Journal of Statistical Modelling: Theory and Applications
Vol. 4, No. 2, 2023, pp. 97-110
Yazd University Press 2023

Research Paper

Estimating Kendall's τ when both times are subject to interval censoring

Fatemeh Ghanbari, Mohammad Hossein Dehghan*
Department of Statistics, Sistan and Baluchestan University, Zahedan, Iran

Received: February 20, 2024/ Revised: June 11, 2024/ Accepted: August 16, 2024

Abstract: A common challenge in working with longitudinal data is dealing with incomplete data. According to the existing studies on the dependence structure of survival times, it is a riveting topic for researchers to estimate survival functions and dependence parameters, especially in biology and medical research. Some researchers have studied the aforementioned subjects with left- or right-truncated or censored data. When the data involves interval censoring, the mentioned issues still need to be solved or modified. In this article, we propose two alternative approaches to the estimation of a dependence parameter and Kendall's τ, given an interesting covariate and interval-censored dataset. More precisely, these approaches include non-parametric and semi-parametric methods to estimate the copula dependence parameter and Kendall's τ, which are evaluated by simulation. Finally, we apply the mentioned approaches to a real-world dataset and copula's goodness-of-fit tests.

Keywords: Generalized Turnbull's estimator, Copula function, Interval-censored data, Semi-parametric estimation.
Mathematics Subject Classification (2010): 62N01, 62N02, 62H20.

1 Introduction

Correlation, one of the most widely used concepts in statistics, can be confusing due to its multiple meanings and statistical interpretations.. The term "correlation" refers to a mutual relationship or association between quantities. In most business, medical, and agricultural fields, it is useful to express the quantity of this relationship. A correlation coefficient is a statistical measure that also describes the association between

*Mohammad Hossein Dehghan: mhdehghan@math.usb.ac.ir

random variables. There are many common correlation coefficients typically used, such as Pearson and Spearman coefficients, and Kendall's τ. The Pearson coefficient is known as the Pearson product-moment correlation coefficientand describes the linear relationship between two variables. Spearman's correlation coefficient can be defined as a special case of Pearson's coefficient, applied to a collection of ranked variables. Of course, unlike Pearson's coefficient, it is not restricted to the linear relationship.

The third correlation coefficient, like Spearman's, is based on the ranks of variables. However, Kendall's τ, proposed by Kendall (1938), differs by considering only the direction of agreement between ranks and ignoring the magnitude of rank differences. Consequently, Kendall's τ is generally more appropriate for discrete data. The formal Kendall's τ is defined by

$$\begin{aligned}\tau_{X,Y} =& Pr\left((X_1 - X_2)(Y_1 - Y_2) > 0\right) - Pr\left((X_1 - X_2)(Y_1 - Y_2) < 0\right)\\ =& E\left(\operatorname{sgn}(X_1 - X_2)\operatorname{sgn}(Y_1 - Y_2)\right),\end{aligned}$$

where (X_2, Y_2) is an independent copy of (X_1, Y_1), and sgn is the sign function. In practice, this measure is estimated by the formula

$$\begin{aligned}\hat{\tau}_{(X,Y)} =& \binom{n}{2}^{-1} \{(\text{the number of concordant pairs})\\ & - (\text{the number of discordant pairs})\},\\ =& \binom{n}{2}^{-1} \sum_{i<j} \operatorname{sgn}((X_i - X_j)(Y_i - Y_j)). \qquad (1)\end{aligned}$$

If the value obtained for a pair is $+1$, the pair is concordant. But, the values -1 and 0 indicate, respectively, that the pair is discordant and the pair is uncorrelated. All the aforementioned parameters are meant to be used for precise data.

In many biomedical applications, the analysis of dependence for bivariate or multivariate survival data with censoring is straightforward. Kendall's τ can be useful as a rank-correlation measure. In fact, unlike Pearson's correlation coefficient, Kendall's τ does not require any knowledge of the parametric shape of the marginal distributions. In addition, its rank-invariance properties make it suitable for measuring the dependence in non-Gaussian lifetime models.

Many estimators of Kendall's τ, with right-censored data, have been proposed and studied by some statisticians. See Lim and Meier (2006); Beaudoin et al. (2007); Wang and Wells (2000); Weier and Basu (1980); Oakes (1982, 2008); Lakhal et al. (2008). Hesieh and Li (2017) studied the estimation of bivariate, left-truncated variables. Also, Betensky and Finkelstein (1999) suggested the estimation of Kendall's τ under interval censoring by using a multiple imputation strategy. Kim (2015) employed a conditional tau statistic to estimate an association of bivariate, interval-censored data; the suggested method performed better in simulation studies compared to Betensky and Finkelstein's multiple imputation method, except in the case of strong associations. Kang and Kim (2021) used the inverse probability censoring weighted (IPCW) method to adjust the effect of inductively dependent censoring and multiple imputation techniques to recover unknown failure times due to interval censoring. According to their simulation studies, the proposed association estimator performs well with moderate sample size.

Recall that a test statistic based on $\hat{\tau}$, the unbiased estimator of τ, which has an asymptotic normal distribution, is known as (Lee, 1990), which is known as $n^{1/2}(\hat{\tau} - \tau)$, and has an asymptotic normal distribution; see Hoeffding (1948). Based on left truncation methods, Tsai (1990) studied the independence of random truncation and failure time, and proposed a hypothesis testing. Martin (2005) proposed a quasi-independence testing for censoring failure time. Derumigny and Fermanian (2019) proposed an estimation which on kernel-based estimation of conditional Kendall's τ.

Various tests of quasi-independence are available for one-sided truncation and for truncation that depends on a measured covariate, but none of them can be applied to more complex truncation schemes. Koziol (1980) proposed a goodness-of-fit test for randomly censored data. Also, Genest et al. (2006) proposed a goodness-of-fit test procedures for copula models based on the probability integral transformation, Jahanshahi et al. (2020) studied goodness-of-fit-tests for the Rayleigh Distribution that could be useful for type II, right-censored data. Still, it is difficult to estimate Kendall's τ and the dependence parameter when both survival times are subject to interval-censored data. Therefore, we propose a simple non-parametric method and a semi-parametric method to estimate Kendall's τ under interval-censored observations.

This paper is organized as follows. In Section 2, we discuss the models of survival under bivariate correlated and interval-censored variables, based on the survival copula. Then, in Section 3, we propose methods for estimating the measure of the dependence parameter under a parametric model. Finally, in Section 4, we present examples with real-world data and the results of simulation studies.

2 Bivariate correlated and censored data modeling

Let $S_1(t_1)$ and $S_2(t_2)$ be the marginal survival functions of T_1 and T_2 longitodinal variables, respectively, where $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ is the joint survival function. The non-parametric estimation of the joint survival function for two correlated variables can be modeled by Sklar's theorem (Sklar, 1959) as follows

$$P(T_1 > t_1, T_2 > t_2) = C_\alpha\big(S_1(t_1), S_2(t_2)\big),$$

where C_α is a couple of joint survival functions with α as the dependence parameter.In the following, we present a brief review of some important families of Archimedean copulas. Here, $u, v \in (0,1)$ and α is the dependence parameter of the copula.
Clayton's family:

$$C_\alpha(u,v) = \max\Big\{\Big(u^{-\alpha} + v^{-\alpha} - 1\Big)^{-1/\alpha}, 0\Big\}; \alpha \in [-1, \infty) \setminus \{0\}.$$

Frank's family:

$$C_\alpha(u,v) = \frac{-1}{\alpha}\log\Big(1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{(e^{-\alpha} - 1)}\Big); \alpha \in \mathbb{R} \setminus \{0\}.$$

Gumbel's family:

$$C_\alpha(u,v) = \exp\Big(- \big((-\log(u))^\alpha + (-\log(v))^\alpha\big)^{1/\alpha}\Big);\ \alpha \in [1, \infty).$$

Remark 2.1. *Let X and Y be continuous random variables with copula C. Then, Kendall's τ is given by*

$$\tau = 4\int_0^1 \int_0^1 C(u,v)dC(u,v)) - 1.$$

For the Kendall's τ for Clayton copula is $\tau = \frac{\alpha}{\alpha+2}$, for Frank copula is $\tau = 1 + \frac{4}{\alpha}[D_1(\alpha) - 1]$ and for Gumbel copula is $\tau = \frac{\theta-1}{\theta}$.

In general, incomplete data may include all possible types of observations. Interval-censored data are among important instances of incomplete data. Recall that, when a variable stochastically occurs in an interval such as (L, R), or $(L \leq T \leq R)$, then we say that the variable T is interval-censored. When we work with two correlated, interval-censored datasets, we may observe the intervals (L_{1i}, R_{1i}) and (L_{2i}, R_{2i}) instead of the variables T_{1i} and T_{2i} which lie in the related intervals, respectively. In this case, the lower and upper bounds are considered as $0 \leq L < R < +\infty$. Also, the variables are said to be right-censored if $R_{1i} = +\infty$ $(R_{2i} = +\infty)$, and left-censored if $L_{1i} = 0$ $(L_{2i} = 0)$. Moreover, the data is said to be complete (exact) if $L_{1i} = R_{1i}$ $(L_{2i} = R_{2i})$ with respect to T_1 and T_2.

For right-censored data, researchers such as Kaplan and Meier (1938) and Beran (1981) proposed a non-parametric estimator of the survival function of a univariate variable. Then, Dabrowska (1988) proposed a non-parametric estimator of the joint survival function of bivariate, right-censored data based on the Kaplan–Meier estimator. Turnbull (1976) proposed a unconditional maximum likelihood estimator (NPMLE) of survival functions. Then, Dehghan and Duchesne (2011) proposed a non-parametric conditional estimator of the survival function given interval-censored data, referred to as generalized Turnbull's estimator (GTE). The latter estimator was adapted for all types of incomplete data and downsized to Kaplan–Meier and unconditional Turnbull's estimators. This approach is able to estimate $S(t \mid z_0) = P[T > t \mid Z = z_0]$, where z_0 is a certain value of a Z covariate, and T is the related variable that can be interval-censored, that is, belongs to an interval like (L, R).

In practice, we have to work with bivariate correlated and interval-censored data. So, we can estimate the marginal survival functions separately. But, it is not still possible to estimate the joint survival function according to interval-censored data. As mentioned above, the copula approaches have proposed the estimation of joint distributions given complete data with dependence parameter(s). In this study, we propose a solution based on the GTE, midpoint imputation, and the copula methods.

Let the data appear as a triple $\{(L_{ki}, R_{ki}, Z_{ki}); k = 1, 2, i = 1, ..., n\}$ instead of two survival times, say $(T_{ki}; k = 1, 2; i = 1, ..., n)$, respectively. Then, by using the GTE (Dehghan and Duchesne , 2015), the normal kernel of the covariate, one can estimate the related survival functions. Let $(t_{1i}^{imp}, t_{2i}^{imp})$ be the corresponding imputed values. Since there is a one-to-one correspondence between survival variables and their related survival functions, one can use the GTE to estimate their corresponding survival function as $(\tilde{S_1}^{imp}, \tilde{S_2}^{imp})$. Finally, the dependence parameter of the copula will be estimated based on the sample size and the length of the interval-censored data, and compared with that of the precise data. In addition, the estimation of Kendall's τ will be explained in the next section.

3 Estimation of Kendall's τ

In this section, we propose an estimator of Kendall's τ. Although the estimator does not have a closed form, it can be useful when the data are interval-censored. First, we explain the non-parametric and semi-parametric estimations of the dependence parameter by the copula model (Sklar, 1959).

3.1 Non-parametric estimation

To estimate the dependence parameter of the copula via Kendall's τ, given interval-censored data, let $(t_{1i}^{imp}, t_{2i}^{imp}); i = 1, ..., n$ be the midpoint-imputed values of the related intervals, respectively. Therefore, one can estimate the related survival functions as $(\tilde{S_1}^{imp}, \tilde{S_2}^{imp})$. Then, to estimate Kandall's tau with the following non-parametric formula, one can put the corresponding estimated survival functions in (1):

$$\hat{\tau}_p = \binom{n}{2}^{-1} \sum_{1 \leq i \leq j \leq n} a_{ij}^{imp} b_{ij}^{imp}.$$

Here, if $\tilde{S_{1i}}^{imp} < \tilde{S_{1j}}^{imp}$ $(\tilde{S_{2i}}^{imp} < \tilde{S_{2j}}^{imp})$, then $a_{ij}^{imp} = 1$ $(b_{ij}^{imp} = 1)$; otherwise, $a_{ij}^{imp} = -1$ $(b_{ij}^{imp} = -1)$. Based on the definition of Kendall's τ, we define a new τ_p estimator, namely, τ_{Imp}:

$$\begin{aligned} \tau_{imp} =& P((\tilde{S_{1i}}^{imp} - \tilde{S_{1j}}^{imp})(\tilde{S_{2i}}^{imp} - \tilde{S_{2j}}^{imp}) > 0) \\ &- P((\tilde{S_{1i}}^{imp} - \tilde{S_{1j}}^{imp})(\tilde{S_{2i}}^{imp} - \tilde{S_{2j}}^{imp}) < 0). \end{aligned}$$

According to the U-statistics theory, $E[a_{ij}^{imp} b_{ij}^{imp}] = \tau$. So,

$$\binom{n}{2}^{-1} \sum_{1 \leq i \leq j \leq n} \left(a_{ij}^{imp} b_{ij}^{imp} - \tau\right) = \hat{\tau}_p - \tau.$$

Therefore, $\sqrt{n}(\hat{\tau}_p - \tau)$ asymptotically converges to the normal distribution with a mean of zero. See Lee (1990) and Van der Vaart (1998).

3.2 Semi-parametric estimation

Pseudo-log-likelihood is a method for estimating the dependence parameter of a copula function, introduced by Genest et al. (1995). This method does not need additional assumptions on the marginal distribution functions, and estimates the dependence parameter of the copula model by replacing the estimated values of the survival function and maximizing the following equation:

$$\ell(\alpha) = \sum \log(\hat{C}_\alpha(\tilde{S_1}^{imp}(T_{1i}), \tilde{S_2}^{imp}(T_{2i}))).$$

Here, $\tilde{S_1}^{imp}$ and $\tilde{S_2}^{imp}$ are correlated survival functions at the midpoints of interval-censored data. Semi-parametric models can be defined based on copula models.

As reviewed above, some common copula models were proposed by Clayton, Gumbel and Frank (Clayton, 1978; Gumbel, 1960; Frank, 1979) in which one could estimate the dependence parameters based on Kendall's τ. For more information, see Genest and Rivest (1993); Genest et al. (1995); Oakes (1982); Wang and Wells (2000). Since there is a one-to-one correspondence between τ and the parameter of the copula, by fitting one of the previous copula families on the incomplete data under consideration, one can estimate Kendall's τ and the dependence parameter of the copula.

4 Simulation results and application of the proposed estimator to real-world data

Although it is possible to apply these methods to datasets of more than 2 dimensions, in this study, we focus on two-dimensional, incomplete datasets. Therefore, we apply the GT estimator, Kendall's τ, Sklar's theorem, and common survival copulas of the Archimedean families, to real, incomplete (interval-censored) data. Obviously, right-censored data would be less accurate than interval-censored data. We believe that when some of the variables are related to each other, it is better to first estimate the upper bound for the right-censored cases. Since the real-world data under consideration contains right-censored cases, we first estimate the related upper bounds according to these cases, and then replace the right-censored cases with interval-censored data.

4.1 Simulation results

Let T_{1i} and T_{2i} be the two random variables of the $\text{Weibull}(\text{Shape} = 3, \text{Scale} = 1 * Z)$, distribution, with Z being covariate. Note that in this study, the distribution of Z was considered to be uniform on the interval (5, 25), but one could take another distribution for the covariate Z, such as the normal distribution given by Epanechnikov or the normal kernel. To extract the desired samples, according to the above distribution and parameters, two samples of size n were considered, including the last visit/inspection time: $\{V_{i1}, V_{i2}, i = 1, ..., n\}$. Since the observations were required to be interval-censored, the sample data appeared to be of the form $\{(L_i, R_i, Z_i); i = 1, ..., n\}$. The simulation was done by using a homogeneous Poisson process with the rate of the interval's length censoring (for example, the rate could be the inverse of E(R-L)), the intervals being (L_{1i}, R_{1i}) and (L_{2i}, R_{2i}), containing T_{1i} and T_{2i} $(i = 1, \cdots, n)$ respectively.

For the joint distribution function of T_1 and T_2, we considered the following two cases:

- (T_1, T_2) follow Clayton's model by replacing their survival function in (**??**) as follows:
$$S(t_1, t_2) = \left(S_1(t_1)^{-\alpha} + S_2(t_2)^{-\alpha} - 1\right)^{-\frac{1}{\alpha}},$$
where according to (**??**), Kendall's τ for Clayton copula is $\tau = \frac{\alpha}{\alpha+2}$.
- (T_1, T_2) follow Frank's model by replacing their survival functions in (**??**) as follows:
$$S(t_1, t_2) = \frac{-1}{\alpha} \log\left\{1 + \frac{(e^{-\alpha S_1(t_1)} - 1)(e^{-\alpha S_2(t_2)} - 1)}{(e^{-\alpha} - 1)}\right\},$$

where according to (**??**), Kendall's τ for Frank copula is $\tau = 1 + \frac{4}{\alpha}[D_1(\alpha) - 1]$, and D_1 is the Debye function Debye (1912), defined for any positive integer k by $D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t-1}dt$.

The simulation results are presented in Tables 1 and 2 for cases 1 and 2, respectively. We set the parameter values as $\tau \in (0.2, 0.4, 0.6, 0.8)$, $n \in (30, 50, 100, 200)$, and Rate $\in (0.2, 0.5, 1, 2)$, and the number of iterations used in this study was 2000. Therefore, we calculated $\hat{\tau}_1 = \sum_{i=1}^{2000} \frac{\hat{\tau}_{1i}}{2000}$ as a non-parametric estimation of Kendall's τ. In the semi-parametric method, we considered the two common copula families of Clayton and Frank as parametric models, and the survival function estimator of the given interval-censored data as a non-parametric method. We first estimated the dependence parameter of the copula models. So we have then calculated $\hat{\tau}_2$ as estimations of Kendall's τ for the given interval-censored data, namely, the semi-parametric method of estimation. Finally, we compared them by using the criterion of mean square error (MSE).

Table 1: Estimation of τ under Clayton's family

		$\hat{\tau}_1$	$MSE_{\hat{\tau}_1}$	$\hat{\tau}_2$	$MSE_{\hat{\tau}_2}$	$\hat{\tau}_1$	$MSE_{\hat{\tau}_1}$	$\hat{\tau}_2$	$MSE_{\hat{\tau}_2}$
Rate	τ	$n = 30$				$n = 50$			
0.2	0.2	0.1429	0.0241	0.1529	0.0243	0.1425	0.0171	0.1500	0.0155
	0.4	0.2433	0.0349	0.2548	0.0321	0.2419	0.0284	0.2462	0.0243
	0.6	0.3696	0.0803	0.3579	0.0753	0.3631	0.0766	0.3445	0.0777
	0.8	0.4272	0.1476	0.4016	0.1537	0.4266	0.1499	0.3847	0.1663
0.5	0.2	0.2020	0.0190	0.2109	0.0179	0.1998	0.0101	0.2012	0.0083
	0.4	0.3536	0.0163	0.3530	0.0118	0.3478	0.0095	0.3399	0.0075
	0.6	0.5618	0.0143	0.5176	0.0164	0.5516	0.0115	0.5073	0.0154
	0.8	0.6771	0.0227	0.5984	0.0395	0.6681	0.0214	0.5951	0.0396
1.0	0.2	0.2198	0.0177	0.2284	0.0153	0.2183	0.0096	0.1216	0.0076
	0.4	0.3808	0.0150	0.3757	0.0102	0.3743	0.0085	0.3608	0.0065
	0.6	0.6197	0.0090	0.5746	0.0078	0.6155	0.0052	0.5747	0.0056
	0.8	0.7721	0.0054	0.6929	0.0121	0.7688	0.0036	0.7010	0.0097
2.0	0.2	0.2238	0.0172	0.2296	0.0148	0.2173	0.0093	0.2152	0.0071
	0.4	0.3847	0.0148	0.3823	0.0101	0.3817	0.0084	0.3728	0.0059
	0.6	0.6359	0.0083	0.5989	0.0063	0.6307	0.0052	0.5990	0.0040
	0.8	0.8071	0.0034	0.7431	0.0048	0.8024	0.0020	0.7493	0.0032
		$n = 100$				$n = 200$			
0.2	0.2	0.1365	0.0137	0.1355	0.0110	0.1336	0.0092	0.1267	0.0071
	0.4	0.2307	0.0243	0.2263	0.0221	0.2276	0.0236	0.2144	0.0222
	0.6	0.3527	0.0786	0.3239	0.0872	0.3488	0.0784	0.3115	0.0931
	0.8	0.4145	0.1548	0.3636	0.1859	0.4081	0.1581	0.3490	0.2011
0.5	0.2	0.2002	0.0049	0.1925	0.0038	0.1936	0.0026	0.1801	0.0019
	0.4	0.3392	0.0056	0.3256	0.0045	0.3352	0.0036	0.3167	0.0033
	0.6	0.5439	0.0096	0.5029	0.0147	0.5368	0.0091	0.4978	0.0174
	0.8	0.6604	0.0225	0.5920	0.0419	0.6522	0.0236	0.5886	0.0438
1.0	0.2	0.2157	0.0045	0.2053	0.0033	0.2114	0.0021	0.1945	0.0016
	0.4	0.3710	0.0039	0.3541	0.0033	0.3657	0.0019	0.3455	0.0017
	0.6	0.6138	0.0028	0.5735	0.0037	0.6016	0.0017	0.5756	0.0027
	0.8	0.7586	0.0032	0.7041	0.0091	0.7523	0.0031	0.7073	0.0083
2.0	0.2	0.2173	0.0043	0.2060	0.0031	0.2973	0.0017	0.2842	0.0015
	0.4	0.3792	0.0031	0.3626	0.0031	0.4600	0.0015	0.4402	0.0016
	0.6	0.6290	0.0023	0.6012	0.0020	0.6289	0.0029	0.6014	0.0021
	0.8	0.7971	0.0009	0.7571	0.0020	0.7975	0.0009	0.7577	0.0009

Table 2: Estimation of τ under Frank's family

		$\hat{\tau}_1$	$MSE_{\hat{\tau}_1}$	$\hat{\tau}_2$	$MSE_{\hat{\tau}_2}$	$\hat{\tau}_1$	$MSE_{\hat{\tau}_1}$	$\hat{\tau}_2$	$MSE_{\hat{\tau}_2}$
Rate	τ	$n = 30$				$n = 50$			
0.2	0.2	0.1453	0.0290	0.1434	0.0303	0.1385	0.0227	0.1334	0.0246
	0.4	0.2472	0.0480	0.2369	0.0518	0.2354	0.0427	0.2265	0.0471
	0.6	0.3796	0.0802	0.3601	0.0854	0.3697	0.0794	0.3536	0.0851
	0.8	0.4437	0.1495	0.4194	0.1539	0.4298	0.1570	0.4106	0.1623
0.5	0.2	0.2055	0.0161	0.2009	0.0151	0.2067	0.0098	0.2027	0.0097
	0.4	0.3577	0.0142	0.3455	0.0143	0.3864	0.0091	0.3436	0.0099
	0.6	0.5736	0.0139	0.5422	0.0164	0.5625	0.0109	0.5381	0.0138
	0.8	0.6863	0.0248	0.6437	0.0291	0.6827	0.0207	0.6501	0.0266
1.0	0.2	0.2171	0.0160	0.2140	0.0143	0.2209	0.0090	0.1275	0.0083
	0.4	0.3872	0.0113	0.3763	0.0096	0.3831	0.0065	0.3744	0.0060
	0.6	0.6337	0.0068	0.6019	0.0064	0.6297	0.0036	0.6047	0.0040
	0.8	0.7833	0.0037	0.7375	0.0061	0.7778	0.0028	0.7429	0.0051
2.0	0.2	0.2226	0.0143	0.2213	0.0130	0.2230	0.0085	0.2209	0.0078
	0.4	0.3934	0.0105	0.3856	0.0088	0.3950	0.0062	0.3874	0.0058
	0.6	0.6539	0.0066	0.6253	0.0055	0.6486	0.0034	0.6266	0.0029
	0.8	0.8179	0.0022	0.7762	0.0024	0.8173	0.0011	0.7851	0.0014
		$n = 100$				$n = 200$			
0.2	0.2	0.1370	0.0118	0.1332	0.01267	0.1328	0.0094	0.1306	0.0100
	0.4	0.2324	0.0278	0.2263	0.0304	0.2278	0.0267	0.2244	0.0284
	0.6	0.3552	0.0835	0.3482	0.0878	0.3522	0.0843	0.3469	0.0868
	0.8	0.4174	0.1585	0.4062	0.1629	0.4124	0.1638	0.4071	0.1659
0.5	0.2	0.1978	0.0050	0.1951	0.0050	0.1957	0.0026	0.1943	0.0027
	0.4	0.3446	0.0052	0.3382	0.0058	0.3392	0.0037	0.3362	0.0041
	0.6	0.5538	0.0095	0.5387	0.0117	0.5473	0.0095	0.5387	0.0109
	0.8	0.6691	0.0229	0.6495	0.0271	0.6618	0.0243	0.6508	0.0269
1.0	0.2	0.2181	0.0045	0.2155	0.0043	0.2131	0.0022	0.2119	0.0021
	0.4	0.3819	0.0036	0.3701	0.0035	0.3741	0.0018	0.3715	0.0018
	0.6	0.6223	0.0051	0.6066	0.0026	0.6144	0.0014	0.6052	0.0018
	0.8	0.7711	0.0026	0.7890	0.0043	0.7650	0.0028	0.7517	0.0039
2.0	0.2	0.2206	0.0043	0.2191	0.0041	0.2174	0.0022	0.2166	0.0021
	0.4	0.3878	0.0036	0.3835	0.0030	0.3853	0.0016	0.3833	0.0015
	0.6	0.6423	0.0034	0.6280	0.0023	0.6389	0.0007	0.6307	0.0007
	0.8	0.8112	0.0005	0.7906	0.0039	0.8055	0.0003	0.7933	0.0005

According to Tables 1 and 2, the non-parametric approach performed slightly better than the semi-parametric approach. The proposed estimators were fully efficient in the independence position (Genest et al., 1995), except in the presence of small dependence between the variables and small sample size (n).

In addition, the MSE values increase the τ values do so, given the amall sample size (n). When the rate is small, that is the length of the interval-censored is large, the MSE values increase by the τ values, but when the rate is big that is the length of the interval-censored is small, the MSE values decrease by the τ values. Finally the MSE of $\hat{\tau}_1$ and $\hat{\tau}_2$ are close together and also have the same behaviour. It seems that the Clayton model is not sensitive to the value of sample size, and its performance is good even if n is small ($n = 30$). Also, in this model, $\hat{\tau}_2$ performs well, when the length of the interval-censored is large. Therefore, the results are close to those obtained by Oakes (2008).

Generally, in interval-censored data, decreasing the length of intervals improves the accuracy, as expected. Note that when the dependence value is big, the non-parametric and semi-parametric methods perform like one another. The estimation of the dependence parameter of the copula is provided in the next table, in which $\hat{\alpha}$ and $\hat{\alpha}_{cen}$ are non-parametric and semi-parametric estimations of α with variances of $\sigma^2(\hat{\alpha})$ and $\sigma^2(\hat{\alpha}_{cen})$, respectively. The coefficients CV and CV_{cen} of variation, as criteria, were subject to the non-parametric ($\hat{\alpha}$) and semi-parametric ($\hat{\alpha}_{cen}$) estimators, respectively.

Since the CV criteria are free of measuring units, they refer to the small variance according to it's related mean, when it is less than 1. Although the variance of $\hat{\alpha}_{cen}$ increases according to the values of α, it can be seen that CV is decreasing by the values of α and the sample size (n). Furthermore, CV is decreasing by the rate of interval censoring. The $\hat{\alpha}_{cen}$ and its variance ($\sigma^2_{\hat{\alpha}_{cen}}$) according the rate values. Also $\sigma^2_{\hat{\alpha}_{cen}}$ increase by the α values. In general the $\hat{\alpha}$ is greater than $\hat{\alpha}_{cen}$. Finally, Table 3 provides the results according to the interval-censored data, given various values of the interval rate. The MSE of the $\hat{\tau}_2$ is slightly better than the MSE of the $\hat{\tau}_1$ when the sample size is small, but they have the same behaviour when the rate and τ values are increasing.

Table 3: Estimation of the dependence parameter of the copula (α)

					$\hat{\alpha}_{cen}$	$\sigma^2(\hat{\alpha}_{cen})$	CV_{cen}	$\hat{\alpha}_{cen}$	$\sigma^2(\hat{\alpha}_{cen})$	CV_{cen}
n	α	$\hat{\alpha}$	$\sigma^2(\hat{\alpha})$	CV	Rate=0.2			Rate=0.5		
100	0.5	0.54	0.0317	0.3297	0.33	0.0263	0.4914	0.48	0.0295	0.3578
	1.3	1.17	0.0664	0.2202	0.60	0.0329	0.3291	0.94	0.0495	0.2366
	3	3.14	0.4153	0.2052	0.98	0.0419	0.2088	2.06	0.1144	0.1642
	8	7.80	1.9600	0.1795	1.16	0.0446	0.1820	2.96	0.1903	0.1473
					Rate=1			Rate=2		
	0.5				0.52	0.0308	0.3374	0.53	0.0312	0.3332
	1.3				1.12	0.0580	0.2150	1.15	0.0613	0.2151
	3				2.73	0.1956	0.1620	3.06	0.2531	0.1644
	8				4.84	0.5607	0.1547	6.40	1.1177	0.1652
					Rate=0.2			Rate=0.5		
200	0.5	0.50	0.0141	0.2375	0.29	0.0114	0.3681	0.45	0.0132	0.2553
	1.3	1.13	0.0453	0.1884	0.56	0.0142	0.2127	0.94	0.0225	0.1595
	3	3.21	0.1471	0.1195	0.91	0.0176	0.1457	2.00	0.0509	0.1128
	8	7.93	0.8565	0.1167	1.08	0.0182	0.1163	2.86	0.0896	0.1043
					Rate=1			Rate=2		
	0.5				0.49	0.0138	0.2397	0.50	0.0139	0.2357
	1.3				1.08	0.0266	0.1510	1.11	0.0284	0.1518
	3				2.71	0.0907	0.1111	3.04	0.1138	0.1109
	8				4.88	0.2397	0.1003	6.56	0.4940	0.1071

4.2 Real-world data

In this subsection, we apply the aforementioned methods to the data in hand to estimate the copula dependence parameter which includes the observed intervals of Cytomegalovirus (CMV). These are samples of urine and blood, taken from 204 patients, which were sampled every 12 and 4 weeks, respectively. The data have already been studied by Finkelestein and Goggins (2000). These datasets include right and interval-censored times. The dataset also includes some information about each patient's baseline and the last CD4 cell count. The interesting subjects on the dataset could be to study the association between CMV shedding in blood and urine times, fit a copula as a joint distribution, and then estimate the dependence parameters given the interval censoring.

To estimate this association, we considered T_1 and T_2 as times of CMV shedding in blood and urine, respectively, and Z to be the common covariate for T_1 and T_2, that is, the last CD4 cell count for every patient. To modify the right-censored data, via the relationship between the baseline CD4 (BASCD4) and the last CD4 (LASTCD4) cell count, we found a linear regression model between them to estimate the upper bounds ($\hat{R} = L + DIFF_{L,R}$), subject to the right-censored ($R = +\infty$) observations, as shown below:

$$DIFF_{L,R} = 3.29 + 0.1 * LASTCD4 + 0.0799 * BASCD4.$$

Here, the difference between L and R, namely $DIFF_{L,R}$, is considered as a response variable, estimated by $BASCD4$ and $LASTCD4$, which have significant testing, for example, p-value = 0 0.01779. Therefore, the right-censored observations were modified such that $[L_{1k}, \hat{R}_{1k}]; k = 1, 2$, according to the variables $T_{1k}; k = 1, 2$. Then, we estimated the correlated survival functions with the non-parametric and semi-parametric methods mentioned in Section 3. So, we estimated Kendall's τ. The non-parametric estimation of τ was 0.2313,

Then, we compared the survival functions of patients with low and high CD4 cell numbers. To do so, we considered two groups; the low group having CD4 cell counts lower than 75, and the high group having CD4 cell counts greater than 75. As expected, the low group had a lower survival probability compared to the high group. Figure 1 graphically demonstrates this fact.

After estimating the right-censored time to event by the method mentioned above, in goodness-of-fit-test analysis, we used the (Schepsmeier and Brechmann, 2013) which the results have presented at Table 4. This shows the results of fitting copula families on the dataset; four copula families, namely Gumble, Frank, Clayton, and Tawn were good fits on the dataset. The results suggest that the Gumble copula is the best fit to the dataset based on the Akaike Information Criterion (AIC). However, if we consider the Log-likelihood (log-like) criterion, the Tawn copula is found to be a better fit than the other three copula families. The Tawn copula is particularly interesting because it is defined based on Pickands dependence functions, which are a function used to model the dependence between random variables.

Table 4: The fitted copula families for dataset

Copula	Association Parameter	Kendal's τ	AIC	BIC	Loglike.
Gumbel	1.85	0.46	-18.4	-16.71	10.2
Tawn2	2.34(0.65)	0.42	-17.99	-14.62	11
Frank	4.98	0.45	-17.03	-15.34	9.51
Joe	2.24	0.40	-16.66	-14.97	9.33

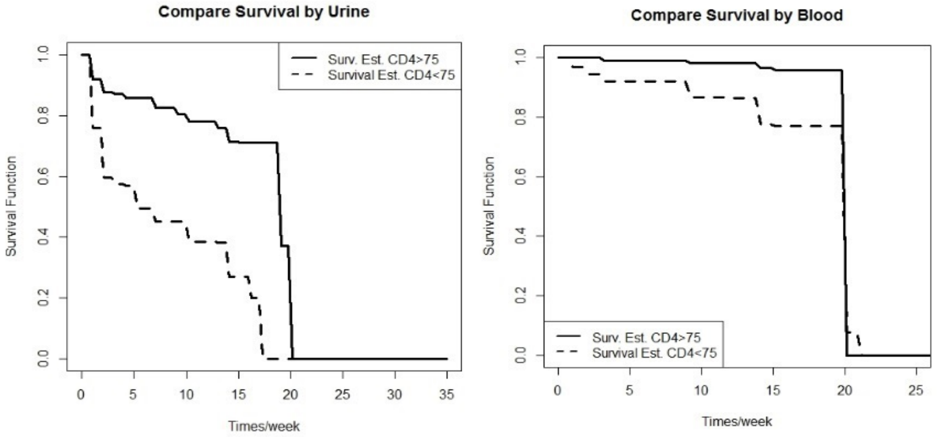

Figure 1: Survival times for CMV shedding in blood and urine

5 Conclusion

Obviously, a parametric method works better than a non-parametric method. Usually, when the dataset is incomplete, it is impossible to use a parametric model. Therefore, both non-parametric and semi-parametric models are needed to work well in simulation. Since the results of non-parametric approaches are slightly better than semi-parametric approaches, for interval-censored data, non-parametric methods may be preferred. As expected, when the sample size and the rate of interval censoring increase, the accuracy of the results is improved regardless of the method that is utilized. Based on our proposed approaches, the estimator of Kendall's τ worked well as a non-parametric method under interval-censored data, when the expectation of $R-L$, $E(R-L)$, was small. The rate of the censoring process was approximately equal to the inverse of $E(R-L)$, and played an essential role in the improvement of the accuracy of the results. Since the dataset under consideration included many right-censored cases, to adapt the position of the dataset for the interval-censored cases and improve the accuracy of the results, we first extracted a regression relationship between some variables, then estimated the upper bounds of the right-censored cases. Because CD4 cell count has a very important role in the patient's survival, we used the non-parametric estimation of the conditional survival function given the CD4 cell count. The research results confirmed the efficiency proposed by (Genest et al., 1995) for estimating Kendall's τ in the independence mode. Note, however, that according to the results of the goodness-of-fit-test of copula models on the modified dataset, by using VineCopula (Schepsmeier and Brechmann (2013)), the Gumbel copula was fit to the given unconditional and conditional data with the value of the AIC. Still, the estimated values of Kendall's τ were 0.22 and 0.46, respectively. As expected, conditioning the survival function on the covariate, compressed the neighbourhood of the covariate centre so that Kendall's τ significantly increased.

Acknowledgment

We gratefully thank Dianne Finkelstein, Harvard T.H. Chan School of Public Health, for her assistance in data acquisition. The authors are grateful to anonymous referees for their helpful comments and suggestions to improve this manuscript.

References

Beaudoin, D., Duchesne, T. and Genest, C. (2007). Improving the estimation of Kendall's τ when censoring affects only one of the variables. *Computational Statistics & Data Analysis Journal*, **51**, 5743–-5764.

Beran, R. (1981). Non-parametric regression with randomly censored survival data. *Statistics and Probability Letters*, **20**, 225–234.

Betensky, R.A. and Finkelstein, D.M. (1999). An extension of Kendall's coefficient of concordance to bivariate interval-censored data. *Statistics in Medicine Journal*, **18**, 3101–3109.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141–-151.

Dabrowska, D.M. (1988). Kaplan-meier estimate on the plane. *The Annals of Statistics*, **16**, 1475–-1489.

Debye, P. (1912). Zur Theorie der spezifischen Wärmen. *Annalen der Physik*, **39**(10), 789-839.

Dehghan, M.H. and Duchesne, T. (2011). A generalization of Turnbull's estimator for non-parametric estimation of the conditional survival function with interval-censored data. *Lifetime Data Analysis*, **17**, 234–-255.

Dehghan, M.H. and Duchesne, T. (2015). Generalization of Turnbull's estimator. *R Package, Available from: http://127.0.0.1:15947/library/gte/html/gte.html.*

Derumigny, A. and Fermanian, J.D. (2019). On kernel-based estimation of conditional Kendall's tau: finite-distance bounds and asymptotic behavior. *Dependence Modeling*, **7**(1), 292-321.

Finkelestein, D.M. and Goggins, G. (2000). Analysis of failure time data with dependent interval censoring. *Biometrics*, **584**, 298–304.

Frank, M.J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae*, **19**, 194–226.

Genest, C. and Rivest, L.P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American statistical Association*, **88**, 1034–1043.

Genest, C., Goudi, K. and Rivest, L. P. (1995). A semi-parametric estimation procedure for dependence parameters in multivariate families of distributions. *Biometrika*, **82**, 543–552.

Genest, C., Quessy, J.F. and Rémillard, B.(2006). Goodness of fit procedures for copula models based on the probability integral transformation. *Scandinavian Journal of Statistics*, **33**, No. 2, 337–366.

Gumbel, E.J. (1960). Bivariate exponential distributions. *American Statistical Association Journal*, **55**, 698–707.

Hesieh, J.J. and Li, Z.J. (2017). Estimation and test of conditional Kendall's τ under bivariate left-truncation data. *Communication in Statistics-Theory and Methods*, **46**, 6635-6644.

Hoeffding, W. (1948). A class of statistics with asymptotic normal distribution. *The Annals of Mathematical Statistics*, **19**, 293–325.

Jahanshahi, S. M.A., Habibirad2, A. and Fakoor, V. (2020). Some new goodness-of-fit tests for Rayleigh distribution. *Pakistan Journal of Statistics*, **16**(2), 305–315.

Kanga,S-H and Kim, Y-J. (2021). Association measure of doubly interval- censored data using a Kendall's τ estimator. *Communications for Statistical Applications and Methods*, **28**, 151–159.

Kaplan, E.L. and Meier, M. (1938). Non-parametric estimation from incomplete observations. *American Statistical Association Journal*, **53**, 457–481.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81–89.

Kim,Yang-Jin (2015). Estimation of conditional Kendall's τ for bivariate interval- censored Data. *Communications for Statistical Applications and Methods*, **22**, 599–604.

Koziol, A.J. (1980). Goodness-of-fit tests for randomly censored data. *Biometrika*, **67**(3), 693–696.

Lakhal, L. and Rivest, L.P. and Abdous, B. (2008). Estimating association and survival in a semi- competing risks model. *Biometrics*, **64**, 180–188.

Lee, A.J. (1990). U-Statistics: Theory and Practice. *Marcel Dekker, New York.*

Lim, J. and Meier, M. (2006). Permutation procedures with censored data. *Computational Statistics & Data Analysis*, **50**, 332–345.

Martin, E.C. and Betensky, R.A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's τ. *Journal of the American Statistical Association*, **100**, 484–492.

Oakes, D. (1982). A concordance test for independence in the presence of censoring. *Biometrika*, **38**, 451–555.

Oakes, D.(2008). On consistency of Kendall's τ under censoring. *Biometrika*, **95**, 997–1001.

Schepsmeier U. and Brechmann E.C. (2013). BiCopSelect and BiCopCompare Copulas Packages.
http://127.0.0.1:13916/library/VineCopula/html/VineCopula-package.html.
http://127.0.0.1:13916/library/VineCopula/html/BiCopCompare.html.

Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Institute Statistique del*, **8**, 229–231.

Tsai, W.Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika*, **77**, 169–177.

Turnbull, B.W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data. *Journal of Royal Statistical Society*, **38**, 290–295.

Van der Vaart, A.W. (1998). *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, W. and Wells, M.T. (2000). Estimation of Kendall's τ under censoring. *Statistica Sinica*, **10**, 1199–1215.

Weier, D.R. and Basu, A.P. (1980). An investigation of Kendall's τ modified for censored data with applications. *Journal of Statistical Planning and Inference*, **4**, 381–390.