	1.86	0.50	1.25	0.31
					0.5	5.74	2.00	2.00	0.71
					0.7	11.74	4.67	3.33	0.89

Parameter specifications for the simulation scenarios with lognormal margins

Theoretical Densities for Scenario 1

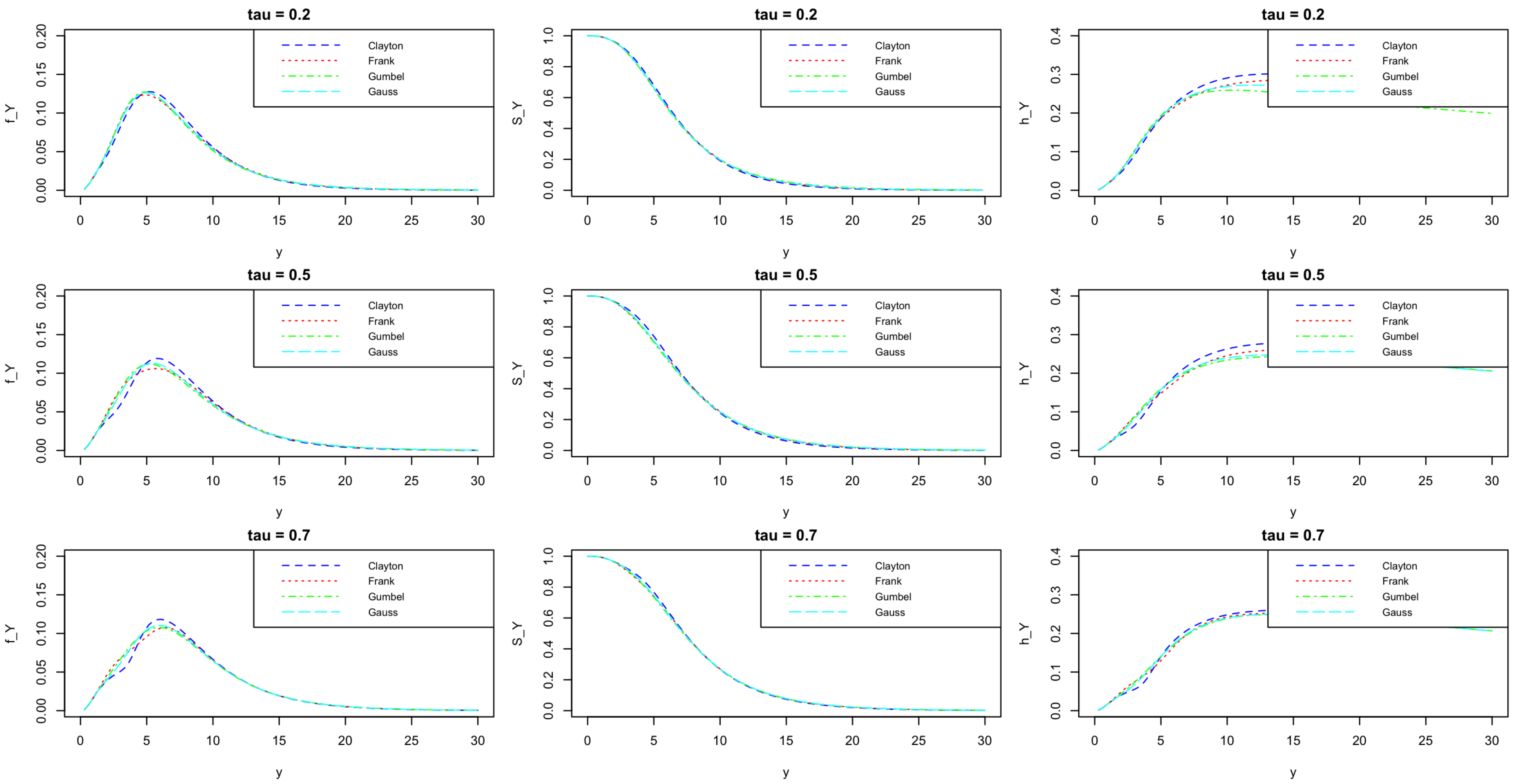


Theoretical density

Survival Function

Hazard Function

Theoretical Densities for Scenario 2



Theoretical density

Survival Function

Hazard Function

Simulation Results for Scenario 1

	τ	μ_T	σ_T	μ_C	σ_C	Estimated τ
n=200						
aver.est	0.7	2.22	1.03	2.00	0.25	0.70
sd.aver.est	0.7	0.11	0.10	0.03	0.02	0.04
aver.asderr	0.7	0.10	0.10	0.03	0.02	0.06
RMSE	0.7	0.11	0.11	0.03	0.02	0.04
aver.est	0.5	2.22	1.02	2.01	0.24	0.44
sd.aver.est	0.5	0.13	0.12	0.05	0.02	0.18
aver.asderr	0.5	0.11	0.10	0.04	0.02	0.14
RMSE	0.5	0.13	0.12	0.05	0.02	0.18
aver.est	0.2	2.26	1.00	2.00	0.25	0.22
sd.aver.est	0.2	0.14	0.09	0.02	0.02	0.12
aver.asderr	0.2	0.12	0.11	0.04	0.02	0.18
RMSE	0.2	0.15	0.08	0.02	0.02	0.12
n=500						
aver.est	0.7	2.19	0.98	2.00	0.25	0.71
sd.aver.est	0.7	0.06	0.06	0.02	0.01	0.05
aver.asderr	0.7	0.06	0.06	0.02	0.01	0.04
RMSE	0.7	0.06	0.07	0.01	0.01	0.04
aver.est	0.5	2.19	0.99	2.00	0.25	0.47
sd.aver.est	0.5	0.08	0.07	0.03	0.02	0.08
aver.asderr	0.5	0.07	0.06	0.02	0.01	0.08
RMSE	0.5	0.08	0.07	0.03	0.02	0.08
aver.est	0.2	2.18	0.96	1.99	0.25	0.25
sd.aver.est	0.2	0.05	0.04	0.02	0.01	0.09
aver.asderr	0.2	0.07	0.06	0.03	0.01	0.11
RMSE	0.2	0.06	0.06	0.03	0.01	0.10

Frank Copula

	τ	μ_T	σ_T	μ_C	σ_C	Estimated τ
n=200						
aver.est	0.7	2.22	1.03	2.00	0.25	0.71
sd.aver.est	0.7	0.08	0.09	0.02	0.02	0.04
aver.asderr	0.7	0.10	0.09	0.03	0.02	0.07
RMSE	0.7	0.08	0.10	0.02	0.02	0.04
aver.est	0.5	2.20	1.02	2.01	0.24	0.45
sd.aver.est	0.5	0.10	0.08	0.05	0.03	0.18
aver.asderr	0.5	0.10	0.09	0.04	0.03	0.17
RMSE	0.5	0.10	0.08	0.05	0.03	0.18
aver.est	0.2	2.16	0.96	1.99	0.26	0.25
sd.aver.est	0.2	0.10	0.07	0.05	0.03	0.19
aver.asderr	0.2	0.11	0.09	0.07	0.03	0.25
RMSE	0.2	0.10	0.08	0.05	0.03	0.20
n=500						
aver.est	0.7	2.22	1.00	2.00	0.25	0.69
sd.aver.est	0.7	0.07	0.07	0.02	0.01	0.04
aver.asderr	0.7	0.06	0.06	0.02	0.01	0.05
RMSE	0.7	0.07	0.07	0.02	0.01	0.04
aver.est	0.5	2.19	1.00	2.00	0.25	0.52
sd.aver.est	0.5	0.07	0.07	0.02	0.02	0.07
aver.asderr	0.5	0.06	0.06	0.02	0.02	0.09
RMSE	0.5	0.07	0.07	0.02	0.02	0.07
aver.est	0.2	2.18	0.98	2.01	0.25	0.19
sd.aver.est	0.2	0.06	0.05	0.04	0.02	0.13
aver.asderr	0.2	0.07	0.06	0.05	0.02	0.19
RMSE	0.2	0.06	0.06	0.04	0.02	0.12

Clayton Copula

	τ	μ_T	σ_T	μ_C	σ_C	Estimated τ
n=500						
aver.est	0.7	2.19	1.00	2.00	0.25	0.71
sd.aver.est	0.7	0.05	0.07	0.02	0.01	0.04
aver.asderr	0.7	0.06	0.06	0.02	0.01	0.04
RMSE	0.7	0.05	0.07	0.02	0.01	0.04
aver.est	0.5	2.20	1.00	2.00	0.25	0.48
sd.aver.est	0.5	0.07	0.05	0.02	0.01	0.08
aver.asderr	0.5	0.07	0.06	0.02	0.02	0.09
RMSE	0.5	0.07	0.05	0.02	0.01	0.08
aver.est	0.2	2.18	0.99	1.99	0.25	0.23
sd.aver.est	0.2	0.05	0.05	0.03	0.02	0.13
aver.asderr	0.2	0.07	0.06	0.03	0.01	0.13
RMSE	0.2	0.05	0.05	0.03	0.02	0.13
n=200						
aver.est	0.7	2.22	1.00	2.00	0.24	0.68
sd.aver.est	0.7	0.09	0.08	0.02	0.02	0.08
aver.asderr	0.7	0.11	0.09	0.03	0.02	0.08
RMSE	0.7	0.09	0.08	0.02	0.02	0.08
aver.est	0.5	2.23	1.03	2.00	0.25	0.46
sd.aver.est	0.5	0.11	0.09	0.03	0.02	0.14
aver.asderr	0.5	0.11	0.10	0.04	0.02	0.15
RMSE	0.5	0.11	0.09	0.03	0.02	0.14
aver.est	0.2	2.19	1.00	1.99	0.25	0.21
sd.aver.est	0.2	0.10	0.11	0.04	0.02	0.16
aver.asderr	0.2	0.12	0.10	0.05	0.02	0.20
RMSE	0.2	0.09	0.11	0.04	0.02	0.16

Gaussian Copula

Conclusion

- Results consistent with the original paper.
- Identifiability of a copula model established under dependent censoring, without presuming knowledge of the copula’s association parameter.
- Demonstrated identifiability for additional marginal distributions, specifically Exponential, Gamma, and truncated Normal distributions
- Simulations performed to verify the results and assess model performance.