ARTICLE TEMPLATE

A semiparametric copula-based estimation of the regression function for right-censored data

Taoufik Bouezmarni[a], Yassir Rabhi[b] and Charles Fontaine[c]

[a]Department of Mathematics, University of Sherbrooke, Sherbrooke Canada. Centre interuniversitaire de recherche en économie quantitative (CIREQ). Email: Taoufik.Bouezmarni@Usherbrooke.ca; [b] State University of New York, Cortland, USA. Email: Yassir.Rabhi@cortland.edu; [c] Institut Universitaire de Recherche Clinique, Université de Montpellier 641, Ave du Doyen Gaston-Giraud 34090 Montpellier France. Email: cfontaine@parisnanterre.fr

ARTICLE HISTORY
Compiled September 20, 2019

ABSTRACT
This paper addresses the semiparametric estimation of the regression function in a situation where the response variable is right-censored and the covariate(s) is completely observed. We present a new copula-based method to estimate the regression function. The key concept presented in this manuscript is to write the regression function in terms of the copula density and marginal distributions. We suppose a parametric model for the copula density with unknown parameter(s) and we estimate the marginal distributions of the response and the covariate by the Kaplan-Meier estimator and the empirical distribution, respectively. We establish the asymptotic properties of our estimator and extend it to the multivariate case. The proposed method is then applied to analyze a data-set on lifetime with lung-cancer.

KEYWORDS
Semiparametric copula-based estimation; Regression function; Censored data; Parametric copula models; Kaplan-Meier estimator.

1. Introduction

Incident cohort design is usually adopted to study the time between an initiating event, say disease onset, to a terminating event, say death. The failure times of individuals recruited in such studies may however be subject to right-censoring due to voluntary retraction, emigration or end of study. Right-censoring on the response variable (failure time) arises when the followed subjects in a study cannot provide us information on the outcome of the study despite of our knowledge of the predictors for these subjects (e.g. age at disease onset).

The relation between these predictors and the response variable is one of the most

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Canssi and Statlab-CRM (Canada) and Le Fonds de recherche du Québec (Nature et technologies). The authors thank Prof Alexander Meister (the editor), the associate editor, and two referees for their helpful and constructive comments.

important problem in statistics. In particular, the regression model

$$Y = m(X) + \varepsilon,$$

where $m(x) = E[Y|X = x]$ is an unknown smooth function and ε is a random error ($E[\varepsilon|X = x] = 0$ and $E[\varepsilon^2|X = x] < \infty$). The estimation of the regression function m have been extensively studied in literature under different data settings. At first, [18] and [11] proposed estimators based on the linear regression model with some methodological weaknesses since they both require a particular censoring pattern. Furthermore, based on the general linear model, the method proposed by [1] is an iterative sequence of estimators and has the drawback of not necessarily converging. In the work of [5,13,24,26–28], data transformations for censored data were presented. The problem with the linear regression models is that they are too restrictive. Nonparametric methods, for example Nadaraya-Watson or local linear kernel (see [6]), provide a flexible method, since they do not assume a specific model for the regression function. However, the nonparametric approaches suffer from the curse of dimensionality. More recently, several copula-based methods were proposed to estimate m for complete data. For instance, see the parametric approach of [25], [12] and [3], and the semiparametric method of [20].

Consider a response variable Y, with distribution F_0 and density f_0, and a vector of d predictors $\boldsymbol{X} = (X_1, ..., X_d)$ with joint distribution $F_{\boldsymbol{X}}$ and joint density $f_{\boldsymbol{X}}$. Let $F_1, \ldots, F_d$ be the respective marginal distributions of $X_1, \ldots, X_d$, $\boldsymbol{x} = (x_1, \ldots, x_d)$ and denote $\boldsymbol{F}(\boldsymbol{x}) = \big(F_1(x_1), \ldots, F_d(x_d)\big)$. From the copula theory, it is known that the multivariate distribution $F_{Y,X_1,...,X_d}$ is given by

$$F_{Y,X_1,...,X_d}(y, x_1, \ldots, x_d) = \mathbb{C}\left(F_0(y), F_1(x_1), \ldots, F_d(x_d)\right),$$

where $\mathbb{C}$ is the copula distribution of $(Y, X_1, ..., X_d)$ with uniform margins on $[0, 1]$. Using this decomposition, the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$ can be written in terms of copula densities as

$$f_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = f_0(y)\, \frac{c\left(F_0(y), \boldsymbol{F}(\boldsymbol{x})\right)}{c_X\left(\boldsymbol{F}(\boldsymbol{x})\right)}, \tag{1}$$

where c and c_X are the respective copula densities of $(Y, X_1, ..., X_d)$ and $(X_1, ..., X_d)$. Using a copula-based parametric approach, [25], [12] and [3] investigated the relationship (1) for various copula families (e.g. Gaussian, Student, Farlie-Gumbel-Morgenstern (FGM), Iterated FGM, Archimedean) in order to estimate the regression function $m(\boldsymbol{x})$ for single and multiple covariates. For instance, if the copula density of (Y, X_1) belongs to the FGM family with parameter θ, i.e. $c(u_0, u_1) = 1 + \theta(1 - 2u_0)(1 - 2u_1)$, then

$$m(x_1) = E[Y] + \theta\left(2\, F_1(x_1) - 1\right) \int F_0(y)\left(1 - F_0(y)\right)\, \mathrm{d}y.$$

For the multiple covariates case ($d \geq 2$), see [12]. Note that for $\theta = 0$ Y and X_1 are independent and $m(x_1) = E[Y]$. Another example, if the copula of $(Y, \boldsymbol{X})$ is Gaussian,

then

$$m(\boldsymbol{x}) = E\left[F_0^{-1}\left\{\Phi\left(\boldsymbol{u}^T\Sigma_{\boldsymbol{X}}^{-1}\boldsymbol{\rho} + Z\sqrt{1-\boldsymbol{\rho}^T\Sigma_{\boldsymbol{X}}^{-1}\boldsymbol{\rho}}\right)\right\}\right],$$

where $\boldsymbol{u} = (\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_d(x_d)))^T$, $\boldsymbol{\rho} = (corr(Y, X_1), \ldots, corr(Y, X_d))^T$, $Z \sim N(0,1)$, Φ is the standard normal cdf and $\Sigma_{\boldsymbol{X}}$ is the correlation matrix of $\boldsymbol{X}$. For $\boldsymbol{\rho} = 0$, Y and $\boldsymbol{X}$ are independent and m is simplified to $m(x_1) = E\left[F_0^{-1}(\Phi(Z))\right] \equiv E[Y]$. A semiparametric approach was taken recently by [20] to estimate m for complete data. The latter assume a parametric model for the copula density and estimate the marginal distributions using nonparametric methods.

In this paper, we propose a semiparametric copula-based estimator for the regression function when the response variable is subject to right-censoring. Here, we assume a parametric model for the copula density and we propose to estimate the marginal distributions F_0 by the product-limit estimator and $F_1, \ldots, F_d$ by their empirical counterparts. To the best of our knowledge, such method has never been proposed or studied in the literature. Assuming a parametric model for the copula density avoids the curse of dimensionality but requires the choice of the adequate copula model.

The paper is organized as follows. In Section 2, we define our estimator for the regression function in the case of right censored data. In Section 3, we establish the asymptotic properties of the proposed estimator by providing its i.i.d. representation, and showing its uniform weak convergence and asymptotic normality. In section 4, we apply our methodology to analyze a data-set on lifetime with lung-cancer. The proofs of the theoretical results are presented in the appendix.

2. Estimator for right-censored data

Suppose that the response Y is subject to right-censoring by a random variable C and one observes the vector $(Z, \delta) = \big(\min(Y, C), \mathrm{I}(Y \leq C)\big)$, where δ indicates if Y is censored or not. We assume that Y and C are independent and we consider uncensored covariates $X_1, \ldots, X_d$. The data has the form $\big\{\big(Z_i, \delta_i, X_{1,i}, \ldots, X_{d,i}\big),\, i = 1, \ldots, n\big\}$. Based on the relationship (1), the regression function $m(\boldsymbol{x}) = \mathbb{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right]$ can be expressed in terms of the copula density and the marginal distributions as

$$m(\boldsymbol{x}) = \mathbb{E}\Big[Y\, w\big(F_0(Y), \boldsymbol{F}(\boldsymbol{x})\big)\Big] = \frac{e(\boldsymbol{F}(\boldsymbol{x}))}{c_X(\boldsymbol{F}(\boldsymbol{x}))}, \tag{2}$$

where $w(u_0, \boldsymbol{u}) = c(u_0, \boldsymbol{u})/c_X(\boldsymbol{u})$ and

$$e(\boldsymbol{u}) = \int_0^1 F_0^{-1}(u_0)\, c(u_0, \boldsymbol{u})\, \mathrm{d}u_0,$$

with F_0^{-1} the inverse of F_0. In the particular cases of a single covariate ($d = 1$) or mutually independent predictors, $c_X = 1$ and $m(x_1)$ is reduced to $e(\boldsymbol{F}(x_1))$.

Formula (2) allows different ways for estimating m. One approach assumes a parametric model for the copula densities and the marginal distributions. The parameters of the copula density and those of the marginal distributions can be estimated by using maximum likelihood (ML) or inference functions for margins method ((IFM)) in two steps [see [23] and [9]]. The IFM method provides results similar to that of maximum likelihood (ML) and is easier to implement This method is however too restrictive, and a misspecification of the parametric model leads to wrong conclusions. A second way is to estimate nonparametrically the copula density and the marginal distributions. For example, the method proposed by [8] and [2] to estimate the copula density can be adapted to right-censored data. However, this method suffers from the curse of dimensionality and requires practical bandwidth parameters. The third approach is semiparametric. This method considers a parametric model for the copula density and nonparametrically estimates the marginal distributions. The semiparametric approach is less restrictive than the parametric method and avoids the curse of dimensionality of the nonparametric estimator. The estimation of the copula distribution based on a semiparametric approach have been proposed by [7], for complete data, and by [23] for right-censored data. Based on extensive simulations, [10] showed that the semiparametric approach outperforms the ML and IFM methods.

To estimate the regression function m, we consider a parametric model $c(\cdot,\cdot;\theta)$ for the copula density, where θ is an unknown parameter vector, and nonparametric estimators for F_0 and F_j $(j=1,\ldots,d)$. First, the marginal distribution of the response is replaced by the Kaplan-Meier estimator, Γ_n, defined as follows :

$$\Gamma_n(t) = 1 - \prod_{\substack{1\le i\le n\\ Z_{(i)}<t}} \left[\frac{n-i}{n-i+1}\right]^{\delta_{(i)}},$$

where $Z_{(1)},\ldots,Z_{(n)}$ are the ordered responses and $\delta_{(1)},\ldots,\delta_{(n)}$ their respective concomitant censoring indicators. The marginal distributions of the covariates are estimated by the rescaled empirical distribution $\widehat{F}_j(t)$:

$$\widehat{F}_j(t) = \frac{1}{n+1}\sum_{i=1}^{n} 1(X_{j,i}\le t),$$

with $X_{j,1},\ldots,X_{j,n}$ i.i.d random copies of X_j $(j=1,\ldots,d)$.
Then, we estimate the copula parameter θ by the estimator $\hat{\theta}$ that maximises the pseudo maximum likelihood given in equation (11) of [23]. Therefore, an estimator for the numerator $e(\boldsymbol{F}(\boldsymbol{x}))$ is

$$\hat{e}(\widehat{\boldsymbol{F}}(\boldsymbol{x})) = \sum_{i=1}^{n} Z_i\, w_i\, c(\Gamma_n(Z_i), \widehat{\boldsymbol{F}}(\boldsymbol{x});\widehat{\theta}), \tag{3}$$

where $\widehat{\boldsymbol{F}}(\boldsymbol{x}) = \big(\widehat{F}_1(x_1),\ldots,\widehat{F}_d(x_d)\big)$ and w_i is the mass attached to Z_i under the Kaplan-Meier estimator, i.e., $w_1 = \Gamma_n(Z_{(1)})$ and $w_i = \Gamma_n(Z_{(i)}) - \Gamma_n(Z_{(i-1)}), i = 2,\ldots,n$. Now, using the fact that $c_X(\boldsymbol{u}) = \mathbb{E}\Big[c\big(F_0(Y),\boldsymbol{u}\big)\Big]$, we estimate the denomi-

nator $c_X(\boldsymbol{F}(\boldsymbol{x}))$ by

$$\widehat{c}_{\boldsymbol{X}}(\widehat{\boldsymbol{F}}(\boldsymbol{x})) = \sum_{i=1}^{n} w_i \, c(\Gamma_n(Z_i), \widehat{\boldsymbol{F}}(\boldsymbol{x}); \widehat{\theta}), \tag{4}$$

Thus, our semiparametric estimator for $m(\boldsymbol{x})$ is

$$\widehat{m}(\boldsymbol{x}) = \frac{\hat{e}(\widehat{\boldsymbol{F}}(\boldsymbol{x}))}{\widehat{c}_{\boldsymbol{X}}(\widehat{\boldsymbol{F}}(\boldsymbol{x}))} = \sum_{i=1}^{n} Z_i \frac{w_i \, c(\Gamma_n(Z_i), \widehat{\boldsymbol{F}}(\boldsymbol{x}); \widehat{\theta})}{\sum_{i=1}^{n} w_i \, c(\Gamma_n(Z_i), \widehat{\boldsymbol{F}}(\boldsymbol{x}); \widehat{\theta})}. \tag{5}$$

In the particular case of a single covariate ($d = 1$), or mutually independent predictors $X_1, \ldots, X_d$, the estimator of $m(\boldsymbol{x})$ is reduced to

$$\widehat{m}(\boldsymbol{x}) = \sum_{i=1}^{n} Z_i \, w_i \, c(\Gamma_n(Z_i), \widehat{\boldsymbol{F}}(\boldsymbol{x})); \widehat{\theta}). \tag{6}$$

Let $L(z) = P[Z \leq z]$ and $\tau = \sup\{z : \, L(z) < 1\} < \infty$. Since the tail region information on the survival function of Y may not be identifiable in $[\tau, \infty)$ due to right censoring, we note that Γ_n is defined on the interval $[0, \tau]$.

Remark 1.
1. We can consider other estimators, $\widetilde{\Gamma}_n$ and $\widetilde{F}_j$ ($j = 1, \ldots, d$), to estimate the response and marginal distributions provided that the following conditions are satisfied :

(i) $\widetilde{\Gamma}_n(t) = \Gamma_n(t) + o_p(n^{-1/2})$.
(ii) $\widetilde{F}_j(t) = \widehat{F}_j(t) + o_p(n^{-1/2})$, for $j = 1, \ldots, d$.

2. The model in (2) can be extended to mixed (continuous and discrete) covariates. However, this model can't be used for a discrete response. The idea proposed in [4], for complete data, can be adapted to our model by assuming a latent variable framework to describe discrete outcomes.

3. In the case where some of the predictors are subject to right-censoring, we may estimate their distributions by the Kaplan-Meier estimator instead.

4. The nonparametric kernel estimator (Nadaraya-Watson type)

$$\widehat{m}^{\mathcal{NP}}(x) = \sum_{i=1}^{n} \frac{w_i \, K_d(\frac{x - X_{\cdot,i}}{h})}{\sum_{\ell=1}^{n} w_\ell \, K_d(\frac{x - X_{\cdot,\ell}}{h})} \, Z_i,$$

with K_d a kernel function defined on $\mathbb{R}^d$ and h the smoothing parameter, is a robust alternative to our semiparametric estimator. This estimator requires less assumptions and is more flexible. However, the a.s. convergence rate of $\widehat{m}^{\mathcal{NP}}$ is $O_{a.s.}\left(n^{-1}\log(n)\right)^{1/(d+2)}$ in the presence of d covariates. This rate becomes slower as the dimension of the covariates increases.

3. Asymptotic results

We start by presenting some notations and assumptions, on the copula and its parameters, needed for establishing the asymptotic properties of the regression function estimator. Denote $\partial_j c = \partial c/\partial u_j$ for $j = 0, \ldots, d$, $\dot{\mathbf{c}} = (\partial c/\partial\theta_1, \ldots, \partial c/\partial\theta_p)^T$ and let $\boldsymbol{x} \in \mathbb{R}^d$ such that $\boldsymbol{F}(\boldsymbol{x}) \in (0,1)^d$.

Assumption A: Let g be either c, $\dot{\mathbf{c}}$ or $\partial_j c$ $(j = 0, \ldots, d)$.

(i) $(\mathbf{u}, \theta) \to g_{u_0}(\mathbf{u}, \theta) \equiv g(u_0, \mathbf{u}; \theta)$ is continuous at $(\boldsymbol{F}(\boldsymbol{x}), \theta_0)$, uniformly on $u_0 \in [0,1]$.

(ii) $u_0 \to g(u_0, \boldsymbol{F}(\boldsymbol{x}); \theta_0)$ is continuous on $[0,1]$.

Assumption B: The parameter estimator $\widehat{\theta}$ satisfies,

$$\widehat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} \zeta_i + o_p(n^{-1/2}),$$

where ζ_i are i.i.d. random variables with zero mean and finite variance.

As noted above, we consider the estimation of θ proposed by Shih & Louis (1995) for a bivariate parametric copula model. This estimator satisfies Assumption B, and by a similar approach, we may extend the latter representation to a multivariate parametric copula model ($d > 2$).

Assumption C:
The copula density is bounded and bounded away from zero on its compact support.

The copula density c is required to be bounded away from 0 for a regression model with multiple covariates. All the parametric copula densities (Gaussian, Student, Archimedean copula, etc) are bounded away from zero. These densities are bounded on $(0,1)^d$, but some of them are unbounded at the corners of $[0,1]^d$. To relax Assumption C, the proofs in the appendix can be adapted by considering a copula density that satisfies the condition

$$c(u_0, u_1, ..., u_d) = O\left(\frac{1}{\sqrt{\prod_{j=0}^{d} u_j(1-u_j)}}\right).$$

This condition is satisfied by many common copula densities (see [21]).

3.1. Main results

We begin by considering the case of one covariate X_1 $(d = 1)$. In Proposition 3.1, we establish the uniform weak convergence of the proposed regression function estimator $\widehat{m}$. The i.i.d. representation of $\widehat{m}$ is derived in Theorem 3.2, while in Corollary 3.3, the asymptotic normality of $\widehat{m}$ is established.

Proposition 3.1. *Under Assumptions A and B we have*

$$\sup_{x_1} \left|\widehat{m}(x_1) - m(x_1)\right| = \mathcal{O}_p\big(\sqrt{n^{-1}\log\log n}\big).$$

Proof. The proof is given in the appendix. □

Denote

$$\begin{aligned}
\eta_{i,1}(x_1) &= \int_0^\tau y\,\xi_i(y)\,\partial_0 c(F_0(y), F_1(x_1);\theta_0)\,\mathrm{d}F_0(y),\\
\eta_{i,2}(x_1) &= \big[\mathrm{I}(X_{1i} \le x_1) - F_1(x_1)\big] \int_0^\tau y\,\partial_1 c(F_0(y), F_1(x_1);\theta_0)\,\mathrm{d}F_0(y),\\
\eta_{i,3}(x_1) &= \int_0^\tau y\,\zeta_i\,\dot{\mathbf{c}}(F_0(y), F_1(x_1);\theta_0)\,\mathrm{d}F_0(y),\\
\eta_{i,4}(x_1) &= \int_0^\tau \xi_i(y)\,\mathrm{d}H_{x_1}(y), \text{ with } H_{x_1}(y) = yc(F_0(y), F_1(x_1);\theta_0).
\end{aligned}$$

with $\xi_i(y)$ the i.i.d. random term of the representation of the Kaplan-Meier estimator Γ_n, see [14]. The additional terms $\eta_{i,1}$, $\eta_{i,2}$ and $\eta_{i,4}$ are needed because the marginal distributions are estimated in the first step.

Theorem 3.2. *Let $\eta_i = \sum_{j=1}^4 \eta_{i,j}$. Under Assumptions A and B, $\widehat{m}$ admits the i.i.d. representation*

$$\widehat{m}(x_1) = m(x_1) + \frac{1}{n}\sum_{i=1}^n \eta_i(x_1) + o_p(n^{-1/2}) \tag{7}$$

The proof is is given in the appendix. The representation of $\widehat{m}$ in (7) leads to the following result.

Corollary 3.3. *Under Assumptions A and B, $\sqrt{n}\big[\widehat{m}(x_1) - m(x_1)\big]$ converges to a normal distribution with zero mean and variance $\sigma^2(x_1) = \mathbb{E}\left[\eta_1^2(x_1)\right]$.*

3.2. Extension to the multivariate case

Following similar steps in proof of Proposition 3.1, we can establish the uniform weak convergence, with the same order as in Proposition 3.1, of $\widehat{m}$ with multiple covariate. Next, we extend the result in Theorem 3.2 to the multivariate case $d \ge 2$, by following similar arguments in Theorem 3.2's proof. First, one can easily check that $\hat{e}$ admits the representation

$$\hat{e}(\hat{\boldsymbol{F}}(\boldsymbol{x})) - e(\boldsymbol{F}(\boldsymbol{x})) = n^{-1}\sum_{i=1}^n \varphi_i(\boldsymbol{x}) + o_p(n^{-1/2}), \tag{8}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)$, $\boldsymbol{F}(\boldsymbol{x}) = (F_1(x_1), \ldots, F_d(x_d))$ and $\varphi_i = \sum_{j=1}^4 \varphi_{i,j}$, with

$$\begin{aligned}
\varphi_{i,1}(\boldsymbol{x}) &= \int_0^\tau y\, \xi_i(y)\, \partial_0 c(F_0(y), \boldsymbol{F}(\boldsymbol{x}); \theta_0)\, \mathrm{d}F_0(y), \\
\varphi_{i,2}(\boldsymbol{x}) &= \sum_{j=1}^d \big[\mathrm{I}(X_{ji} \le x_j) - F_j(x_j)\big] \int_0^\tau y\, \partial_j c(F_0(y), \boldsymbol{F}(\boldsymbol{x}); \theta_0)\, \mathrm{d}F_0(y), \\
\varphi_{i,3}(\boldsymbol{x}) &= \int_0^\tau y\, \zeta_i^T\, \dot{\mathbf{c}}(F_0(y), \boldsymbol{F}(\boldsymbol{x}); \theta_0)\, \mathrm{d}F_0(y)
\end{aligned}$$

and

$$\varphi_{i,4}(\boldsymbol{x}) = \int_0^\tau \xi_i(y)\, \mathrm{d}H^*_{\boldsymbol{x}}(y) \text{ with } H^*_{\boldsymbol{x}}(y) = c(F_0(y), \boldsymbol{F}(\boldsymbol{x}); \theta_0).$$

Second, one can show that $\widehat{c}_{\boldsymbol{X}}$ has the representation

$$\widehat{c}_{\boldsymbol{X}}(\hat{\boldsymbol{F}}(\boldsymbol{x})) - c_X(\boldsymbol{F}(\boldsymbol{x})) = n^{-1} \sum_{i=1}^n \phi_i(\boldsymbol{x}) + o_p(n^{-1/2}), \tag{9}$$

where $\phi_i = \sum_{j=1}^4 \phi_{i,j}$, with

$$\begin{aligned}
\phi_{i,1}(\boldsymbol{x}) &= \int_0^\tau \xi_i(y)\, \partial_0 c(F_0(y), \boldsymbol{F}(\boldsymbol{x}); \theta_0)\, \mathrm{d}F_0(y), \\
\phi_{i,2}(\boldsymbol{x}) &= \sum_{j=1}^d \big[\mathrm{I}(X_{ji} \le x_j) - F_j(x_j)\big] \int_0^\tau \partial_j c(F_0(y), \boldsymbol{F}(\boldsymbol{x}); \theta_0)\, \mathrm{d}F_0(y), \\
\phi_{i,3}(\boldsymbol{x}) &= \int_0^\tau \zeta_i^T\, \dot{\mathbf{c}}(F_0(y), \boldsymbol{F}(\boldsymbol{x}); \theta_0)\, \mathrm{d}F_0(y)
\end{aligned}$$

and

$$\phi_{i,4}(\boldsymbol{x}) = \int_0^\tau \xi_i(y)\, \mathrm{d}H^*_{\boldsymbol{x}}(y).$$

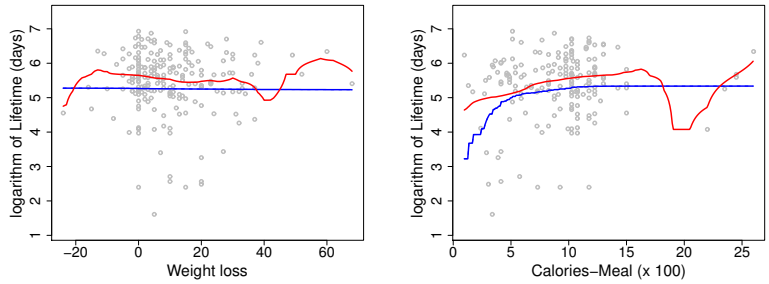
Combining (8) with (9) leads to the next main result.

Theorem 3.4. *Under Assumptions A, B and C, we have*

$$\widehat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) = n^{-1} \sum_{i=1}^n \frac{1}{c_X(\boldsymbol{F}(\boldsymbol{x}))} \Big[\varphi_i(\boldsymbol{x}) - m(\boldsymbol{x})\phi_i(\boldsymbol{x})\Big] + o_p(n^{-1/2}).$$

Theorem 3.4 implies that $\sqrt{n}\big[\widehat{m}(\boldsymbol{x}) - m(\boldsymbol{x})\big]$ follows asymptotically a normal distribution with mean zero and variance

$$\sigma^2(\boldsymbol{x}) = \mathbb{V}ar\left(\frac{1}{c_X(\boldsymbol{F}(\boldsymbol{x}))}\Big[\varphi_i(\boldsymbol{x}) - m(\boldsymbol{x})\phi_i(\boldsymbol{x})\Big]\right).$$



(a) lifetime vs Weight-Loss. (b) lifetime vs Calories-Meals.

Figure 1. Regression function estimators of patients with lung cancer: (a) lifetime vs Weight-Loss, (b) lifetime vs Calories-Meals. Copula-based estimator (blue line) and nonparametric kernel estimator (red line).

4. Data analysis

4.1. Survival with lung cancer

We illustrate the methodology described in Section 2 by analysing a set of survival data from patients with advanced lung cancer collected by the "North Central Cancer Treatment Group" (see [16]). The primary goal of [16] was the investigation of whether descriptive information from a patient-completed questionnaire could provide prognostic information that was independent from the one obtained by the physician. The questionnaire (which contains different questions about age, calories intake, weight-loss in the last six months, etc) was completed by the patients before entering the study of lung cancer. This study was mainly focused on dietary and daily activities factors and their possible relation to risk of lung cancer. The data set contains the lifetimes of $n = 228$ individuals (men and women), among whom 165 died and 63 were right-censored during the follow-up. The survival time Y is defined as the time between onset of lung cancer and death. In this example, we consider two covariates: X_1 as the weight loss in the last six months prior to entering the study and X_2 as the average of meal calories per day before entering the study; hence the two covariates are completely observed.

To choose the copula function for our semiparametric regression estimator, we consider the copula model $\{c(.;\theta), \theta \in \Theta\}$ that minimizes the weighted sum of squared residuals defined by

$$\sum_{i=1}^{n} w_i \left[Z_i - \widehat{m}_c(X_i) \right]^2. \tag{10}$$

This leads to select the Gaussian and Clayton copulas for estimating the regression functions of Lifetime vs Weight-Loss and Lifetime vs Calories-Meals, respectively. Figure 1 displays the plots of the semiparametric copula-based estimator (blue line) and the nonparametric kernel estimator (red line) defined by

$$\widehat{m}^{\mathcal{NP}}(x) = \sum_{i=1}^{n} \frac{w_i \, K\Big(\dfrac{x - X_i}{h_n}\Big)}{\displaystyle\sum_{\ell=1}^{n} w_\ell \, K\Big(\dfrac{x - X_\ell}{h_n}\Big)} \, Z_i, \tag{11}$$

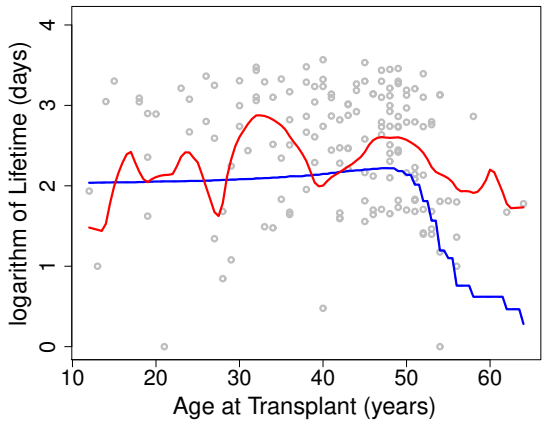


Figure 2. Regression function estimators of 157 heart transplant patients. Copula-based estimator with Joe's copula (blue line) and nonparametric kernel estimator (red line).

where $K(x) = 0.75(1-x^2)\,\mathrm{I}(-1 \le x \le 1)$ and $h_n > 0$ is a sequence of bandwidth tending to 0. Both estimators show similar behaviors and fluctuate about the survival-time $e^{5.3} \approx 200$ days. The horizontal tendency of the copula-based estimators can be explained by the weak dependence between the response (lifetime) and the two covariates. In [22], we found that the estimates of Kendall's tau measure of association is $\widehat{\tau}_{Y,X_1} = -0.0848$, for Lifetime vs Weight-Loss, and $\widehat{\tau}_{Y,X_2} = 0.1335$, for Lifetime vs Calories-Meals. This indicates weak association between the response and the covariates. These results concur with the findings of [16]. Note that in (11) we select the bandwidth h_n that minimizes the weighted integrated squared error $\mathrm{WISE}(h) = \int_{x>0} \left[\widehat{m}^{NP}(x;h) - m(x)\right]^2 \mathrm{d}F_X(x)$, given by

$$\widehat{h} = \arg\min_h \left\{ \sum_{i=1}^{n} w_i \left[\widehat{m}^{NP}_{-i}(X_i;h) - Z_i\right]^2 \right\},$$

where $\widehat{m}^{\mathcal{N}P}_{-i}(x)$ is a leave-one-out estimate of m at x;

$$\widehat{m}^{\mathcal{N}P}_{-i}(x;h) = \sum_{\substack{k=1 \\ k\neq i}}^{n} \frac{w_k\, K\Big(\dfrac{x-X_k}{h}\Big)}{\displaystyle\sum_{\substack{j=1 \\ j\neq i}}^{n} w_j\, K\Big(\dfrac{x-X_j}{h}\Big)}\, Z_k.$$

4.2. Heart transplant data

The Stanford survival data were extracted from a program of heart transplant which began in October 1967 (see [19]). The patients had to be selected to take part of that program and therefore received a transplant. Some people die between the selection of the patient and the transplant, which leads their survival time to be 0. The cut-off date

of the study was February 1980, and at this moment, the data about heart transplant of 184 patients was collected. The variables of interest in this analysis are the survival time Y (in days), the failure status δ (1 if dead and 0 if alive and censored) and the age at the time of the first transplant X. We only keep patients who their mismatch score is not missing. Therefore, we get a sample of $n = 157$ data with a censorship of 35%. Figure 2 illustrates the semiparametric copula-based estimator (blue line) and the nonparametric kernel-estimator (red line), given by (11). The copula-based estimator shows a slight increase up to Age-of-Transplant = 50, and then a decrease after that. This indicates a rapid decline in the survival of patients who received a transplant after the age of fifty. The Age-of-Transplat variable (X) seems to have an impact on the survival time (Y). In [22], the estimation of Spearman's rho measure of association is $\widehat{\rho}_{Y,X} = -0.626$. This reflects a moderate strong association between X and Y, and concurs with the finding of [19]. Note that we employed the weighted sum of squared residuals defined in (10) to select the Joe copula for this data.

5. Conclusion

We introduced a new copula-based estimator for the regression function when the response variable is subject to right-censoring and the covariate(s) is completely observed. The key element in the paper is to write the regression function in terms of the copula density and marginal distributions. The estimation method uses a parametric model for the copula density, with unknown parameter(s), and estimate nonparametrically the marginal distributions of the response and covariate(s). We studied the asymptotic behavior of our estimator analytically, and we extend it to the multivariate case. The proposed estimation method have shown satisfactory results in analyzing two real data-sets, concerning survival with heart-transplant and lifetime with lung-cancer.

Acknowledgement(s)

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Canssi and Statlab-CRM (Canada) and Le Fonds de recherche du Québec (Nature et technologies).

References

[1] Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, pages 429–436, 1979.

[2] S. X. Chen and T. Huang. Nonparametric estimation of copula functions for dependent modeling. *Canadian Journal of Statistics*, 35:265–282, 2007.

[3] G.J. Crane and J. Van Der Hoek. Conditional Expectation Formulae for Copulas. *Australian and New Zealand Journal of Statistics*, 50:53–67, 2008.

[4] A. R. de Leon and B. Wu. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics and Medecine*, 30:175–185, 2011.

[5] K. Doksum and B. Yandell. Properties of regression estimates based on censored survival data. *In A festschrift for Erich L. Lehmann (eds P. J. Bickel, K. Doksum and J.L. Hodges Jr.)*, pages 140–156, 1983.

[6] J. Fan and I. Gijbels. Censored-regression: local linear approximations and their applications. *Journal of the American Statistical Association*, 89:560–570, 1994.

[7] C. Genest, K. Ghoudi, and L.P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552, 1995.

[8] I. Gijbels and J. Mielniczuk. Estimating the density of a copula function. *Communications in Statistics - Theory and Methods*, 19, 1990.

[9] H. Joe and J. Xu. The estimation method of inference functions for margins for multivariate models. *Technical Report, Department of Statistics, University of British Columbia*, 166, 1996.

[10] G. Kim, M. Silvapulle, and P. Silvapulle. Comparison of the semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis*, 51: 2836–2850, 2007.

[11] H. Koul, V. Susarla, and J. Van Ryzin. Regression analysis with randomly right censored data. *Annals of Statistics*, 9:1276–1288, 1981.

[12] Yeo Keng Leong and Emiliano A Valdez. Claims Prediction with Dependence using Copula Models. Technical report, 2005.

[13] S. Leurgans. Linear models, random censoring and synthetic data. *Biometrika*, 74:301–309, 1987.

[14] S. Lo, Y. Mack, and J. Wang. Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probability Theory and Related Fields*, 80:461–473, 1989.

[15] S.H. Lo, Y.P Mack, and J.L Wang. Density and hazard rate estimation for censored data via strong reptresentation of the kaplan-meier estimator. *Probability Theory and Relatd Fields*, 80:473–473, 1989.

[16] C.L. Loprinzi et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *J. Clin. Oncol.*, 12:601–607, 1994.

[17] X.L. Meng, Y. Bassiakos, and S.H. Lo. Large-sample properties for a general estimator of the treatment effect in the two-sample problem with right censoring. *Ann. Statist.*, 19: 1786–1812, 1991.

[18] R. Miller. Least squares regression with censored data. *Biometrika*, 63:449–464, 1976.

[19] R. Miller and J. Halpern. Regression with Censored Data. *Biometrika*, 69:521–531, 1982.

[20] H. Noh, A. El Ghouch, and T. Bouezmarni. Copula-Based Regression Estimation and Inference. *Journal of the American Statistical Association*, 108:676–688, 2013.

[21] M Omelka, I. Gijbels, and N Veraverbeke. Improved kernel estimation of copulas: Weak convergence and goodnees-of-fit testing. *Annals of Statistics*, 37:3023–3058, 2009.

[22] Y. Rabhi and T. Bouezmarni. Nonparametric inference for copula density function under random censoring. Technical report, Dept. of Mathematics, University of Sherbrooke., 2016. (Technical Report 2016-150).

[23] J. Shih and T. Louis. Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. *Biometrics*, 51:1384–1399, 1995.

[24] C. Srinivasan and M. Zhou. Linear regression with censoring. *Journal of Multivariate Analysis*, 49:179–201, 1994.

[25] E. A. Sungur. Some observations on copula regression functions. *Communications in Statistics-Theory and Methods*, 34:1967–1978, 2005.

[26] Z. Zheng. Regression analysis with censored data. *PhD Dissertation*, Columbia University, 1984.

[27] Z. Zheng. A class of estimatord for the parameters in linear regression with censored data. *Acta Math. Appl. Sin.*, 3:231–241, 1987.

[28] M. Zhou. Asymptotic normality of the "synthetic data" regression estimator for censored survival data. *Annals of statistics*, 20:1002–1021, 1992.

Appendix A. Appendix: Proofs of main results

For a function S defined on the set $\mathbb{S}$, we denote by $||S|| = \sup_{x\in\mathbb{S}} |S(x)|$ in the proofs.

Proof of Proposition 3.1.
Recall that the copula density c is a smooth function. The difference $\widehat{m} - m$ is equal to

$$\widehat{m}(x_1) - m(x_1) = \int_0^\tau y \left[c\big(\Gamma_n(y), \widehat{F}_1(x_1); \widehat{\theta}\big) - c\big(F_0(y), F_1(x_1); \theta\big)\right] \mathrm{d}\Gamma_n(y) + \int_0^\tau y\, c\big(F_0(y), F_1(x_1); \theta\big)\, \mathrm{d}\big[\Gamma_n(y) - F_0(y)\big]$$

Let $H_{x_1}(y) = y\, c\big(F_0(y), F_1(x_1); \theta\big)$, $\widehat{F}_0(y) = \Gamma_n(y) - F_0(y)$, $\Delta H_{x_1}(t) = H_{x_1}(t) - H_{x_1}(t^-)$ and $\Delta\widehat{F}_0(t) = \widehat{F}_0(t) - \widehat{F}_0(t^-)$. By partial integration formula for the Lebesgue-Stieltjes integral

$$\begin{aligned}\int_0^\tau H_{x_1}(y^-)\, \mathrm{d}\widehat{F}_0(y) &= H_{x_1}(\tau)\widehat{F}_0(\tau) - H_{x_1}(0)\widehat{F}_0(0) - \int_0^\tau \widehat{F}_0(y^-)\, \mathrm{d}H_{x_1}(y) - \sum_{0\le t\le\tau} \Delta H_{x_1}(t)\, \Delta\widehat{F}_0(t) \\ &= H_{x_1}(\tau)\widehat{F}_0(\tau) - \int_0^\tau \widehat{F}_0(y^-)\, \mathrm{d}H_{x_1}(y),\end{aligned}$$

because $H_{x_1}(0) = 0$ and $\Delta H_{x_1}(t) = 0$, since H_{x_1} is continuous on $[0, \tau]$. Hence

$$\widehat{m}(x_1) - m(x_1) = \int_0^\tau y \left[c\big(\Gamma_n(y), \widehat{F}_1(x_1); \widehat{\theta}\big) - c\big(F_0(y), F_1(x_1); \theta\big)\right] \mathrm{d}\Gamma_n(y) + H_{x_1}(\tau)\widehat{F}_0(\tau) - \int_0^\tau \widehat{F}_0(y^-)\, \mathrm{d}H_{x_1}(y). \tag{A1}$$

By using Mean Value Theorem for a multivariate real-valued differentiable functions in the 1^{st} term on the R.H.S. of (A1), and the uniform convergence results $\|\Gamma_n - F_0\| = \mathcal{O}_{a.s.}\big(\sqrt{n^{-1}\log\log(n)}\big)$, $\|\widehat{F}_1 - F_1\| = \mathcal{O}_{a.s.}\big(\sqrt{n^{-1}\log\log(n)}\big)$ and $|\widehat{\theta} - \theta| = \mathcal{O}_p\big(1/\sqrt{n}\big)$, the result follows.

□

Proof of Theorem 3.2.
Notice that $\widehat{m}(x_1) - m(x_1)$ can be expressed as

$$\begin{aligned}\widehat{m}(x_1) - m(x_1) &= \int_{y\ge0} y \left[c\big(\Gamma_n(y), \widehat{F}_1(x_1); \widehat{\theta}\big) - c\big(F_0(y), F_1(x_1); \theta\big)\right] \mathrm{d}[\Gamma_n(y) - F_0(y)] \\ &+ \int_{y\ge0} y \left[c\big(\Gamma_n(y), \widehat{F}_1(x_1); \widehat{\theta}\big) - c\big(F_0(y), F_1(x_1); \theta\big)\right] \mathrm{d}F_0(y) \\ &+ \int_{y\ge0} y\, c\big(F_0(y), F_1(x_1); \theta\big)\, \mathrm{d}[\Gamma_n(y) - F_0(y)].\end{aligned}$$

Let $r_n(x_1)$, I_2 and I_3 denote, respectively, the first, second and thirds terms on the R.H.S. of the latter equality. Using Taylor expansion of first order on c in I_2, partial integration on I_3 and the uniform convergence results $\|\Gamma_n - F_0\| = \mathcal{O}_{a.s.}\big(\sqrt{n^{-1}\log\log(n)}\big)$, $\|\widehat{F}_1 - F_1\| = \mathcal{O}_{a.s.}\big(\sqrt{n^{-1}\log\log(n)}\big)$ and $|\widehat{\theta} - \theta| = \mathcal{O}_p\big(1/\sqrt{n}\big)$, we obtain

$$\begin{aligned}
\widehat{m}(x_1) - m(x_1) = r_n(x_1) &+ \int_{y\geq 0} y\left[\Gamma_n(y) - F_0(y)\right]\partial_1 c\big(F_0(y), F_1(x_1);\theta\big)\,\mathrm{d}F_0(y) \\
&+ \int_{y\geq 0} y\left[\widehat{F}_1(x_1) - F_1(x_1)\right]\partial_2 c\big(F_0(y), F_1(x_1);\theta\big)\,\mathrm{d}F_0(y) \\
&+ \int_{y\geq 0} y\left[\widehat{\theta} - \theta\right]\dot{c}\big(F_0(y), F_1(x_1);\theta\big)\,\mathrm{d}F_0(y) \\
&+ \int_{y\geq 0} \left[\Gamma_n(y) - F_0(y)\right]\mathrm{d}\big[y\,c\big(F_0(y), F_1(x_1);\theta\big)\big] + \mathcal{O}_p\big(n^{-1}\log\log(n)\big). \qquad \text{(A2)}
\end{aligned}$$

Now, let's focus on $r_n(x_1)$. Divide $[0,\tau]$ into r sub-intervals $[0, y_1]$, $[y_1, y_2], \ldots, [y_{r-1}, y_r]$ of equal length $\ell = a_0 n^{-1/2}(\log n)^q$ ($q \geq 1/2$ and $a_0 > 0$ is some constant), so r is of order $\mathcal{O}\left(n^{1/2}(\log n)^{-q}\right)$. We have

$$\begin{aligned}
\big|r_n(x_1)\big| &\leq \sum_{i=0}^{r-1}\left|\int_{y_i}^{y_{i+1}} y\Big[c\big(\Gamma_n(y), \widehat{F}_1(x_1);\widehat{\theta}\big) - c\big(F_0(y), F_1(x_1);\theta\big)\Big]\,\mathrm{d}[\Gamma_n(y) - F_0(y)]\right| \\
&\leq \sum_{i=0}^{r-1}\tau\left\{\|\Gamma_n - F_0\|.\|\partial_1 c\| + \|\widehat{F}_1 - F_1\|.\|\partial_2 c\| + |\widehat{\theta} - \theta|.\|\dot{c}\|\right\}\int_{y_i}^{y_{i+1}}\big|\mathrm{d}\big[\Gamma_n(y) - F_0(y)\big]\big| \\
&\leq \tau\left\{\|\Gamma_n - F_0\|.\|\partial_1 c\| + \|\widehat{F}_1 - F_1\|.\|\partial_2 c\| + |\widehat{\theta} - \theta|.\|\dot{c}\|\right\} \\
&\quad\times \sum_{i=0}^{r-1}\sup_{u,v\in[y_i, y_{i+1}]}\big|\big[\Gamma_n(v) - F_0(v)\big] - \big[\Gamma_n(u) - F_0(u)\big]\big|\,. \qquad \text{(A3)}
\end{aligned}$$

The sup-norm term, inside the summation, on the R.H.S. of (A3) is of order $\mathcal{O}_{a.s.}\big(n^{-3/4}(\log n)^{\frac{1+q}{2}}\big)$, as $n \to \infty$, by the oscillation result in [17] (see proposition 1, page 6). Since r and the 1^{st} term in (A3) are of order $\mathcal{O}\big(n^{-1/2}(\log n)^{-q}\big)$ and $\mathcal{O}_p\big(n^{-1/2}\big(\log\log n\big)^{1/2}\big)$, respectively, the term on the R.H.S. of (A3) is of order $\mathcal{O}_{a.s.}\left(n^{-3/4}(\log n)^{\alpha_1}\right)$ ($\alpha_1 \geq 1$). Hence,

$$\sup_x |r_n(x_1)| = \mathcal{O}_{a.s.}\left(n^{-3/4}(\log n)^{\alpha_1}\right),$$

The result then follows from equation (A2) by using the i.i.d. representations of $\Gamma_n - F_0$, in [15], and $\widehat{\theta} - \theta$ in Assumption B2. □

Proof of Corollary 3.3. The result follows from Theorem 3.2 using Central Limit Theorem. □

Proof of Theorem 3.4. First, remark that

$$\widehat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) = \frac{1}{\widehat{c}_{\mathcal{X}}(\boldsymbol{x})}[\widehat{e}(\boldsymbol{x}) - e(\boldsymbol{x})] + \frac{e(\boldsymbol{x})}{\widehat{c}_{\mathcal{X}}(\boldsymbol{x})\, c_{\mathcal{X}}(\boldsymbol{x})}[\widehat{c}_{\mathcal{X}}(\boldsymbol{x}) - c_{\mathcal{X}}(\boldsymbol{x})].$$

Analogously to the proof of Proposition 3.1, by using the uniform results $\|\Gamma_n - F_0\| = \mathcal{O}_{a.s.}\big(\sqrt{n^{-1}\log\log(n)}\big)$, $\|\widehat{F}_i - F_i\| = \mathcal{O}_{a.s.}\big(\sqrt{n^{-1}\log\log(n)}\big)$ $(i = 1, \ldots, d)$ and $|\widehat{\theta} - \theta| = \mathcal{O}_p\big(1/\sqrt{n}\big)$, we obtain $\|\widehat{e} - e\| = \mathcal{O}_p\big(\sqrt{n^{-1}\log\log(n)}\big)$ and $\|\widehat{c}_{\mathcal{X}} - c_{\mathcal{X}}\| = \mathcal{O}_p\big(\sqrt{n^{-1}\log\log(n)}\big)$. Hence,

$$\widehat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) = \frac{1}{c_{\mathcal{X}}(\boldsymbol{x})}[\widehat{e}(\boldsymbol{x}) - e(\boldsymbol{x})] + \frac{e(\boldsymbol{x})}{c^2_{\mathcal{X}}(\boldsymbol{x})}[\widehat{c}_{\mathcal{X}}(\boldsymbol{x}) - c_{\mathcal{X}}(\boldsymbol{x})] + \mathcal{O}_p\left(\frac{\log\log(n)}{n}\right).$$

By employing similar arguments to that of Theorem 3.2's proof, one can derive the representations (8) and (9) of $\widehat{e}(\boldsymbol{x})$ and $\widehat{c}_{\mathcal{X}}(\boldsymbol{x})$ using the representation of the Kaplan-Meier estimator Γ_n in [15]. This completes the proof. □