A Nonparametric Conditional Copula Model For Successive Duration Times

with application to insurance subscription

Yohann Le Faou & Olivier Lopez

Sorbonne Université, CNRS, LPSM, Paris, France

yohann.lefaou@courriel.upmc.fr

github.com/YohannLeFaou

Keywords : Successive times, Censoring, Copula, Insurance

Introduction

We consider the estimation of a **conditional copula** function $\mathfrak{C}$ of a couple of duration variables T and U, in a framework where these **times are observed successively** and suffer from **right censoring**.

Applications: Biostatistics (T = Infection time, U = Recovery time), **Insurance** (T = Effective time of a contract, U = Termination time of a contract).

Goal: Study the dependence structure between T and U, in presence of covariates $X \in \mathbb{R}^d$ - e.g. age of the policyholder, sex, level of insurance - that may have impact on the joint distribution.

Theorem: Sklar's theorem

Let $F(t,u) = \mathbb{P}(T \le t, U \le u)$. Then:

$$F(t,u) = \mathfrak{C}(F_T(t), F_U(u)).$$

Obstacles: Censoring variable C. Dependence studied conditionally on X.

Contributions: Mathematical justification of our method. Application on a real dataset of information on a portfolio of health insurance contracts.

Censored Observations

- We consider i.i.d. realizations $(T_i, U_i, X_i, C_i)_{1\le i\le n}$ of a random vector (T, U, X, C).
- **Censoring of the data**: the variables T and U are not directly observed. Instead of (T_i, U_i), one observes

$$\begin{cases} Y_i &= \min(T_i, C_i), \\ Z_i &= \min(U_i, C_i - T_i), \\ \eta_i &= \mathbf{1}_{T_i \le C_i}, \\ \gamma_i &= \mathbf{1}_{U_i + T_i \le C_i}. \end{cases}$$

Conditional copula estimation

Let $F(t,u|x) = \mathbb{P}(T \le t, U \le u | X = x)$. By Sklar Theorem, $F(t,u|x) = \mathfrak{C}^{(x)}(F_T(t|x), F_U(u|x))$, where $\mathfrak{C}^{(x)}$ denotes the copula of the conditional distribution of (T,U) conditionally on $X = x$.

Assumption 1: Fundamental Assumptions

(a) Assume that C is independent from (T, U, X).
(b) Let $\mathcal{C} = \{\mathfrak{C}_\theta : \theta \in \Theta\}$, with Θ a compact subset of $\mathbb{R}^k$, denote a parametric family of copula functions. Assume that, for all $x \in \mathcal{X}$, there exists $\theta(x) \in \Theta$ such that

$$\mathfrak{C}^{(x)} = \mathfrak{C}_{\theta(x)}.$$

Let $M(x,\theta) = E[\log c_\theta(F_T(T|X), F_U(U|X))|X = x]$, where $c_\theta(a,b) = \partial^2_{a,b}\mathfrak{C}_\theta(a,b)$ denotes the copula density associated with copula function $\mathfrak{C}_\theta$. We have, by definition of $\theta(x)$,

$$\theta(x) = \arg\max_{\theta\in\Theta} M(x,\theta).$$

Consider a function ϕ such that $E[|\phi(T,U,X)|] < \infty$, and $\phi(t,u,x) = 0$ for $t + u \ge \tau_{U+T}(x)$. Under Assumption 1.a, elementary computations show that

$$E\left[\frac{\delta\phi(Y,Z,X)}{S_C(Y+Z)}\middle| X\right] = E\left[\phi(T,U,X)|X\right], \tag{1}$$

where $S_C(t) = \mathbb{P}(C > t)$, and $\delta = \eta\gamma$.

Let

$$M_{n,h}(x,\theta) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)\frac{\delta_i \log c_\theta(\hat{A}_i, \hat{B}_i)}{\hat{S}_C(Y_i + Z_i)}\omega_{i,n} \tag{2}$$

$$= \frac{1}{nh^d}\sum_{i=1}^n W_{i,n} K\left(\frac{X_i - x}{h}\right)\log c_\theta(\hat{A}_i, \hat{B}_i)\omega_{i,n}, \tag{3}$$

where K is a kernel function, h is a bandwidth parameter, $\hat{A}_i = \hat{F}_T(Y_i|X_i)$ and $\hat{B}_i = \hat{F}_U(Z_i|X_i)$ are pseudo-observations, $\hat{S}_C$ is an estimator of S_C, and $w_{i,n}$ is a trimming function defined as $\omega_{i,n} = \mathbf{1}_{\min(\hat{A}_i, \hat{B}_i, 1-\hat{A}_i, 1-\hat{B}_i)\ge\nu_n}$ for a sequence ν_n tending to zero.

We define our final estimator of $\theta(x)$ as

$$\hat{\theta}_h(x) = \arg\min_{\theta\in\Theta} M_{n,h}(x,\theta), \tag{4}$$

Assumption 2: Technical Assumptions

(a) Regularity assumptions (C^2) on the model when x varies.
(b) Assumptions that ensure that the Hessian matrix of $c_\theta(x,\theta)$ taken at the point $\theta(x)$ is positive-definite.
(c) Integrability assumptions (L^2) required to obtain results on convergence speed.
(d) Speed of convergence of pseudo-observations:

$$\sup_{1\le i\le n} |\hat{A}_i - A_i| + |\hat{B}_i - B_i| = O_P(\varepsilon_n),$$

with $\varepsilon_n = o(\nu_n)$.

Main Results

Theorem 1: Bias term

Let

$$\theta^*_h(x) = \arg\max_{\theta\in\Theta} \frac{1}{h^d} E\left[K\left(\frac{X_i - x}{h}\right)\log c_\theta(F_T(T_i|X_i), F_U(U_i|X_i))\right].$$

Then:

$$\sup_{x\in\mathcal{X}} \|\theta^*_h(x) - \theta(x)\| = O(h^2).$$

Theorem 2: Stochastic term

$$\sup_{x\in\mathcal{X}} \|\hat{\theta}_h(x) - \theta^*_h(x)\| = O_P(\nu_n + [\log n]^{1/2} n^{-1/2} h^{-d/2}).$$

Application of the method to insurance data

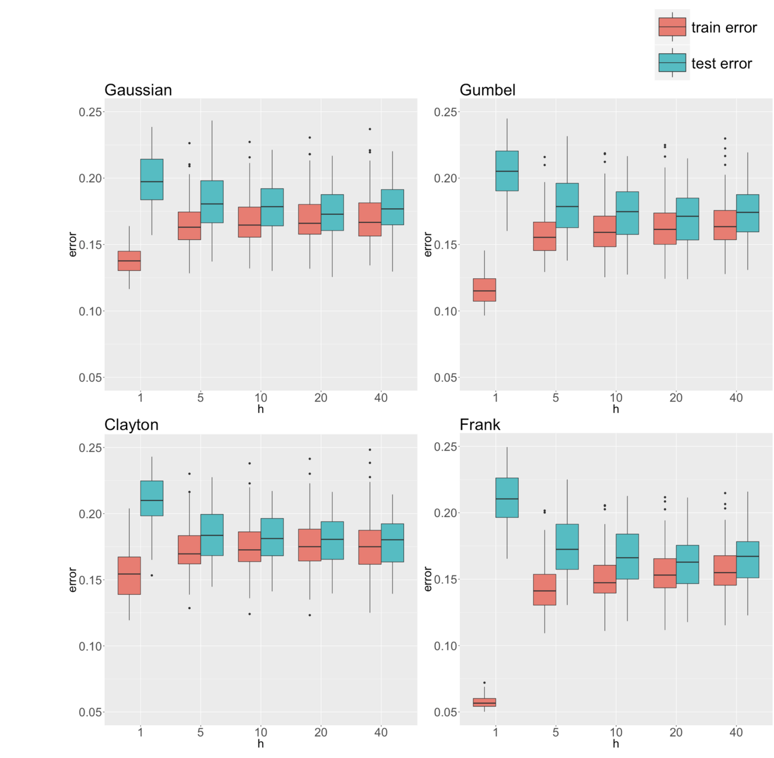
- T corresponds to the effective time of a contract (i.e. the duration between the date of subscription and the date of effect of the contract),
- U is the termination time of the contract (i.e. the duration between the date of effect and the date of termination),
- C is the age of a contract (i.e. the duration between the date of subscription of the contract and the date of the end of observation),
- X is the age of the contract holder at the subscription.

Four families of parametric copulas are tested to model the dependence between T and U: Gaussian, Clayton, Gumbel and Frank copulas.

Results

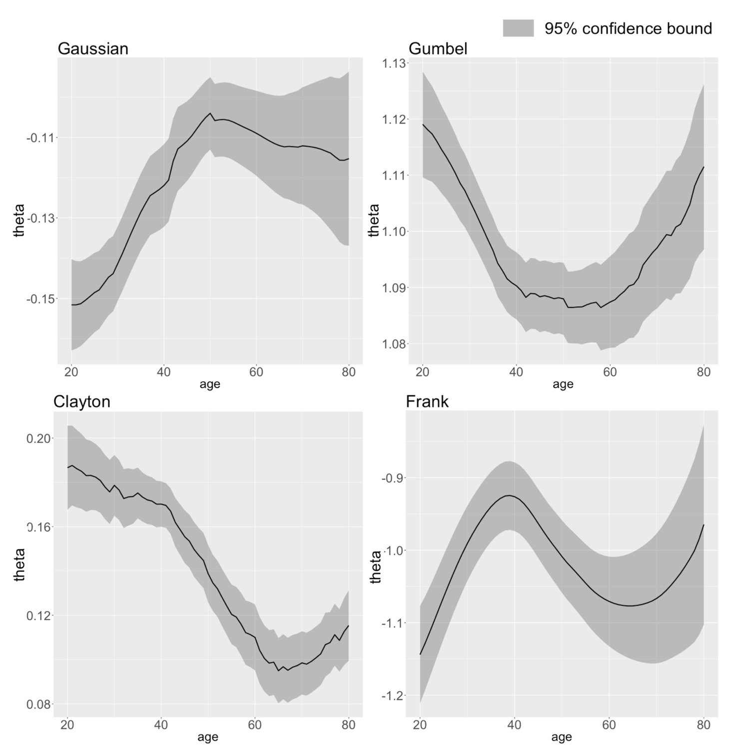
(a) For each copula family and each bandwidth value h, box plots of the train and test square root errors $\sqrt{\epsilon_{h,tr}}$ and $\sqrt{\epsilon_{h,te}}$ ($n = 10000$, 100 repetitions). $\sqrt{\epsilon_{h,t.}}$ is a distance calculated from Kendall τ, using the one to one relations between the parameter θ and the Kendall τ for the different copula families.

(b) For each copula family, mean value of the conditional copula parameter as a function of the age x ($h = 20$, $n = 10000$, 100 repetitions).

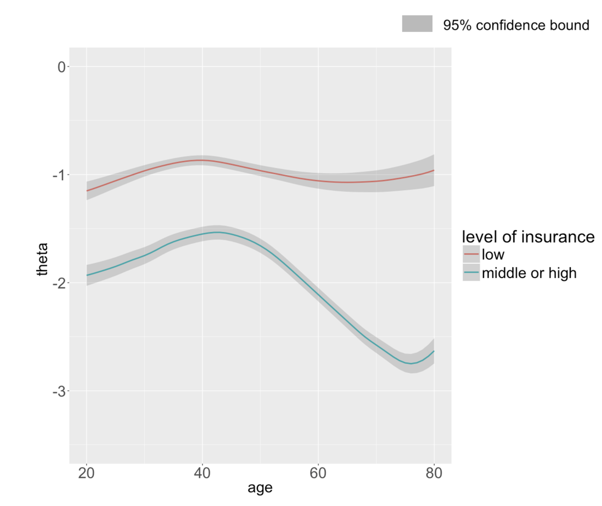


(a) Copulas train and test errors

(c) Impact of the variable *level of insurance* on the conditional dependence between T and U, given the age of the prospect (Frank copula, $h = 20$, 100 repetitions).



(b) Copulas conditional parameters



(c) Impact of the level of insurance

Conclusion

- We proposed a methodology to estimate a conditional copula function under random censoring, when the two variables linked through the copula are right censored successive times.
- From a numerical point of view, the procedure is simple, since it relies on a weighted log-likelihood approach.
- In our paper, we provide a mathematical justification of the method. In particular, we provide conditions on the censoring which allow to understand the behavior of the method even in the tail of the distribution.

References

[1] Y. Le Faou and O. Lopez. A nonparametric conditional copula model for successive duration times, with application to insurance subscription.