1-1-2019

Dependence measure for length-biased survival data using copulas

Rachid Bentoumi
Zayed University

Mhamed Mesfioui
Université du Québec à Trois-Rivières

Mayer Alvo
University of Ottawa, Canada

Research Article **Open Access**

Rachid Bentoumi, Mhamed Mesfioui*, and Mayer Alvo

Dependence measure for length-biased survival data using copulas

https://doi.org/10.1515/demo-2019-0018
Received February 20, 2019; accepted September 19, 2019

Abstract: The linear correlation coefficient of Bravais-Pearson is considered a powerful indicator when the dependency relationship is linear and the error variate is normally distributed. Unfortunately in finance and in survival analysis the dependency relationship may not be linear. In such case, the use of rank-based measures of dependence, like Kendall's tau or Spearman rho are recommended. In this direction, under length-biased sampling, measures of the degree of dependence between the survival time and the covariates appear to have not received much intention in the literature. Our goal in this paper, is to provide an alternative indicator of dependence measure, based on the concept of information gain, using the parametric copulas. In particular, the extension of the Kent's [18] dependence measure to length-biased survival data is proposed. The performance of the proposed method is demonstrated through simulations studies.

Keywords: Length-biased sampling, covariate distribution, length-biased distribution, information gain, dependence measure, kernel density estimation, copulas

MSC: 62F40, 62F12, 62G07, 62H05, 62H12, 60H20

1 Introduction

Survival data occur in many areas such as medicine, epidemiology, biology, economics and manufacturing. The principal goal in survival analysis is the study of the occurrence of a specific event. Most of the literature on length-biased sampled data concentrates on statistical methods for the survival function (e.g., [7]; [32], estimating the density function (e.g., [4]; [17]), kernel smoothing [33], proportional hazards models [35] and covariate bias induced by length-biased sampling of failure times (e.g., [3]). The phenomena of length-biased sampling appears naturally in many areas of research, see for instance [24] in land economics, [36] in screening and early detection of disease, [34] in epidemiology and geriatric medicine. There are many situations where length-biased data arise without censoring, for example quality control problems for estimating fiber length distribution [7], shopping center sampling and mall intercept surveys [25]. For further examples of length-biased sampling see for example [26].

The analogue of Kent's measure for length-biased survival data (see, e.g., [18]) has not received much attention in the literature. In this context, for example, it is of interest to know if there exists any correlation between survival times with dementia and associated covariates such as age at onset, sex and years of education. In this sense, for more general regression models used in survival analysis a measure of dependence can be defined using the concept of information gain (see, e.g., [18]; [19]). This concept generalizes more common measures such as the multiple correlation coefficient. The purpose of this paper is to extend the dependence

Rachid Bentoumi: Department of Mathematics and Statistics, Zayed University,P.O. Box 144534, Abu Dhabi, United Arab Emirates
***Corresponding Author: Mhamed Mesfioui:** Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières, C.P. 500, Trois-Rivières (Québec) Canada G9A 5H7, E-mail: mhamed.mesfioui@uqtr.ca
Mayer Alvo: Department of Mathematics and Statistics, University of Ottawa, STEM Complex, room 336, 150 Louis-Pasteur Pvt Ottawa, ON, Canada K1N 6N5

measure of [18] under length-biased sampling. More specifically, we propose a new measure of dependence between survival time and one continuous covariate without censoring. The main idea consists in expressing the extended dependence measure in terms of the underlying copula under length-biased sampling.

The remainder of the paper is organized as follows: In Section 2, we introduce notations and present some preliminaries. In Section 3, we derive the dependence measure for length-biased data without censoring for the case of one continuous covariate. We develop an estimation procedure for the proposed measure based on parametric copulas and bootstrap technique. Section 4 presents a simulation study allowing to investigate the performance of the proposed method.

2 Notations and preliminaries

In this section, we first introduce the concept of information gain and then, under length-biased sampling, we review distributions for length-biased data and we expose some general notions of copulas.

2.1 Concept of information gain

Let (X, Y) be a random vector with true joint density $g(x, y)$ modelled by a parametric family $\{f(x, y; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_1\}$. Suppose that X and Y are modelled as independent random variables under $\Theta_0 \subset \Theta_1$. For the comparison between the best fitting models under Θ_0 and Θ_1, [18] used Fraser information [10] to extend the work of [21] and provided the joint information gain to be

$$\Gamma = 2\left\{\Phi(\boldsymbol{\theta}_1) - \Phi(\boldsymbol{\theta}_0)\right\},$$

where $\Phi(\boldsymbol{\theta}) = \int\int \log\left\{f(x, y; \boldsymbol{\theta})\right\} g(x, y)dxdy$ and $\boldsymbol{\theta}_i$ maximizes $\Phi(\boldsymbol{\theta})$ over Θ_i. As information gain increases, the model under Θ_1 gets closer to the true density $g(x, y)$ compared with the model under Θ_0. [18] proposed

$$\rho_J^2(X, Y) = 1 - \exp\{-\Gamma\},$$

as a measure of dependence between X and Y. On the other hand, if X is modelled conditionally on Y by a parametric family $\{f(x|y; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_1\}$, [18] used conditional Fraser information [10] on the expected conditional log-likelihood $\Phi_C(\boldsymbol{\theta}) = \int\int \log\left\{f(x|y; \boldsymbol{\theta})\right\} g(x, y)dxdy$ in order to adapt the joint information gain to a conditional information gain, defined as

$$\Gamma_C = 2\left\{\Phi_C(\boldsymbol{\theta}_1) - \Phi_C(\boldsymbol{\theta}_0)\right\},$$

and the conditional dependence measure of [18] is

$$\rho_C^2(X|Y) = 1 - \exp\{-\Gamma_C\}.$$

Note that, if $g(x, y) = f(x, y; \boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^* \in \Theta_1$, then the information gain with respect to [18] reduces to twice the [20] information gain (see, e.g., [18]; [19]). When the concept of information is used, we need to assume that $\Gamma < \infty$ ($\Phi(\boldsymbol{\theta}_i) < \infty$) and $\Gamma_C < \infty$ ($\Phi_C(\boldsymbol{\theta}_i) < \infty$). Furthermore, since $\Theta_0 \subset \Theta_1$, Γ and Γ_C are always nonnegative. The measures ρ_J^2 and ρ_C^2 have the following properties (see, [18]):

- if X and Y are two independent random variables, then $\rho_J^2 = 0$ ($\rho_C^2 = 0$ in conditional models);
- $0 \le \rho_J^2 < 1$ in continuous models. This is also true for ρ_C^2;
- under normal models, ρ_J^2 reduces to the product-moment correlation and ρ_C^2 is the squared multiple correlation coefficient.

2.2 Length-biased sampling and Length-biased distributions

Length-biased sampling occurs when one naturally collects samples from a given population, but the sampling distribution is different from the target population. It such case, not every unit in the population has

an equal chance to be sampled when the natural sampling plan is adopted. For example, suppose in a boy school, data are collected on the number of brothers and sisters in the family of each boy in this school. Since this is a boy school and each family has at least one boy, the collected data are clearly a biased representation of the target population. We will give examples of length-biased distribution derived from the discrete and continuous distributions. To do this, let X be a discrete random variable representing the size of some group from a target population with probability mass function

$$f(k) = P(X = k), \quad k = 1, 2, \ldots.$$

Suppose that a group from this target population is observed only when at least one of the individuals in the group is sighted and each individual has an independent probability p of being observed. From [27] the probability that the observed group has $X = k$ individuals is given as

$$P(\text{A group is sighted}|\text{group size} = k) = 1 - (1 - p)^k =: w(k).$$

The distribution of the observed group size is

$$P(X = k|\text{A group is sighted }) = \frac{w(k)f(k)}{\sum_{k=1}^{\infty} w(k)f(k)} =: f^w(k).$$

If $p \to 0$, then $w(k) \approx kp$. Consequently,

$$f^w(k) \to \frac{kf(k)}{\sum_{k=1}^{\infty} kf(k)}.$$

This distribution is called the length-biased distribution derived from $f(k)$.

Next we give an example of length-biased distribution derived from a continuous distribution, (see, [1]). Let U be a continuous random variable taking values in $(0, c)$, with density function $f_U(u)$. Let T be the left truncation time, with density function $g(t)$, and independent of U. Suppose that a unit U of size u in the population is recorded only if $U > T$. Then, the joint density of (U, T) given $U \geq T$ can be expressed as

$$f_{U,T}(u, t|U \geq T) = \frac{f_{U,T}(u, t)}{\mathbb{P}(U \geq T)} = \frac{f_U(u)\, g(t)}{\mathbb{P}(U \geq T)},$$

if $U \geq T$ and 0 otherwise. Now,

$$\mathbb{P}(U \geq T) = \int_0^{\infty} \mathbb{P}(U \geq t|T = t)\, g(t)dt = \int_0^{\infty} S_U(t)g(t)dt.$$

If the onset times follow a stationary Poisson process, the truncation times are uniformly distributed over the interval $(0, c)$ and $\mathbb{P}(U \geq c) = 0$, see [35]. It follows that

$$\mathbb{P}(U \geq T) = \frac{\mu}{c},$$

where μ is the mean failure time. Therefore,

$$f_{U,T}(u, t|U \geq T) = \frac{f_U(u)}{\mu}.$$

The density function of U conditional on $U \geq T$ is then

$$f(u|U \geq T) = \int_0^{u} f_{U,T}(u, t|U \geq T)dt = \int_0^{u} \frac{f_U(u)}{\mu} dt = \frac{uf_U(u)}{\mu}.$$

Note that, $f(u|U \geq T)$ is a length-biased density derived from $f_U(u)$.

[7] discussed several procedures used in sampling of textile fibres. One procedure is called length-biased and occurs when the chance of selection is proportional to fibre length. From [7], the length-biased density of a positive random variable (r.v) U, which denotes the failure lifetime or survival time, is defined by

$$f_{LB}(u) = \frac{uf_U(u)}{\mu}, \tag{1}$$

where $f_U(u)$ is the unbiased density and $\mu = \int uf_U(u)du < \infty$. According to (1), we define the length-biased density of U conditional on the covariate $Z = z$ as

$$f_{LB}(u|z) = \frac{uf_U(u|z)}{\mu(z)}, \tag{2}$$

where $\mu(z) = \int uf_U(u|z)du < \infty$ and $f_U(u|z)$ denotes the unbiased density corresponding to $f_{LB}(u|z)$. Under length-biased sampling, the covariate associated with the survival time follows a biased density

$$f_B(z) = \frac{\mu(z)f_Z(z)}{\mu}, \tag{3}$$

where $f_Z(z)$ is the unbiased density of the covariate (see, [3]).

2.3 Some general notions of copulas and goodness-of-fit procedures

In this section, we recall some basic definitions and properties of copulas. Also, we provide some examples of parametric copulas and we discuss goodness-of-fit procedures.

In several research areas such as finance, medicine and biology, researchers are constantly striving to understand the dependence structure between two or more random variables. The relationship is described by the joint cumulative distribution function (CDF). However, determining this joint CDF can be a very tedious task. The concept of copulas is an innovative tool for modeling this dependence structure. Indeed, the knowledge of this concept is essential to understanding many areas of application in particular, survival analysis. Thus, whenever it is necessary to model the dependence structure, we can use the copulas.

Let H be a joint distribution function of a random pair (X, Y) and let F and G be the marginal distributions of X and Y, respectively. The copula C is simply the distribution corresponding to the random vector (U, V) with uniform margins defined by $U = F(X) \sim \mathcal{U}_{[0,1]}$ and $V = G(Y) \sim \mathcal{U}_{[0,1]}$. Note that [31] provides an important link between the joint CDF H, the marginal distributions F and G, and copula C described by the following representation

$$H(x, y) = C(F(x), G(y)), \quad \forall (x, y) \in \mathbb{R}^2. \tag{4}$$

If F and G are continuous then C is unique; otherwise, C is uniquely determined on $\mathrm{Ran}F \times \mathrm{Ran}G$, where $\mathrm{Ran}F$ is the range of F. Moreover, if a copula C is twice differentiable then it admits a density defined by

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}. \tag{5}$$

From Sklar's theorem [31] with representation in (4), we can see that the copula C is independent of the marginal distributions. In addition, C is considered as the dependence function associated to the random vector (X, Y). In practice, Sklar's Theorem is very interesting because it models F, G and the dependence structure separately. The following two examples illustrate some applications of this theorem.

Example 2.1. *(Construction of bivariate distribution): Consider the following copula which is given in [23]*

$$C(u, v) = \frac{uv}{u + v - uv}.$$

If the marginal distributions of the random variables X and Y are given by $F(x) = G(x) = 1 - e^{-x}$, $x \geq 0$ *then from (4), we get the next joint distribution of the random vector* (X, Y)

$$H(x, y) = C\left(1 - e^{-x}, 1 - e^{-y}\right) = \left(\frac{1}{1 - e^{-x}} + \frac{1}{1 - e^{-y}} - 1\right)^{-1}.$$

Example 2.2. *(Extraction of copula from a given joint distribution). Let $H_\theta(x, y)$ be the joint distribution function of Gumbel's bivariate exponential distribution [16] given by*

$$H_\theta(x, y) = \begin{cases} 1 - e^{-x} - e^{-y} - e^{-(x+y+\theta xy)}, & x, y > 0, \\ 0, & \text{otherwise}. \end{cases}$$

where θ is a parameter in $[0, 1]$. Clearly, the marginals are exponentially distributed: $F(x) = H(x, \infty) = 1 - e^{-x}$, $G(y) = H(\infty, y) = 1 - e^{-y}$, with inverses $F^{-1}(u) = -\ln\{1 - u\}$, $G^{-1}(v) = -\ln\{1 - v\}$, $u, v \in [0, 1]$. Hence the corresponding copula is

$$C_\theta(u, v) = H_\theta(F^{-1}(u), G^{-1}(v)) = u + v - 1 + (1 - u)(1 - v)e^{-\theta \ln\{1-u\} \ln\{1-v\}}.$$

The parameter $\theta \in [0, 1]$ of the copula C_θ can be viewed as a dependence parameter.

An important property of copulas comes from the fact that for strictly monotone transformations of the random variables, copulas are invariant. In other words, If f and g are strictly increasing transformations on Ran$X$ and Ran$Y$, respectively, then the random vectors (X, Y) and $(f(X), g(Y))$ have the same copula.

Next, we discuss an important class of copulas known as Archimedean copulas defined in [11]. In fact, these copulas find a wide range of applications, in practice, for number of reasons: the ease with which they can be constructed; the great variety of families of copulas which belong to this class; the many nice properties possessed by the members of this class. Furthermore, the dependency structure depends on a single parameter of the generator function ϕ defined below. Formally, an Archimedean copulas is defined by the relation

$$C_\phi(u, v) = \phi^{-1}\{\phi(u) + \phi(v)\}, \quad u, v \in [0, 1], \tag{6}$$

where ϕ denotes a continuous, strictly decreasing convex function defined from $[0, 1]$ to $[0, \infty)$ such that $\phi(1) = 0$. The function ϕ^{-1} represents the inverse of ϕ. The mapping ϕ is so-called the generator of the copula C_ϕ. In what follows, we present some relevant examples of copulas widely used in practice. These parametric copulas will be utilized, in simulation part, to illustrate the proposed estimation method.

- (Clayton copula). The family of Clayton copulas are expressed by

$$C_\theta(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}} \quad \text{for} \quad \theta \in [-1, \infty)\backslash\{0\}, \tag{7}$$

  where the generator of this family is given by $\phi_\theta(t) = \theta^{-1}t^{-\theta} - 1$.
- (Frank's copula). The analytic expression of Frank's copula is

$$C_\theta(u, v) = -\frac{1}{\theta} \ln\left\{1 - \frac{(1 - e^{-\theta u})(1 - e^{-\theta v})}{1 - e^{-\theta}}\right\} \quad \text{for} \quad \theta \in \mathbb{R}\backslash\{0\}, \tag{8}$$

  where the generator of this family is $\phi_\theta(t) = -\ln\left\{(1 - e^{-\theta})(1 - e^{-\theta t})^{-1}\right\}$.
- (Gumbel's copula). Gumbel's copula is formulated as

$$C_\theta(u, v) = \exp\left\{-\left((-\log\{u\})^\theta + (-\log\{v\})^\theta\right)^{\frac{1}{\theta}}\right\} \quad \text{for} \quad \theta \in [1, \infty[, \tag{9}$$

  with generator $\phi_\theta(t) = (-\log\{t\})^\theta$.

Let us now discuss goodness-of-fit (GOF) procedures for copula. The concept of copulas, particulary Archimedean copulas, is frequently used as a good tool for describing the dependence between two random variables X and Y with continuous marginal distributions F and G,respectively. Given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with joint CDF

$$H(x, y) = C(F(x), G(y)), \quad \forall (x, y) \in \mathbb{R}^2,$$

the most frequent question is which copula family is associated with $H(x, y)$? The GOF procedures for copula, which we explain below, can be considered as a good practical answer to this question.

Consider a continuous random vector (X, Y) with margins F, G and bivariate CDF H. Assume further that the copula C of (X, Y) belongs to a class of parametric copula $C_0 = \{C_\theta, \theta \in \Theta\}$, where Θ is the parameter space. Let (X_i, Y_i), $i = 1, \ldots, n$ denote independent copies of (X, Y). Suppose one wants to choose between the null and alternative hypotheses of belonging or not to a given parametric family, namely

$$\mathcal{H}_0 : C \in C_0 \quad \text{versus} \quad \mathcal{H}_1 : C \notin C_0. \tag{10}$$

Several goodness-of-fit procedures allowing to test $\mathcal{H}_0$ versus $\mathcal{H}_1$ have been developed in the literature, e.g. [12], [29], [5], [8], [13], [28], [22], [15] and [2].

The formal GOF tests are rank-based. In other words, instead of using the observations (X_i, Y_i), $i = 1, \ldots, n$, one considers the pseudo-observations $\mathbf{U}_i = (U_{i1}, U_{i2}) = \big(R_i/(n+1), S_i/(n+1)\big)$, $i = 1, \ldots, n$, where $R_i = nF_n(X_i)$ is the rank of X_i among $X_1, \ldots, X_n$ and $S_i = nG_n(Y_i)$ is the rank of Y_i among $Y_1, \ldots, Y_n$. Here, F_n and G_n denote empirical CDF of X and Y, respectively. Note that, the pseudo-observations can be expressed as

$$\mathbf{U}_i = \left(\frac{n}{n+1}F_n(X_i), \frac{n}{n+1}G_n(Y_i)\right) \quad \text{for} \quad i = 1, \ldots, n, \tag{11}$$

and considered as a sample from the copula C. In addition, they are not mutually independent and their components are only approximately uniform on $(0, 1)$. We note that, the factor $n/(n+1)$ in (11) is introduced to avoid problems with C_θ blowing up at the boundary $[0, 1]^2$. The idea behind using the pseudo-observations is that the copula C of a random vector is invariant by continuous, strictly increasing transformations of its components.

The study of some GOF tests for copula and their implementation using the copula package leads to describe one method which is very useful in survival analysis. This approach gave the best results overall, as mentioned by [15] and later, [2] confirmed this remark resulting from examination and comparison of several GOF tests. In what follows, we describe a copula GOF based on the empirical copula:

For testing $\mathcal{H}_0 : C \in C_0$, [15] used the pseudo-observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$ and proposed to work with a consistent estimation of an unknown copula C. In particular, the empirical copula

$$C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(U_{i1} \le u_1, U_{i2} \le u_2\right), \quad \mathbf{u} = (u_1, u_2) \in [0, 1]^2. \tag{12}$$

[9] showed under various conditions that C_n is a consistent estimator of the true underlying copula C. The idea in this approach is to compare C_n with an estimator of C under $\mathcal{H}_0 : C \in C_0$. In a goodness-of-fit setting, [15] suggested to use the empirical process

$$\mathbb{C}_n = \sqrt{n}\left(C_n - C_{\hat{\theta}_n}\right), \tag{13}$$

where $\hat{\theta}_n = Y_n(\mathbf{U}_1, \ldots, \mathbf{U}_n)$ is an estimator of θ. From Genest et al. (2009), a Cramér-von Mises statistic for this approach is

$$S_n^{(E)} = \int_{[0,1]^2} \mathbb{C}_n(\mathbf{u})^2 \, d\mathbb{C}_n(\mathbf{u}) = \sum_{i=1}^{n}\left\{C_n(\mathbf{U}_i) - C_{\hat{\theta}_n}(\mathbf{U}_i)\right\}^2. \tag{14}$$

[14] established the convergence of (13), and showed that the test based on $S_n^{(E)}$ is consistent. A specific parametric bootstrap procedure, developed by [15] can be used to approximate the P-value for this statistic. The validity of this method has been shown in [14].

3 Information gain under length-biased sampling based on copulas

The objective of this section is to exploit the concept of information gain, based on the parametric copulas method, to derive the joint and conditional dependence measures, under length-biased sampling without censoring for the case of one continuous covariate and provide an estimation method for these measures. To do this, let U denote length-biased survival time with CDF $G_{LB}(u, \boldsymbol{\lambda})$ and PDF $g_{LB}(u, \boldsymbol{\lambda})$ while Z represents a continuous covariate with CDF $F_B(z, \boldsymbol{\psi})$ and PDF $f_B(z, \boldsymbol{\psi})$. Suppose that the random vector (U, Z) has a parametric copula C_α. Using Sklar's Theorem, a joint length-biased CDF of (U, Z) is

$$F_{LB}(u, z; \boldsymbol{\theta}) = C_\alpha \left(G_{LB}(u; \boldsymbol{\lambda}), F_B(z; \boldsymbol{\psi}) \right), \tag{15}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\lambda}, \boldsymbol{\psi})$ is the parameter vector of the model. The corresponding joint length-biased density of (U, Z) is given as

$$f_{LB}(u, z; \boldsymbol{\theta}) = c_\alpha \left(G_{LB}(u; \boldsymbol{\lambda}), F_B(z; \boldsymbol{\psi}) \right) g_{LB}(u; \boldsymbol{\lambda}) f_B(z; \boldsymbol{\psi}), \tag{16}$$

where c_α is the parametric copula density given in (5). Consequently, the conditional density of U given $Z = z$ can be expressed in terms of the parametric copula density as

$$g_{LB}(u|z; \boldsymbol{\theta}) = c_\alpha \left(G_{LB}(u; \boldsymbol{\lambda}), F_B(z; \boldsymbol{\psi}) \right) g_{LB}(u; \boldsymbol{\lambda}). \tag{17}$$

The most copula families C_α, $\alpha \in \Theta$, contain the independence copula, that is, C_{α_0} coincides with the independence copula for some $\alpha_0 \in \Theta$. This means that the r.v.'s U and Z are independent which implies that $f_B(z; \boldsymbol{\psi}_0) = f_Z(z; \boldsymbol{\psi}_0)$ and $F_B(z; \boldsymbol{\psi}_0) = F_Z(z; \boldsymbol{\psi}_0)$, where $F_Z(z; \boldsymbol{\psi}_0)$ and $f_Z(z; \boldsymbol{\psi}_0)$ are, respectively, CDF and PDF of the unbiased covariate under the independence model. Therefore, if the covariate sample from the incident cases is available, one can estimate $\boldsymbol{\psi}_0$ by the MLE $\hat{\boldsymbol{\psi}}_0$. In this case, the parameter of the independence model becomes $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\lambda}_0)$ which leads to

- $C_{\alpha_0} \left(G_{LB}(u; \boldsymbol{\lambda}_0), F_Z(z; \hat{\boldsymbol{\psi}}_0) \right) = G_{LB}(u; \boldsymbol{\lambda}_0) F_Z(z; \hat{\boldsymbol{\psi}}_0)$.
- $c_{\alpha_0} \left(G_{LB}(u; \boldsymbol{\lambda}_0), F_Z(z; \hat{\boldsymbol{\psi}}_0) \right) = 1$.
- $f_{LB}(u, z; \boldsymbol{\theta}_0) = g_{LB}(u; \boldsymbol{\lambda}_0) f_Z(z; \hat{\boldsymbol{\psi}}_0)$.
- $g_{LB}(u|z; \boldsymbol{\theta}_0) = g_{LB}(u; \boldsymbol{\lambda}_0)$.

When the covariate sample from the incident cases is not available, one can use the bootstrap techniques to obtain a new sample $Z_1^\star, \ldots, Z_n^\star$ following approximately $f_Z(z)$. First, consider a random sample (U_i, Z_i), $i = 1, \ldots, n$, from $f_{LB}(u, z)$. In particular, $\mathcal{U} = (U_1, \ldots, U_n)$ from $f_{LB}(u)$. Then, use the bootstrap techniques with replacement for the original sample $\mathcal{U}$ to obtain a new sample $\mathcal{U}^\star = \left(U_1^\star, \ldots, U_n^\star \right)$ following approximately $f_U(u)$. The idea is that, U_i is chosen to be included in the new sample $\mathcal{U}^\star$ with probability p_i. For $j = 1, \ldots, n$, the probability p_i, $i = 1, \ldots, n$ can be found using (1) as

$$\begin{aligned} p_i &= \mathbb{P}\left(U_j^\star = U_i | U_1, \ldots, U_n \right) \\ &= \hat{\mu} \frac{\mathbb{P}\left(U_j^\star = U_i \right)}{Y_i} \\ &= \hat{\mu} \frac{1/n}{U_i} \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} U_i^{-1} \right)^{-1} \frac{n^{-1}}{U_i}. \\ &= \frac{U_i^{-1}}{\sum_{i=1}^{n} U_i^{-1}}. \end{aligned} \tag{18}$$

Here, $\hat{\mu} = n \left(\sum_{i=1}^{n} U_i^{-1} \right)^{-1}$ is an estimator of μ in (1) (see [7]). Note that, the bootstrap technique described above converges as shown by [7].

Now, from (U_i, Z_i), $i = 1, \ldots, n$, find $\mathcal{Z}^\star = \left(Z_1^\star, \ldots, Z_n^\star\right)$ associated with $\mathcal{U}^\star = \left(U_1^\star, \ldots, U_n^\star\right)$. Therefore, given this new sample $\mathcal{Z}^\star$, one can use the standard kernel density estimator method in order to estimate the unbiased PDF $f_Z(z)$:

$$\hat{f}_Z(z) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(z - Z_i^\star\right), \tag{19}$$

where the function $K_h(s) = h^{-1}K\left(h^{-1}s\right)$, h is the bandwidth of the estimator and $K : \mathbb{R} \to \mathbb{R}$ is defined to be any smooth function satisfying the following assumptions.

Assumptions 3.1. *(a) K is a PDF; (b) K is symmetric; (c) $\int sK(s)ds = 0$; (d) $\|K\|_2^2 = \int K^2(s)ds < \infty$; (e) $\mu_2(K) = \int s^2K(s) < \infty$.*

A practical estimator of the optimal bandwidth was proposed by Silverman (1986) as $\hat{h}_{opt} = 0.9\hat{\sigma}n^{-5}$, where $\hat{\sigma} = \min\left(s, R/1.34\right)$. Here, s and R are the standard deviation and interquartile range of the data, respectively. Note that, the standard normal density is a very useful kernel function satisfying Assumptions 3.1.

3.1 Copula-based modelling of conditional information gain under length-biased sampling

Hereafter, we express the conditional information gain under length-biased sampling in terms of the underlying copula of (U, Z). The resulting formula is used to estimate the conditional measure of dependence.

Proposition 3.2. *Let (U, Z) be a pair of random variables possibly dependent with true density $f_{LB}(u, z; \boldsymbol{\theta}_1)$ given in (16). Under length-biased sampling, the conditional information, based on the parametric copula density, can be expressed as*

$$\begin{aligned}\Gamma_C &= 2\left\{\iint \log\left\{c_{\alpha_1}\left(G_{LB}(u;\boldsymbol{\lambda}_1), F_B(z;\boldsymbol{\psi}_1)\right) g_{LB}(u;\boldsymbol{\lambda}_1)\right\} f_{LB}(u, z;\boldsymbol{\theta}_1)dudz\right.\\ &\quad \left. - \int \log\left\{g_{LB}(u;\boldsymbol{\lambda}_0)\right\} g_{LB}(u;\boldsymbol{\lambda}_1)du\right\}.\end{aligned} \tag{20}$$

Proof. By testing the two hypotheses $H_0 : \alpha = \alpha_0$ versus $H_1 : \alpha \neq \alpha_0$, the twice Kullback-Leibler (1951) information gain is

$$\begin{aligned}\Gamma_C &= 2\left\{\iint \log\left\{g_{LB}(u|z;\boldsymbol{\theta}_1)\right\} f_{LB}(u, z;\boldsymbol{\theta}_1)dudz\right.\\ &\quad \left. - \iint \log\left\{g_{LB}(u|z;\boldsymbol{\theta}_0)\right\} f(u, z;\boldsymbol{\theta}_1)dudz\right\}\\ &= 2\left\{\iint \log\left\{c_{\alpha_1}\left(G_{LB}(u;\boldsymbol{\lambda}_1), F_B(z;\boldsymbol{\psi}_1)\right) g_{LB}(u;\boldsymbol{\lambda}_1)\right\} f_{LB}(u, z;\boldsymbol{\theta}_1)dudz\right.\\ &\quad \left. - \int \log\left\{g_{LB}(u;\boldsymbol{\lambda}_0)\right\} g_{LB}(u;\boldsymbol{\lambda}_1)du\right\},\end{aligned}$$

where $g_{LB}(u|z;\boldsymbol{\theta}_1)$ is given by (17) and we used the fact that under the independence model: $g_{LB}(u|z;\boldsymbol{\theta}_0) = g_{LB}(u;\boldsymbol{\theta}_0) = g_{LB}(u;\boldsymbol{\lambda}_0)$. ■

Consequently, from Proposition 3.2, the conditional dependence measure with respect to the work of [18] is

$$\rho_C^2(U|Z) = 1 - \exp\{-\Gamma_C\}. \tag{21}$$

In order to estimate the conditional information gain and conditional dependence measure, let (U_i, Z_i), $i = 1, \ldots, n$ be a random sample from $f_{LB}(u, z; \theta_1)$ given in (16). Based on Proposition 3.2, the conditional information gain can be formulated as

$$\Gamma_C = 2\left\{\mathrm{E}\left[\log\left\{c_{\alpha_1}\left(G_{LB}(U;\boldsymbol{\lambda}_1), F_B(Z;\boldsymbol{\psi}_1)\right) g_{LB}(U;\boldsymbol{\lambda}_1)\right\}\right] - \mathrm{E}\left[\log\left\{g_{LB}(U;\boldsymbol{\lambda}_0)\right\}\right]\right\}. \tag{22}$$

An estimator of Γ_C is

$$\begin{aligned}\hat{\Gamma}_C &= \frac{2}{n}\left\{\sum_{i=1}^{n}\log\left\{c_{\hat{\alpha}_1}\left(G_{LB}\left(U_i;\hat{\boldsymbol{\lambda}}_1\right),F_B\left(Z_i;\hat{\boldsymbol{\psi}}_1\right)\right)g_{LB}\left(U_i;\hat{\boldsymbol{\lambda}}_1\right)\right\}\right.\\ &\quad\left.-\sum_{i=1}^{n}\log\left\{g_{LB}\left(U_i;\hat{\boldsymbol{\lambda}}_0\right)\right\}\right\},\end{aligned} \tag{23}$$

where $\hat{\boldsymbol{\theta}}_1 = \left(\hat{\alpha}_1,\hat{\boldsymbol{\lambda}}_1,\hat{\boldsymbol{\psi}}_1\right)$ and $\hat{\boldsymbol{\theta}}_0 = \left(\alpha_0,\hat{\boldsymbol{\lambda}}_0\right)$ are the parameter values that maximize, respectively, the observed log-likelihood

$$\sum_{i=1}^{n}\log\left\{c_{\alpha_1}\left(G_{LB}(U_i;\boldsymbol{\lambda}_1),F_B(Z_i;\boldsymbol{\psi}_1)\right)g_{LB}(U_i;\boldsymbol{\lambda}_1)\right\},$$

and

$$\sum_{i=1}^{n}\log\left\{g_{LB}\left(U_i;\boldsymbol{\lambda}_0\right)\right\}.$$

Therefore, an estimator of the conditional measure of dependence is then

$$\hat{\rho}_C^2\left(U|Z\right) = 1-\exp\left(-\hat{\Gamma}_C\right), \tag{24}$$

where $\hat{\Gamma}_C$ is given by (23).

3.2 Copula-based modelling of joint information gain under length-biased sampling

In this section, we provide a new way to estimate the joint information gain under length-biased sampling. To this end, we first establish an expression of the twice [20] information gain in terms of parametric copula density.

Proposition 3.3. *Let (U,Z) be a pair of random variables possibly dependent with true density $f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)$ given in (16). Under length-biased sampling, the joint information gain, based on the parametric copula density, is*

$$\begin{aligned}\Gamma &= 2\left\{\iint\log\left\{c_{\alpha_1}\left(G_{LB}(u;\boldsymbol{\lambda}_1),F_B(z;\boldsymbol{\psi}_1)\right)g_{LB}(u;\boldsymbol{\lambda}_1)f_B(z;\boldsymbol{\psi}_1)\right\}f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)dudz\right.\\ &\quad\left.-\iint\log\left\{g_{LB}\left(u;\boldsymbol{\lambda}_0\right)f_Z(z;\hat{\boldsymbol{\psi}}_0)\right\}f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)dudz\right\}.\end{aligned} \tag{25}$$

Proof. The twice [20] information gain, by testing $H_0:\alpha=\alpha_0$ versus $H_1:\alpha\neq\alpha_0$, would be

$$\begin{aligned}\Gamma &= 2\left\{\iint\log\left\{f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)\right\}f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)dudz\right.\\ &\quad\left.-\iint\log\left\{f_{LB}\left(u,z;\boldsymbol{\theta}_0\right)\right\}f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)dudz\right\}\\ &= 2\left\{\iint\log\left\{c_{\alpha_1}\left(G_{LB}(u;\boldsymbol{\lambda}_1),F_B(z;\boldsymbol{\psi}_1)\right)g_{LB}(u;\boldsymbol{\lambda}_1)f_B(z;\boldsymbol{\psi}_1)\right\}f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)dudz\right.\\ &\quad\left.-\iint\log\left\{g_{LB}\left(u;\boldsymbol{\lambda}_0\right)f_Z(z;\hat{\boldsymbol{\psi}}_0)\right\}f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)dudz\right\},\end{aligned}$$

where $f_{LB}\left(u,z;\boldsymbol{\theta}_1\right)$ is given by (16) and we used the fact that under the independence model: $f_{LB}\left(u,z;\boldsymbol{\theta}_0\right) = g_{LB}\left(u;\boldsymbol{\lambda}_0\right)f_Z\left(z;\boldsymbol{\psi}_0\right)$. ■

From Proposition 3.3, the joint dependence measure with respect to the work of Kent (1983), is

$$\rho_f^2\left(U,Z\right) = 1-\exp\left\{-\Gamma\right\}. \tag{26}$$

It can be shown that from Proposition 3.3, one can write

$$\Gamma = \Gamma_C + \Gamma_B, \tag{27}$$

where Γ_C is given by (20) and Γ_B is expressed by

$$2\left\{\int \log\left\{f_B(z;\boldsymbol{\psi}_1)\right\} f_B(z;\boldsymbol{\psi}_1)dz - \int \log\left\{f_Z(z;\hat{\boldsymbol{\psi}}_0)\right\} f_B(z;\boldsymbol{\psi}_1)dz\right\} \tag{28}$$

is the information gain obtained through knowledge of the bias of covariate.

In order to estimate the conditional information gain and conditional dependence measure, let (U_i, Z_i), $i = 1, \ldots, n$ be a random sample from $f_{LB}(u, z; \theta_1)$ given in (16). There exist two ways for estimating the joint information. The first method is based on (25). An estimator of Γ is

$$\begin{aligned}\hat{\Gamma} \;=\; & 2\left\{\frac{1}{n}\sum_{i=1}^{n}\log\left\{c_{\hat{\alpha}_1}\left(G_{LB}\left(U_i;\hat{\boldsymbol{\lambda}}_1\right), F_B\left(Z_i;\hat{\boldsymbol{\psi}}_1\right)\right) g_{LB}\left(U_i;\hat{\boldsymbol{\lambda}}_1\right) f_B\left(Z_i;\hat{\boldsymbol{\psi}}_1\right)\right\}\right. \\ & \left. -\frac{1}{n}\sum_{i=1}^{n}\log\left\{g_{LB}\left(U_i;\hat{\boldsymbol{\lambda}}_0\right) f_Z\left(Z_i;\hat{\boldsymbol{\psi}}_0\right)\right\}\right\},\end{aligned} \tag{29}$$

where $\hat{\boldsymbol{\theta}}_1 = \left(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\psi}}_1\right)$ and $\hat{\boldsymbol{\theta}}_0 = \left(\alpha_0, \hat{\boldsymbol{\lambda}}_0\right)$ are the parameter values that maximize the observed log-likelihood, respectively,

$$\sum_{i=1}^{n}\log\left\{c_{\alpha_1}\left(G_{LB}(U_i;\boldsymbol{\lambda}_1), F_B(Z_i;\boldsymbol{\psi}_1)\right) g_{LB}(U_i;\boldsymbol{\lambda}_1) f_B(Z_i;\boldsymbol{\psi}_1)\right\},$$

and

$$\sum_{i=1}^{n}\log\left\{g_{LB}\left(U_i;\boldsymbol{\lambda}_0\right) f_Z\left(Z_i;\hat{\boldsymbol{\psi}}_0\right)\right\}.$$

The second method is based on (27). In this direction, an estimator of the joint information gain is

$$\hat{\Gamma} = \hat{\Gamma}_C + \hat{\Gamma}_B, \tag{30}$$

where $\hat{\Gamma}_C$ is given by (23) and the estimator of Γ_B is

$$\hat{\Gamma}_B = \frac{2}{n}\left\{\sum_{i=1}^{n}\log\left\{f_B\left(Z_i;\hat{\boldsymbol{\psi}}_1\right)\right\} - \sum_{i=1}^{n}\log\left\{f_Z\left(Z_i;\hat{\boldsymbol{\psi}}_0\right)\right\}\right\}. \tag{31}$$

Hence, an estimator of the joint measure of dependence is

$$\hat{\rho}_J^2(U, Z) = 1 - \exp\left\{-\left(\hat{\Gamma}_C + \hat{\Gamma}_B\right)\right\}. \tag{32}$$

We note that, in the case where the covariate sample from the incident cases is not available, a natural estimator of the unbiased density of the covariate, f_Z, is given by (19) as

$$\hat{f}_Z(z) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(z - Z_i^{\star}\right).$$

4 Simulation study

In this section, we develop some useful simulation algorithms in order to simulate length-biased survival times with one continuous covariate using parametric copula method. Also, we investigate the performance of this method by providing some results of several simulations assessing the behaviour of the estimated information gain and dependence measure given length-biased data.

4.1 Simulation algorithms

The following algorithm allows us to simulate length-biased survival times from the length-biased distribution, if the CDF G_{LB} and its inverse G_{LB}^{-1} admit a closed form.

Algorithm 4.1.
For $i = 1, \ldots, n$
1. $W_i \sim U(0, 1)$.
2. $U_i = G_{LB}^{-1}(W_i)$.

Often, it is difficult to simulate length-biased data directly from length-biased distribution because in general the CDF $G_{LB}(u)$ and its inverse $G_{LB}^{-1}(u)$ may not have a closed form. In this case, we can use the following algorithm which is based on the bootstrap techniques.

Algorithm 4.2.
1. Simulate a large sample $U_1^\star, \ldots, U_N^\star$ from a given unbiased density $f_U(u)$.
2. For $i = 1, \ldots, n$, use the bootstrap techniques from the original sample $U_1^\star, \ldots, U_N^\star$ with probability $p_i = U_i^\star \left(\sum U_i^\star\right)^{-1}$ to obtain a new sample $U_1, \ldots, U_n$ from the length-biased density $g_{LB}(u)$.

We note that the probability p_i, given in the latter algorithm, can be obtained in the same way as for (18).

Now, we develop some useful algorithms for the parametric copula method. On one hand, we develop a method allowing to simulate data from the joint unbiased density through its underlying copula. On the other hand, we derive a practical approach, based on the bootstrap techniques, for simulating length-biased data from the joint length-biased density. Specifically, let $f_U(u; \boldsymbol{\lambda})$ and $F_U(u; \boldsymbol{\lambda})$ denote unbiased PDF and unbiased CDF of the continuous r.v. U (survival time). Also, let $f_Z(z; \boldsymbol{\psi})$ and $F_Z(z; \boldsymbol{\psi})$ be unbiased PDF and unbiased CDF of the continuous r.v. Z (covariate). From Theorem 4, the joint unbiased CDF of the random vector (U, Z) can be written as a function of a parametric copula as follows

$$F_U(u, z, \boldsymbol{\theta}) = C_\alpha\left(F_U(u; \boldsymbol{\lambda}), F_Z(z; \boldsymbol{\psi})\right), \quad \forall (u, z) \in \mathbb{R}^2, \tag{33}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\lambda}, \boldsymbol{\psi})$. The joint unbiased density of the random vector (U, Z), denoted by $f_U(u, z, \boldsymbol{\theta})$, can be derived from (4) provided $\partial^2 C_\alpha(u, v)/\partial u \partial v$ exists. Algorithm 4.3 can be used to simulate a random sample $\left(U_i^\star, Z_i^\star\right)$, $i = 1, \ldots, N$ from $f_U(u, z; \boldsymbol{\theta})$.

Algorithm 4.3.
For $i = 1, \ldots, N$
1. $(V_i, W_i) \sim C_\alpha(v, w)$.
2. $U_i^\star = F_U^{-1}(V_i, \boldsymbol{\lambda})$.
3. $Z_i^\star = F_Z^{-1}(W_i, \boldsymbol{\psi})$.
4. The desired observation from $f_U(u, z; \boldsymbol{\theta})$ is $\left(U_i^\star, Z_i^\star\right)$.

Note that, Algorithm 4.3 allows us to simulate a random sample $\left(U_i^\star, Z_i^\star\right)$ $i = 1, \ldots, N$ directly from the joint unbiased density $f_U(u, z; \boldsymbol{\theta})$. However, as we will show, we cannot simulate a random sample (U_i, Z_i) $i = 1, \ldots, n$ directly from the joint length-biased density $f_{LB}(u, z; \boldsymbol{\theta})$. A bootstrap techniques will be proposed as a simple solution for this simulation problem. Using the fact that $\mu(z) = \int_0^\infty u c_\alpha\left(F_U(u), F_Z(z)\right) f_U(u) du$, it follows from (3) that

$$f_B(z) = \frac{\int_0^\infty u c_\alpha\left(F_U(u), F_Z(z)\right) f_U(u) du}{\mu} f_Z(z). \tag{34}$$

The length-biased density of U conditional on $Z = z$ becomes

$$g_{LB}(u|z) = \frac{u f_U(u|z)}{\mu(z)} = \frac{u c_\alpha\left(F_U(u), F_Z(z)\right) f_U(u)}{\int_0^\infty u c_\alpha\left(F_U(u), F_Z(z)\right) f_U(u) du}. \tag{35}$$

Even for a given parametric copula associated with some known unbiased CDF $F_U(u,z)$, equations (34) and (35) cannot be used to simulate, directly, a random sample $(U_1,Z_1),\ldots,(U_n,Z_n)$ from $f_{LB}(u,z)$. Because, in most cases, there is no closed forms of $f_B(z)$, $F_B(z)$, $F_B^{-1}(z)$, $g_{LB}(u|z)$, $G_{LB}(u|z)$ and $G_{LB}^{-1}(u|z)$. Hereafter, we describe an alternative way based on bootstrap techniques enables to simulate length-biased data.

Algorithm 4.4.
For $i=1,\ldots,N$
1. Use Algorithm 4.3 to simulate $\left(U_i^{\star},Z_i^{\star}\right)$ from the joint unbiased density $f_U(u,z)$.
2. Use Algorithm 4.2 to obtain length-biased survival times $U_1,\ldots,U_n$.
3. From $\left(U_i^{\star},Z_i^{\star}\right)$, $i=1\ldots,N$ find $Z_1,\ldots,Z_n$ associated with $U_1,\ldots,U_n$.
4. The desired random sample from $f_{LB}(u,z)$ is $\left(U_j,Z_j\right)$, $j=1\ldots,n$.

If real data is available, then its application becomes very simple using the following suggested algorithm (for the conditional dependence measure):

Algorithm 4.5.
Given real data are (U_i,Z_i) $i=1,\ldots,n$
1. use GOF procedures for copulas to find the parametric copula family C_α associated with this data set;
2. use GOF procedures to find the appropriate distributions of the survival times and covariate, respectively $G_{LB}(u,\boldsymbol{\lambda})$ and $F_B(z,\boldsymbol{\psi})$;
3. find $\hat{\boldsymbol{\theta}}_1=\left(\hat{\alpha}_1,\hat{\boldsymbol{\lambda}}_1,\hat{\boldsymbol{\psi}}_1\right)$ and $\hat{\boldsymbol{\theta}}_0=\left(\alpha_0,\hat{\boldsymbol{\lambda}}_0\right)$ the parameter values that maximize, respectively, the observed log-likelihood
$$\sum_{i=1}^{n}\log\left\{c_{\alpha_1}\left(G_{LB}(U_i;\boldsymbol{\lambda}_1),F_B(Z_i;\boldsymbol{\psi}_1)\right)g_{LB}(U_i;\boldsymbol{\lambda}_1)\right\},$$
and
$$\sum_{i=1}^{n}\log\left\{g_{LB}\left(U_i;\boldsymbol{\lambda}_0\right)\right\}.$$
4. estimate the conditional information and conditional dependence measure using respectively equations (23) and (24).

Note that, a similar algorithm can be developed for the joint dependence measure.

4.2 Simulation study results

Let U be a positive random variable which follows a generalized gamma distribution $GG(r,p,k)$ defined by
$$f(u)=\frac{r}{p\Gamma(k)}\left(\frac{u}{p}\right)^{rk-1}\exp\left\{-\left(\frac{u}{p}\right)^{r}\right\},\quad r,\ p \text{ and } k>0. \tag{36}$$
The latter, includes $Gamma(k,p)$, $Weibull(r,p)$ and $Exp(p)$ by letting in (36), $r=1$, $k=1$, and $r=k=1$, respectively. As shown by [6], if $GG(r,p,k)$ denotes the unbiased density then, its corresponding length-biased density is $GG\left(r,p,k+r^{-1}\right)$. Let $G_{LB}(u)$ denote the length-biased distribution function corresponding to $GG\left(r,p,k+r^{-1}\right)$. Using Algorithm 4.1, we can easily generate length-biased survival times directly from $GG\left(r,p,k+r^{-1}\right)$ given that the corresponding unbiased density is $GG(r,p,k)$.
Figure 1 describes the histogram of the length-biased survival times, $U_1,\ldots,U_n$, obtained from Algorithm 4.4, true length-biased density $GG(r,p,1+r^{-1})$ and corresponding $GG(\hat{r},\hat{p},\hat{k})$. Here, the unbiased density of the covariate , $f_Z(z)$, is $U(0,1)$ and the structure dependence of the joint unbiased CDF $F_U(u,z)$ is assumed to be described by a Clayton copula.

In what follows, we examine the performance of the parametric copula method when the data come from a length biased density. To do this, we suppose that

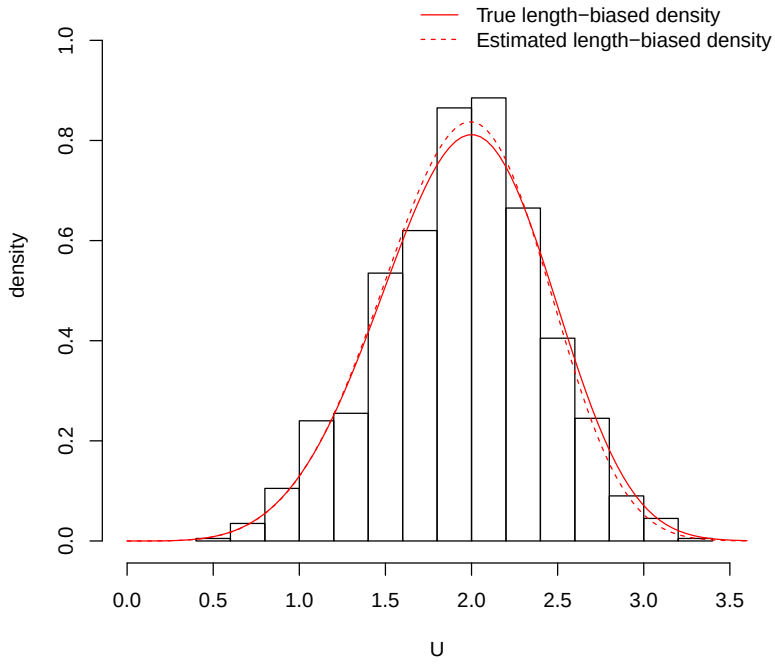

Figure 1: Observed frequencies of the length-biased survival times, true length-biased density $GG(r, p, 1 + r^{-1})$ and $GG(\hat{r}, \hat{p}, \hat{k})$ with $N = 5000$, $n = 1000$, $r = 4$, $p = 2$ and $\alpha = 8$.

- the true unbiased density of the survival time is $Weibull(r, p)$, where $r = 4$ and $p = 2$;
- the true unbiased density of the covariate is $U(0, 1)$;

and consider

- C_α: Clayton copula associated with the joint unbiased CDF $f_U(u, z)$;
- $\boldsymbol{\theta} = (\alpha, r, p)$: parameter of the model;
- $N = 5000$: number of the simulated observations;
- $m = 1000$: number of the simulated samples.

Given length-biased data (U_i, Z_i), $i = 1 \ldots, n$ generated from Algorithm 4.4, the question is which copula family is associated with the joint CDF $F_{LB}(u, z)$? A practical answer to this question is to use the goodness-of-fit procedures for copula to find the parametric copula is associated with that length-biased data. In such a case, we suggest to use the goodness-of-fit statistic computed from the empirical copula processes, $S_n^{(E)}$, given in (14).

Table 1 leads to the conclusion that the test based on $S_n^{(E)}$ confirms that the Clayton family copula associated with the unbiased CDF $F_U(u, z)$ is the same as for the length-biased CDF $F_{LB}(u, z)$, but with different estimated values of dependence parameter, denoted by $\hat{\alpha}_{LB}$, as shown by the Table 2.

Table 1: Percentage of rejection at 5%, based on 1000 replicates, of the null hypothesis of belonging to a given family of copulas with $N = 5000$, $n = m = 1000$, $r = 4$ and $p = 2$.

True copula		Families under H_0				
		Clayton	Frank	Gumbel	Gaussian	Student
Clayton	α	$S_n^{(E)}$	$S_n^{(E)}$	$S_n^{(E)}$	$S_n^{(E)}$	$S_n^{(E)}$
	0.005	4,3	4.4	-	5.4	49.7
	0.5	4.7	99.5	100	88.6	97.5
	2	4.2	100	100	100	100
	10	3.8	100	100	100	100

Table 2: Av. estimated dependence parameters $\hat{\alpha}$ and $\hat{\alpha}_{LB}$, based on 1000 replicates, for Clayton copula associated with the CDF's $F_U(u, z)$ and $F_{LB}(u, z)$, respectively, for $N = 5000$, $n = m = 1000$, $r = 4$ and $p = 2$.

True α	Av. $\hat{\alpha}$	Av. $\hat{\alpha}_{LB}$
0.005	0.0052	0.0048
0.5	0.4996	0.3853
2	1.9980	1.5219
10	10.021	7.5899

Now, based on the next Algorithm, our principal objective is to examine, for different values of α given in Table 2, the behavior of information gain and dependence measure estimators. Recall that, the copula family under length-biased sampling is Clayton copula with dependence parameter, denoted by α_{LB}, and the length-biased density of the survival time $g_{LB}(u)$ is $GG(r, p, k)$, where $k = 1 + r^{-1}$. For simplicity, a simple choice used to estimate the unbiased density $f_Z(z)$ and the biased density $f_B(z)$ is the kernel density estimator as follows

$$\hat{f}_Z(z) = \frac{1}{N}\sum_{i=1}^{N} K_h\left(z - Z_i^{\star}\right),$$

$$\hat{f}_B(z) = \frac{1}{n}\sum_{i=1}^{n} K_h(z - Z_i).$$

Let $\boldsymbol{\theta}_{LB} = (\alpha_{LB}, r, p, k)$ denote the parameter of the model under length-biased sampling.

Algorithm 4.6.

For $k = 1, \ldots, m$ and for $i = 1, \ldots, n$

1. For the conditional model, find $\hat{\boldsymbol{\theta}}_{LB,1} = \left(\hat{\alpha}_{LB}, \hat{r}_1, \hat{p}_1, \hat{k}_1\right)$ and $\hat{\boldsymbol{\theta}}_{LB,0} = \left(\hat{r}_0, \hat{p}_0, \hat{k}_0\right)$ that maximize, respectively, the observed log likelihood

$$\sum_{i=1}^{n} \log\left\{c_{\hat{\alpha}_{LB}}\left(G_{LB}(U_i; \hat{r}_1, \hat{p}_1, \hat{k}_1), F_B(Z_i)\right) g_{LB}(U_i; \hat{r}_1, \hat{p}_1, \hat{k}_1)\right\},$$

and

$$\sum_{i=1}^{n} \log\left\{g_{LB}\left(U_i; \hat{r}_0, \hat{p}_0, \hat{k}_0\right)\right\}.$$

2. For the conditional model, calculate $\hat{\Gamma}_{C,k}$ and $\hat{\rho}^2_{C,k}(U|Z)$, respectively, by (23) and (24).

To estimate the joint dependence measure, Algorithm 4.6 can be used provided $\hat{\Gamma}_B = (2/n)\left\{\sum_{i=1}^{n} \log\left\{\hat{f}_B(Z_i)\right\} - \sum_{i=1}^{n} \log\left\{\hat{f}_Z(Z_i)\right\}\right\}$, $\hat{\Gamma} = \hat{\Gamma}_C + \hat{\Gamma}_B$ and $\hat{\rho}^2_J(U, Z) = 1 - e^{-\hat{\Gamma}}$.

Table 3 indicates the average maximum likelihood estimators of $\boldsymbol{\theta}_{LB}$ under hypotheses H_1 and H_0, using parametric copula method.

Table 3: A.v MLE's for $\boldsymbol{\theta}_{LB}$, using parametric copula method, under hypotheses H_1 and H_0 for $N = 5000$, $n = m = 1000$, $r = 4$ and $p = 2$.

Av. $\hat{\boldsymbol{\theta}}_{LB,1}$				Av. $\hat{\boldsymbol{\theta}}_{LB,0}$		
$\hat{\alpha}_{LB,1}$	$\hat{r}_1$	$\hat{p}_1$	$\hat{k}_1$	$\hat{r}_0$	$\hat{p}_0$	$\hat{k}_0$
0.0021	3.9751	1.9749	1.2788	3.9766	1.9755	1.2776
0.3553	3.9396	1.9597	1.3266	3.9932	1.9799	1.2692
1.4696	3.9430	1.9628	1.3235	3.9811	1.9753	1.2783
7.4200	3.8997	1.931	1.3600	3.9770	1.9742	1.2802

Table 4 exhibits, for different estimated values of dependence parameter $\hat{\alpha}_{LB}$ the average of information gain and dependence measure estimators under length-biased sampling.

Table 4: Av. estimated information gain and dependence measure given simulated length-biased data, using parametric copula method, for $N = 5000$, $n = m = 1000$, $r = 4$ and $p = 2$.

Av. $\hat{\alpha}_{LB}$	Av. $\hat{\Gamma}_C$	Av. $\hat{\Gamma}_B$	Av. $\hat{\Gamma}$	Av. $\hat{\rho}_C^2(U\|Z)$	Av. $\hat{\rho}_J^2(U,Z)$
0.0048	0.0010	0.0028	0.0299	0.0010	0.0295
0.3853	0.0951	0.0145	0.1097	0.0905	0.1036
1.5219	0.6472	0.0215	0.6688	0.4758	0.4871
7.5899	2.5431	0.0426	2.5858	0.9211	0.9245

The most important remark from Table 4 is that the estimated conditional and joint dependence measures are slightly different due to the small values of $\hat{\Gamma}_B$ for all estimated values of $\hat{\alpha}_{LB}$. This can be explained, simply, by the initial choice of the model parameters. In particular, if the shape parameter $r = 0.6$, the parametric copula associated with the CDF $F_{LB}(u,z)$ is always Clayton copula.

Table 5: Percentage of rejection at 5%, based on 1000 replicates, of the null hypothesis of belonging to a given family of copulas for $N = 5000$, $n = m = 1000$, $r = 0.6$ and $p = 2$.

True copula		Families under H_0				
		Clayton	Frank	Gumbel	Gaussian	Student
Clayton	α	$S_n^{(E)}$	$S_n^{(E)}$	$S_n^{(E)}$	$S_n^{(E)}$	$S_n^{(E)}$
	0.005	5	5	-	5.2	49.6
	0.5	5.4	22.7	48.3	17.6	62.5
	2	6.9	99.4	100	97.7	99.7
	10	4	100	100	100	100

Table 6: Av. estimated information gain and dependence measure given simulated length-biased data, using parametric copula method, for $N = 5000$, $n = m = 1000$, $r = 0.6$ and $p = 2$.

Av. $\hat{\alpha}_{LB}$	Av. $\hat{\Gamma}_C$	Av. $\hat{\Gamma}_B$	Av. $\hat{\Gamma}$	Av. $\hat{\rho}_C^2(U\|Z)$	Av. $\hat{\rho}_J^2(U,Z)$
0.0009	0.0009	0.0288	0.0298	0.0009	0.0293
0.1300	0.0145	0.0664	0.0809	0.0143	0.0776
0.5017	0.1410	0.4576	0.5987	0.1312	0.4501
2.5136	1.1282	1.0504	2.1786	0.6757	0.8866

Table 5 indicates that, the new value of the shape $r = 0.6$ influences considerably $\hat{\Gamma}_B$ and $\hat{\Gamma}_C$. Consequently, from Table 6, the difference between estimated conditional and joint dependence measure is very significant and hence we can conclude that given length-biased data we cannot ignore the potential effect of the covariate on the survival time.

5 Discussion

This paper provides a measure of dependence for length-biased survival data, by extending the dependence measure of [18], under length-biased sampling. More specifically, we looked at a measure of dependence between survival time (without censoring) and one continuous covariate. In this direction, we developed parametric copulas method based on information gain. It would be interesting to adapt this approach for several continuous covariates especially under censoring and consider other types of covariates in the model. This can be done using the concept of copulas which takes into account censored data.

Acknowledgments: Mhamed Mesfioui acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada No 06536-2018. Mayer Alvo was supported by the Natural Sciences and Engineering Research Council of Canada (grant OGP0009068).

References

[1] Bentoumi, R. (2017). Measure of Dependence for Length-biased Survival Data. PhD thesis, University of Ottawa, Canada.
[2] Berg, D. (2009). Copula goodness-of-fit testing: an overview and power comparison. *Eur. J. Finance 15*(7-8), 675–701.
[3] Bergeron, P.-J, M. Ashgarian, and D.B. Wolfson (2008). Covariate bias induced by length-biased sampling of failure times. *Amer. Statist. Assoc. 103*(482), 737–742.
[4] Bhattacharyya, B.B., L.A. Franklin, and G.D. Richardson (1988). A comparison of nonparametric unweighted and length-biased density estimation of fibres. *Comm. Statist. Theory Methods 17*(11), 3629–3644.
[5] Breymann, W., A. Dias, and P. Embrechts (2003). Dependence structures for multivariate high-frequency data in finance. *Quant. Finance. 3*(1), 1–14.
[6] Correa J.A. and D.B. Wolfson (1999). Length-bias: some characterizations and applications. *Stat. Comput. Simul. 64*(3), 209–219.
[7] Cox, D.R. (1969). Some sampling problems in technology. In N. L. Johnson and H. Smith (Eds.), *New Developments in Survey Sampling*, pp. 506–527. John Wiley, New York.
[8] Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *J. Multivariate Anal. 95*(1), 119–152.
[9] Fermanian, J.-D., D. Radulovic, and M. Wegkamp (2004). Weak convergence of empirical copula processes. *Bernoulli 10*(5), 847–860.
[10] Fraser, D.A.S. (1965). On information in statistics. *Ann. Math. Statist. 36*(3), 890–896.
[11] Genest, C. and R.J. Mackay (1986). The joy of copulas: bivariate distributions with uniform marginals. *Amer. Statist. 40*(4) 280–283.
[12] Genest, C. and L.-P. Rivest (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Amer. Statist. Assoc. 88*(423), 1034–1043.
[13] Genest, C., J.-F. Quessy, and B. Rémillard (2006). Goodness-of-fit procedures for copula models based on the probability integral transform. *Scand. J. Stat. 33*(2), 337–366.
[14] Genest, C. and B. Rémillard (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Ann. Inst. H. Poincaré Probab. Statist. 44*(2), 1096–1127.
[15] Genest, C., B. Rémillard, and D. Beaudoin (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance Math. Econom. 44*(2), 199–213.
[16] Gumbel E.J. (1960). Bivariate exponential distributions. *J. Amer. Statist. Assoc. 55*(292), 698–707.
[17] Jones, M.C. (1991). Kernel density estimation for length biased data. *Biometrika. 78*(3), 511–519.
[18] Kent, J.T. (1983). Information gain and a general measure of correlation. *Biometrika 70*(1), 163–173.
[19] Kent, J.T. and J. O'Quigley (1988). Measure of dependence for censored survival data. *Biometrika 75*(3), 525–534.
[20] Kullback, S. and R.A. Leibler (1951). On information and sufficiency. *Ann. Math. Statist. 22*(1), 79–86.
[21] Linfoot, E.H. (1957). On informational measure of correlation. *Inform. Control 1*(1), 85–89.

[22] Mesfioui, M., J.-F. Quessy, and M.-H. Toupin (2009). On a new goodness-of-fit process for families of copulas. *Canad. J. Statist. 37*(1), 80–101.
[23] Nelsen, R.B. (2006). *An Introduction to Copulas.* Second edition. Springer, New York.
[24] Nowell, C., M. A. Evans, and L. McDonald (1988). Length-biased sampling in contingent valuation studies. *Land Econon. 64*(4), 367–371.
[25] Nowell, G. and R.S. Linda (1991). Length-biased sampling in mall intercept surveys. *J. Mark. Res. 28*(4), 475–479.
[26] Patil, G.P., and C.R. Rao (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics 34*(2), 179–189.
[27] Qin, J. (2017). *Biased sampling, Over-identified Parameter Problems and Beyond.* Springer, Singapore.
[28] Scaillet, O. (2007). Kernel-based goodness-of-fit tests for copulas with fixed smoothing parameters. *J. Multivariate Anal. 98*(3), 533–543.
[29] Shih, J.H. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika 85*(1), 189–200.
[30] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.
[31] Sklar, A. (1959). Fonction de répartion á *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris. 8*, 229–231
[32] Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika 76*(4), 751–61.
[33] Wand, M.P. and M.C. Jones (1995). *Kernel Smoothing.* Chapman & Hall, London.
[34] Wolfson, C., D.B. Wolfson, M. Asgharian, C.E. M'Lan, T. Østbye, K. Rockwood, and D.B. Hogan (2001). A reevaluation of the duration of survival after the onset of dementia. *N. Engl. J. Med. 344*(15), 1111–1116.
[35] Wang, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika 83*(2), 343–354.
[36] Zelen, M. (1993). Optimal scheduling of examinations for the early detection of disease. *Biometrika 80*(2), 279–293.