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

Keywords: Successive times, Censoring, Copula, Insurance

Introduction

We consider the estimation of a conditional copula function $C$ 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 $X \in \mathbb{R}^p$, e.g. of the policyholder, sex, level of insurance - that may have an impact on the joint distribution.

Theorem: Sklar’s theorem

Let $F(t, u) = P(T \leq t, U \leq u)$, Then: $F(t, u) = C(F_U(t), F_U(u))$.

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 $(T_i, U_i, X_i, C_i)_{i \in \mathbb{N}}$ 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 $\left( Y_i = \min(T_i, C_i), Z_i - \min(U_i, C_i - T_i) \right)$, $\theta_i = \log \frac{C(Y_i, Z_i)}{C(Y_i, C_i - T_i)}, \gamma_i = \log \frac{C(Y_i, Z_i)}{C(Y_i, Z_i)}$ such that $C(Y_i, Z_i)$ is the copula density associated with copula function $C_i$. We have, by definition of $\theta_i$, $\gamma_i = \max \{ \log \frac{M(x, \theta)}{M(x, \gamma)} \} \sup \theta \in \Theta$.

Conditional copula estimation

Let $F(t, u) = P(T \leq t, U \leq u | X = x)$. By Sklar Theorem, $F(t, u) = C\left( F_U(t), F_U(u) \right)$, where $C^{(1)}$ denotes the copula of the conditional distribution of $(T, U)$ conditionally on $X = x$.

Assumption 1: Fundamental Assumptions

(a) Assume that $C$ is independent of $(T, U, X)$.
(b) Let $C \subset (\Theta, \emptyset \in \Theta)$, with $\Theta$ a compact subset of $\mathbb{R}^p$, denote a parametric family of copula functions. Assume that, for all $x \in X$, there exists $\theta(x) \in \Theta$ such that $C^{(1)}(\cdot, x) = \arg \max \{ C(y, x) \}$.

Conclusion

Main Results

Theorem 1: Bias term

Let $\theta_i(x) = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{k=1}^n K \left( \frac{X_k - x}{h} \right) \log \frac{C(Y_k, Z_k)}{C(Y_k, \nu_k)}$.

Then: $\sup_{x \in \mathbb{R}^p} \left| \theta_i(x) - \theta(x) \right| = O_p(h^2)$.

Theorem 2: Stochastic term

$\sup_{x \in \mathbb{R}^p} \left| \theta_i(x) - \theta(x) \right| = O_P((\nu_i \log n)^{1/2} h^{-1/2} \sqrt{D})$.

Application of the method to the 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).
- $T$ 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, $\sqrt{\frac{1}{n} \sum_{k=1}^n K \left( \frac{X_k - x}{h} \right) \log \frac{C(Y_k, Z_k)}{C(Y_k, \nu_k)}}$ and $\sqrt{\frac{1}{n} \sum_{k=1}^n K \left( \frac{X_k - x}{h} \right) \log \frac{C(Y_k, Z_k)}{C(Y_k, \nu_k)}}$ (n = 10000, 100 repetitions).

(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).

References