文章编号:1000-5862(2019)01-0013-09

Evaluating the Correlation Coefficient Between Bivariate Survival Times
——a Copula-Based Approach

HAN Xiaozhen[1],FU Pingfu[2,3*]

(1. Department of Quantitative Health Sciences,Cleveland Clinic,Cleveland Ohio 44195,USA;
2. Departments of Population and Quantitative Health Sciences,Case Western Reserve University,Cleveland Ohio 44106,USA;
3. Case Comprehensive Cancer Center,Case Western Reserve University,Cleveland Ohio 44106,USA)

Abstact: The analysis of correlations within pairs of survival times is of great interest to many researchers in biology and medicine. The analysis objective is to investigate the association of bivariate survival data under the setting of low-moderate percentage of censoring through Monte Carlo simulations using a copula approach. Here the association of bivariate survival data is estimated using Spearman's correlation coefficient. The results from simulation studies show that when the percentage of censoring is low,Gumbel-based estimation procedure is much more robust,and the stronger a positive association is,the more accurate estimate can be obtained when the censoring percentage is 0% and 30%. This is true for the Frank,Gumbel and Clayton-based estimation procedures under the condition that the copula assumption made here is the same as the true one.

Key words: bivariate survival data; copula approach; correlation analysis; correlation coefficient; censoring percentage

中图分类号: O 211; Q 332 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2019.01.03

0 Introduction

In medical research,it is useful for physicians to know the correlation among pairs of survival times in terms of prognosis,patient care including treatment decision making. For example,early stage patients with head and neck squamous cell carcinoma(HNSCC) are at high risk of recurrence and having second primary tumor after being treated and having an excellent prognosis [1]. The survival times during the second treatment period in this special case are censored for some. It is also of interest to get an accurate estimate for the correlations of bivariate survival times under censoring in general.

Here we are interested in the correlation between the time until recurrence of the cancer during the first application (T_1,Q_1) and the time until the patient dies during the second application (T_2,Q_2). T_1 is the initial desease free interval and T_2 is the overall survival after the salvage surgery. To be more accurate,T_1 is defined as the minimum of (X_1,C_1) where X_1 is the time to recurrence and C_1 is the censoring time independent of X_1. T_2 is defined as the minimum of (X_2,C_2) where X_2 is the time to death and C_2 is the censoring time independent of X_2. Both (T_1,Q_1) and (T_2,Q_2) are survival times under censoring where T_k represents time to the event or time to censoring,and Q_k is a censoring indicator. When the correlation of such survival times is a concern,the dependence of the two is mostly assumed monotonic. Under this circumstance, quantifying the correlation using rank correlation coefficients instead of using Pearson's correlation coefficient seems more appropriate [2-3]. One of the popular rank correlation coefficients is the Spearman's rank-order correlation coefficient(Spearman's r_s). Most medical researchers are

收稿日期: 2018-10-30
基金项目: 凯斯综合癌症中心生物统计学和生物信息学(P30CA43703)资助项目.
通信作者: 付平福(1963-),男,江西樟树人,副教授,主要从事生物统计方面的研究. E-mail: pxf16@case.edu

very familiar with Spearman's r_s and intuitively know how "large" the correlation is when they get the result.

One method among other alternatives to get an accurate estimation for the Spearman's r_s in the setting of bivariate survival times is to use maximum likelihood estimation within common copulas, a semiparametric approach developed by J. H. Shih et al [4]. In this paper, we use a copula-based approach to evaluate the correlation coefficient of bivariate survival times under the setting of low-moderate percentage of censoring. By making comparisons with other existing approaches, we find that the copula-based approach has advantages in some aspects. In addition, the results from simulation studies provide significant references for physicians.

0.1 Three Common Measures of Association

Whether and how the variables of interest are related with each other is attractive for many investigators. Three commonly used indices are the Pearson product moment correlation, Spearman's rank-order correlation and Kendall's tau correlation. The Pearson product moment correlation is commonly used to measure the association between two continuous variables and often denoted as the Greek letter ρ. It ranges from -1 to 1 and is calculated through dividing the covariance of the two variables by the product of their standard deviations. A positive ratio indicates a positive linear association and a negative ratio indicates a negative linear association. When the ratio is zero, it could indicate either the absence of linear association or the absence of any kind of association depending on whether the data have a bivariate normal distribution [5-6]. As for the other two popular measures which are Spearman's rank-order correlation coefficient and Kendall's tau correlation coefficient, they can be used for the association between two ordinal or interval variables. Their absolute values could indicate how strong and weak the monotonic relationship between the two variables is [6]. In medical studies with survival endpoints, the estimating method becomes more sophisticated due to censoring. Under this circumstance, many recent works have demonstrated the advantages of using Kendall's tau and Spearman's rank-order correlation coefficient. The former is more easily generalized for censored data and the latter could be innovatively adapted in censoring cases by a semiparametric approach incorporating a copula [4].

0.2 Copula

It is Abe Sklar(1959) who first used the word copula——a Latin noun meaning "a link, tie, bond" [7], in the statistical world for the Sklar's Theorem to describe the functions that "join together" one-dimensional distribution functions to form multivariate distribution functions [8]. Since then, there have been applications of copula in several fileds including medicine, finance, engineering and climate research among others. In this paper, only bivariate versions of the copulas are considered. Thus, informally, if (X, Y) is a pair of continuous random variables with distribution function $H(x, y)$ and marginal distributions $F_x(x)$ and $F_y(y)$ respectively, then $U = F_x(x) \sim U(0,1)$ and $V = F_y(y) \sim U(0,1)$ and the distribution function of (U, V) is a copula. Copulas can be used as very powerful tools for modeling dependence between random variables with their unique advantages like studying non-linear dependence, being able to measure dependence for heavy tail distributions, being able to study asymptotic properties of dependence structures and flexible usage with parametric, semi-parametric or non-parametric assumption [8-9]. Copulas also work well when the random variables of interest represent the lifetimes of observations with censoring in some population. In this case, the probability of an individual living beyond time x is valued and always specified by the survival function which is $\bar{F} = P(X > x) = 1 - F(x)$, here, F denotes the cumulative distribution function of X. For a pair (X, Y) of random variables whose joint distribution is H, the joint survival function can be given as $\bar{H}(x, y) = P(X > x, Y > y)$. The margins of joint survival function are univariate survival functions $\bar{F}$ and $\bar{G}$. Assuming the copula of X and Y is C, we can have

$$\bar{H}(x,y) = 1 - F(x) - G(x) + H(x, y) = \bar{F}(x) + \bar{G}(y) - 1 + C(F(x), G(y)) = \bar{F}(x) + \bar{G}(y) - 1 + C(1 - \bar{F}(x), 1 - \bar{G}(y)).$$

If a function $\hat{C}(u, v) = u + v - 1 + C(1 - u, 1 - v)$ could be defined from I^2 into I, we then have $\bar{H}(x, y) = \hat{C}(\bar{F}(x), \bar{G}(y))$. I^2 is the product $I \times I$ where $I = [0, 1]$. The function $\hat{C}$ is a copula and referred as the sur-

vival copula of X and Y[8].

0.3 Maximum Likelihood Estimation for Spearman's r_s within Common Copulas

Given a sample of n pairs of possibly censored times (t_{1i}, t_{2i}), $1 \leqslant i \leqslant n$, and corresponding status indicators (Q_{1i}, Q_{2i}), we can use the Nelson-Aalen method to get the probability of marginal survival function $S_k(t_k)$, $k = 1, 2$, at observed survival or censoring time t_k. Then we define $u_i = 1 - S_1(t_{1i})$ and $v_i = 1 - S_2(t_{2i})$, and approximately, they are all uniformly distributed. This transformation guarantees the use of copulas later, because a copula is a mathematically well-developed bivariate distribution and it needs its marginal distributions to be uniformly distributed. Since u_i and v_i are uniformly distributed, we can construct copula $C_\theta(u, v)$ after choosing a common type of copula, where θ is dependence parameter. In this paper, we will try four of the most commonly used copulas. They are Gumbel, Frank, Clayton and Normal. The parameter θ for the chosen copula can be estimated by maximum likelihood estimation. After that, Spearman's r_s can be obtained by integration over the specified copula distribution. Based on the explanation above, we develop four estimation procedures using R in which four different copula assumptions that are Gumbel, Normal, Frank and Clayton are made. The procedures develop based on the properties of rank correlation (here we consider Kendall's tau and Spearman's r_s), Both Kendall's tau $(\rho_T(X, Y))$ and Spearman's $r_s(\rho_s(X, Y))$ can be expressed in terms of copulas as follows:

$$\rho_T(X, Y) = 4\int_0^1\int_0^1 C(u, v)\, \mathrm{d}C(u, v) - 1,$$

$$\rho_s(X, Y) = 12\int_0^1\int_0^1 \{C(u, v) - uv\}\, \mathrm{d}u\mathrm{d}v.$$

1 Data Set Description

We applied the methods and the procedures developed to two well-known publically available data sets, which are Diabetic retinopathy[10] and Infections under dialysis[11]. The Diabetic retinopathy analysis contains a sample of 197 paired survival times because both eyes of an individual are observed at the same time and it is of interest to know how times to blindness of a treated and an untreated eye correlated with each other when the patient had diabetic retinopathy[10]. The infections under dialysis analysis focuses on whether and how time until infection of the first application of a portable dialysis machine correlates with the time until infection of the second application. This data set is of 38 patients with their times until infection of the two time periods and corresponding censoring indicators. The indicator 1 means the event infection occurred and the indicator 0 means the catheter is removed because of other reasons than infection so that the time to infection is censored. There are 23 patients having both uncensored times, 20 patients have one censored time and 3 patients have both censored times[11]. The censoring percentage is 24%.

2 Approach

2.1 Spearman's r_s Diabetic Retinopathy and Infections under Dialysis Analysis

Again, given a sample of n pairs of possibly censored times (t_{1i}, t_{2i}), $1 \leqslant i \leqslant n$, and corresponding status indicators (Q_{1i}, Q_{2i}), we can use the Nelson-Aalen method to get the probability of survival $S_k(t_k)$, $k = 1, 2$, at observed survival or censoring times t_k. The R function used to get the Nelson-Aalen estimator is basehaz function from R survival package as the Breslow hazard estimator for a Cox model can be reduced to the Nelson-Aalen estimator when there are no covariates. The relationship used to get the probability of survival is $S(t) = \exp(-H(t))$, where $H(t)$ is called the integrated or cumulative hazard[12]. Then we get u_i and v_i by using the definition $u_i = 1 - S_1(t_{1i})$ and $v_i = 1 - S_2(t_{2i})$. The corresponding censored information for u_i and v_i is the same as the censored indicator which are Q_{1i} and Q_{2i}. Both u_i and v_i are all uniformly distributed.

After getting data pairs (u_i, v_i) with their corresponding censored information (Q_{1i}, Q_{2i}), we constructed a R function for the (pseudo-) likelihood of the copula parameter θ. The likelihood of the copula parameter θ is different according to the chosen copula $C_\theta(u, v)$[5]. Four common types of copulas i. e. Clayton, Frank, Gumbel and Normal were considered in this application.

The distribution function of the chosen copula and

its derivatives are used to express the (pseudo-) likelihood defined in M. Schemper et al's paper and their technical report [5].

In order to get the maximum likelihood estimator (MLE) for θ which could realize the maximum value of the (pseudo-) likelihood, we use the R general-purpose optimization function optim and chose the option "SANN" as the method under the condition that our chosen copula is Clayton and Gumbel. The method "SANN" is by default a variant of simulated annealing and very useful in getting to a good value on a rough surface [13]. As for the Frank copula, the method specified in the option is "BFGS", a quasi-Newton method which is published simultaneously in 1970 by Broyden, Fletcher, Goldfarb and Shanno [14-16]. The above approach didn't fit for the Normal Copula. A self-coded R function for the (pseudo-) likelihood of the Normal Copula is thus prepared and a customized while loop was used to get θ which could ensure the maximum value of the prepared likelihood.

The last step is to get Spearman's r_s through (numerical) integration over the chosen copula's distribution function by inserting the estimated copula parameter θ in the equation: $r_s = 12\int_0^1\int_0^1 C_\theta(u,v)\,\mathrm{d}u\mathrm{d}v - 3$. For Clayton, Gumbel and Frank copulas, the two-fold integration can be accomplished by using a custom-constructed R function. However, it is different to get the r_s by integration over the normal copula because the problem is about a four-fold integration and all four folds are on lines (not regions). A Monte Carlo simulation approach was implemented in this case.

For the confidence interval, there are five kinds of confidence intervals based on the bootstrap method. They are bootstrap-t interval, standard normal interval, percentile interval, bias corrected and accelerated interval and approximate bootstrap confidence interval. Among them, the percentile interval was chosen because it has both a transformation-respecting property and a range-preserving property. The range-preserving property is desirable since the values of r_s lie in the interval $[-1,1]$ [17].

In order to know which estimation procedure is the optimal among 4 copulas for each application, A_{IC} value is calculated as follows [18-19].

$A_{IC} = 2K - 2\ln L$, where K is the number of estimated parameters which is one and L is the (pseudo-) likelihood value calculated based on the principles illustrated before.

2.2 Simulation Study

In order to evaluate our four proposed estimation procedures above, we carried out a series of simulation studies.

Simulation study 1:

Using the package copula in R, we firstly generate 1 000 pairs of times (t_{1i}, t_{2i}), $1 \leqslant i \leqslant 1\,000$. The function $m_{vdc}()$ could be used to define a multivariate distribution with given margins. Within that function, we could specify the copula type using the option copula and assign values to the parameters both of the bivariate copula and of its corresponding margins [20]. The result returned is an m_{vdc} object. After generating 1 000 pairs of times as realizations of random variables T_1 and T_2 which are unit exponentially distributed. However, the joint distribution of $U = \exp(-T_1)$ and $V = \exp(-T_2)$ is a bivariate normal copula with U and V uniformly distributed. By changing the option accordingly, four bivariate copulas are used: Normal, Clayton, Frank and Gumbel. And 1 000 pairs of times are generated respectively.

Since our aim is to illustrate how our four estimation procedures will perform under the influence of low-moderate percentage of censoring, we use the method mentioned in M. Schemper et al's paper which is assuming pairs of individuals to enter the study at a constant rate [6]. Thus, we assume pairs of individuals went into study constantly during the time period $(0,m)$. Under this assumption, follow-up times O will be uniformly distributed in $(0,m)$. Then we use function runif to generate 1 000 data pairs. If $O_i < T_{1i}$ or $O_i < T_{2i}$, then censoring indicators are defined. We specify the value m in each simulation in order to achieve the overall censoring proportions of 0% and 30%.

By using the principles above, for each combination of underlying copula, underlying Spearman's r_s and censoring percentage, 1 000 pairs of times (t_{1i}, t_{2i}), $1 \leqslant i \leqslant 1\,000$, and their corresponding censoring indicators were generated. There are 36 combinations in total. For example, the first combination is underlying

Normal Copula, underlying Spearman's r_s of 0.00 and the underlying percentage of censoring of 0%. Four proposed estimation procedures are then used to get the estimated Spearman's r_s denoted as NCE, FCE, CCE and GCE for normal, Frank, Clayton and Gumbel copula estimations, respectively.

Secondly, for each combination of underlying copula, underlying Spearman's r_s and censoring percentage, 5 000 pairs of times (t_{1i}, t_{2i}), $1 \leqslant i \leqslant 1\ 000$, and their corresponding censoring indicators were generated. Four proposed estimation procedures are used to get the estimated Spearman's r_s again and denoted as NCE, FCE, CCE and GCE.

Simulation study 2:

Compared with simulation study 1, for each combination of underlying copula, underlying Spearman's r_s and censoring percentage, instead of using a single sample of $n = 1\ 000$ or 5 000 pairs of times. We simulated 1 000 samples with $n = 50$ and $n = 200$ separately. The sample mean for the r_s is used as the point estimator and the confidence interval is calculated as

$$(\bar{X} \pm t_{\alpha/2}(n-1)\, s/\sqrt{n})\ [8].$$

Here, s is the sample standard deviation.

3 Results

Table 1 summarizes all point estimates of the Spearman's r_s for the two well-known data sets based on the estimation assuming Normal, Frank, Clayton and Gumbel copulas respectively, and is also a summary of all 95% confidence intervals using percentile interval based on bootstrap method. From the results of the two well-known data sets, our normal-based estimation procedure tends to have a larger 95% confidence interval and a lower biased estimate than the other three estimation procedures.

Table 2 and table 3 deliver the estimates of Spearman's r_s for each simulated single sample generated with both known underlying copula structure and known underlying r_s value in addition to the percentage of censoring. By comparing the true underlying r_s value with the estimates, we could have a general idea to evaluate how the four proposed estimation procedures perform when the underlying copula strucuture are Normal, Frank, Clayton and Gumbel. In other words, if the data set we got had one of this four copula structures after model selection and testing, we could know how the corresponding proposed estimation procedure would perform by taking the percentage of censoring into account. After observing table 2 and table 3 together, we found that our Normal copula-based estimation procedure always gave a downward estimate. By looking at table 2 alone, it is hard to tell how the proposed estimation procedures perform. However, when the sample size increased to 5 000 time pairs (table 3), there is a trend that if the copula assumption is just the same as the underlying copula structure, the estimate using that copula estimation procedure would be the most accurate one. This situation is true but there were a few exceptions happened because we only conduct one single sample. Table 3 illustrates that when the underlying copula structure is Frank, Clayton or Gumbel, normal-based copula estimation procedure would never give a better estimate compared with the other three. In addition, normal-based copula estimation procedure took much more time to get the estimate compared with the other three copula-based estimation procedures.

Table 4 illustrates the results for the estimates of Spearman's r_s by maximum likelihood estimators using three estimation procedures and four different types of underlying copulas based on 1 000 samples of sample size 50 and 200 under the various censoring conditions. It's ideal to get the data without censoring (i.e. 0% censoring). In this case, no censored information could be fog the true underlying copula structure. Our Frank-based estimation procedure perfoms well for data from Frank copula except when the Spearman's r_s is around zero, because under this condition, the estimate we got was upward biased. For data from Clayton copula, Clayton-based estimation procedure produced upward biased estimates no matter what the true underlying Spearman's r_s were. Gumbel-based estimation procedure is more likely to get downward biased estimates when the Spearman's r_s is around zero for the data from Gumbel copula. There is a trend that when the true underlying Spearman's r_s increase, the mean squared error (MSE) decreases under the condition that we choose

the right copula-based estimation procedure which means the copula assumption we made is the same as the true one. The same thing happens when the percentage of censoring is 30%. This illustrates that copula-based estimation procedure performs well especially when a strong positive association exists (tables 5,table 6).

Table 4 also demonstrates the results under the condition that the censoring percentage is 30%. Frank-based estimation procedure tent to have downward biased estimates for Frank-distributed data except when the underlying Spearman's r_s is around zero. Gumbel-based estimation procedure is less sensible to the censoring and could perform well under 30%.

Table 1 Estimates of r_s by the Maximum Likelihood Copula Estimators for Three Studies

study	NCE	FCE	CCE	GCE
Diabetic retinopathy	0.26	0.37	0.22	0.42
95% CI	(0.12,0.59)	(0.18,0.55)	(0.08,0.36)	(0.21,0.60)
Infections under dialysis	0.26	0.34	0.32	0.34
95% CI	(−0.04,0.70)	(0.00,0.63)	(0.04,0.59)	(−0.25,0.55)

Note: NCE,FCE,CCE and GCE denote estimation assuming underlying Normal,Frank,Clayton and Gumbel copulas respectively. CI denotes confidence interval. All our 95% confidence interval estimates are percentile interval based on 1 000 bootstrap samples.

Table 2 Estimates of r_s by Maximum Likelihood Estimators Assuming Various Copulas Using a Single Sample of n = 1 000

Underlying Copula	Underlying r_s	0% censoring				30% censoring			
		NCE	FCE	CCE	GCE	NCE	FCE	CCE	GCE
Normal	0.000	−0.021	0.070	0.053	0.050	−0.047	0.045	0.053	0.044
	0.300	0.265	0.363	0.361	0.316	0.230	0.321	0.327	0.331
	0.600	0.562	0.668	0.648	0.617	0.545	0.640	0.584	0.665
Frank	0.000	−0.077	0.020	−0.002	−0.006	−0.085	0.011	0.005	−0.003
	0.300	0.163	0.284	0.293	0.220	0.148	0.255	0.291	0.246
	0.600	0.442	0.593	0.621	0.504	0.436	0.561	0.562	0.566
Clayton	0.000	−0.025	0.060	0.061	0.032	−0.030	0.048	0.055	0.041
	0.300	0.248	0.337	0.338	0.270	0.267	0.321	0.322	0.330
	0.600	0.509	0.631	0.623	0.534	0.569	0.623	0.605	0.654
Gumbel	0.000	−0.032	0.063	0.088	0.033	−0.052	0.039	0.121	0.015
	0.300	0.224	0.323	0.347	0.291	0.203	0.294	0.322	0.308
	0.600	0.502	0.615	0.613	0.584	0.452	0.550	0.533	0.576

Table 3 Estimates of r_s by Maximum Likelihood Estimators Assuming Various Copulas Using a Single Sample of n = 5 000

Underlying Copula	r_s	0% censoring				30% censoring			
		NCE	FCE	CCE	GCE	NCE	FCE	CCE	GCE
Normal	0.000	0.008	0.099	0.098	0.073	0.001	0.088	0.101	0.095
	0.300	0.289	0.387	0.401	0.341	0.275	0.362	0.368	0.380
	0.600	0.578	0.684	0.661	0.634	0.567	0.658	0.604	0.685
Frank	0.000	−0.054	0.042	0.030	0.031	−0.065	0.029	0.030	0.028
	0.300	0.181	0.304	0.323	0.246	0.174	0.281	0.284	0.282
	0.600	0.455	0.604	0.641	0.518	0.456	0.581	0.571	0.589
Clayton	0.000	−0.049	0.037	0.040	0.027	−0.056	0.026	0.037	0.024
	0.300	0.222	0.311	0.313	0.250	0.243	0.302	0.304	0.314
	0.600	0.494	0.613	0.606	0.516	0.552	0.610	0.591	0.640
Gumbel	0.000	−0.067	0.022	0.025	0.017	−0.071	0.015	0.041	0.014
	0.300	0.212	0.304	0.290	0.302	0.167	0.264	0.242	0.285
	0.600	0.500	0.612	0.603	0.601	0.451	0.564	0.508	0.590

Table 4　Point Estimates and 95% Confidence Interval Assuming Three Copulas under Various Percentage of Censoring

		0% censoring $n=50$			30% censoring $n=50$		
	r_s	FCE	CCE	GCE	FCE	CCE	GCE
NC	0.00	0.10(0.10,0.11)	0.19(0.18,0.20)	0.05(0.04,0.06)	0.10(0.09,0.11)	0.22(0.21,0.24)	0.09(0.08,0.10)
	0.30	0.38(0.38,0.39)	0.45(0.44,0.46)	0.34(0.33,0.35)	0.37(0.36,0.38)	0.44(0.43,0.45)	0.39(0.38,0.40)
	0.60	0.68(0.67,0.68)	0.69(0.69,0.70)	0.63(0.62,0.64)	0.66(0.66,0.67)	0.66(0.65,0.66)	0.69(0.69,0.70)
FC	0.00	0.04(0.03,0.05)	0.15(0.14,0.16)	−0.01(−0.02,0.01)	0.04(0.03,0.05)	0.20(0.19,0.21)	0.03(0.02,0.04)
	0.30	0.30(0.29,0.31)	0.37(0.36,0.38)	0.25(0.24,0.26)	0.29(0.28,0.30)	0.37(0.36,0.38)	0.30(0.29,0.31)
	0.60	0.60(0.59,0.60)	0.63(0.63,0.64)	0.53(0.52,0.53)	0.58(0.58,0.59)	0.59(0.58,0.60)	0.60(0.60,0.61)
CC	0.00	0.04(0.03,0.05)	0.09(0.09,0.10)	−0.02(−0.03,−0.01)	0.04(0.03,0.05)	0.10(0.09,0.10)	0.02(0.01,0.04)
	0.30	0.31(0.30,0.32)	0.34(0.33,0.34)	0.24(0.23,0.24)	0.31(0.30,0.32)	0.33(0.32,0.34)	0.31(0.30,0.33)
	0.60	0.60(0.60,0.61)	0.61(0.61,0.62)	0.51(0.50,0.52)	0.62(0.61,0.62)	0.60(0.60,0.61)	0.64(0.63,0.65)
GC	0.00	0.02(0.01,0.03)	0.13(0.12,0.14)	−0.03(−0.05,−0.02)	0.01(0.00,0.02)	0.18(0.17,0.19)	0.00(−0.01,0.01)
	0.30	0.31(0.30,0.31)	0.37(0.35,0.38)	0.30(0.29,0.31)	0.28(0.27,0.29)	0.35(0.34,0.36)	0.31(0.30,0.32)
	0.60	0.60(0.60,0.61)	0.63(0.63,0.64)	0.59(0.58,0.60)	0.57(0.56,0.57)	0.57(0.56,0.58)	0.60(0.59,0.61)
		0% censoring $n=200$			30% censoring $n=200$		
	r_s	FCE	CCE	GCE	FCE	CCE	GCE
NC	0.00	0.10(0.09,0.10)	0.11(0.11,0.12)	0.07(0.07,0.07)	0.09(0.09,0.10)	0.12(0.11,0.13)	0.09(0.09,0.10)
	0.30	0.39(0.38,0.39)	0.42(0.42,0.42)	0.34(0.34,0.34)	0.37(0.36,0.37)	0.39(0.39,0.40)	0.38(0.38,0.39)
	0.60	0.68(0.68,0.68)	0.67(0.67,0.68)	0.63(0.63,0.64)	0.66(0.66,0.67)	0.62(0.62,0.63)	0.69(0.69,0.69)
FC	0.00	0.04(0.04,0.04)	0.05(0.05,0.06)	0.01(0.01,0.02)	0.04(0.03,0.04)	0.07(0.06,0.07)	0.04(0.03,0.04)
	0.30	0.30(0.30,0.31)	0.33(0.33,0.34)	0.24(0.24,0.25)	0.29(0.28,0.29)	0.32(0.31,0.32)	0.29(0.28,0.29)
	0.60	0.60(0.60,0.60)	0.62(0.62,0.62)	0.52(0.52,0.52)	0.58(0.58,0.58)	0.57(0.56,0.57)	0.59(0.59,0.60)
CC	0.00	0.04(0.03,0.04)	0.04(0.04,0.05)	0.01(0.01,0.02)	0.04(0.04,0.04)	0.04(0.04,0.05)	0.03(0.03,0.04)
	0.30	0.31(0.31,0.31)	0.32(0.31,0.32)	0.24(0.24,0.25)	0.31(0.31,0.32)	0.31(0.31,0.31)	0.32(0.31,0.32)
	0.60	0.61(0.61,0.61)	0.60(0.60,0.61)	0.51(0.51,0.51)	0.62(0.61,0.62)	0.59(0.59,0.59)	0.64(0.64,0.64)
GC	0.00	0.01(0.01,0.02)	0.03(0.03,0.04)	0.00(−0.00,0.00)	0.01(0.01,0.02)	0.05(0.04,0.05)	0.01(0.01,0.02)
	0.30	0.31(0.30,0.31)	0.32(0.31,0.32)	0.30(0.30,0.31)	0.28(0.27,0.28)	0.28(0.27,0.28)	0.30(0.29,0.30)
	0.60	0.61(0.61,0.61)	0.61(0.61,0.62)	0.60(0.60,0.60)	0.57(0.57,0.57)	0.54(0.53,0.54)	0.60(0.59,0.60)

Table 5　Mean Squared Error (MSE) Assuming Three Copula Types

Underlying Copula	r_s	0% Censoring			30% Censoring		
		FCE	CCE	GCE	FCE	CCE	GCE
Normal	0.0	0.029	0.061	0.027	0.032	0.089	0.034
	0.3	0.022	0.042	0.015	0.023	0.045	0.025
	0.6	0.012	0.017	0.007	0.012	0.015	0.014
Frank	0.0	0.024	0.049	0.028	0.027	0.074	0.029
	0.3	0.019	0.032	0.020	0.022	0.038	0.022
	0.6	0.010	0.013	0.014	0.014	0.017	0.011
Clayton	0.0	0.023	0.021	0.031	0.027	0.022	0.029
	0.3	0.019	0.016	0.023	0.023	0.017	0.025
	0.6	0.011	0.008	0.018	0.013	0.009	0.014
Gumbel	0.0	0.022	0.038	0.033	0.026	0.069	0.032
	0.3	0.020	0.032	0.018	0.024	0.037	0.022
	0.6	0.010	0.015	0.010	0.015	0.021	0.011

4　Conclusion

From our simulation study, our normal-based estimation procedure always tends to have downward biased estimates when the underlying copula structure is Normal, Clayton, Gumbel and Frank. When the percentage of censoring increases to 30%, all the four estimation procedures tend to get less accurate estimates comparing to the setting of 0% censoring when the underlying copula structure are Normal, Clayton, Frank and Gumebl. However, Gumbel-based estimation procedure is much more robust for data from Gumbel copula in terms of percentage of censoring. Finally, we also

found that Frank, Gumebl, and Clayton-based estimation procedures perfom better with the increase of a positive association under the condition that we choose the right copula-based estimation procedure which means the copula assumption we made is the same as the true one.

Table 6 Mean Squared Error (MSE) Assuming Three Copula Types

Underlying Copula	r_s	0% Censoring			30% Censoring		
		FCE	CCE	GCE	FCE	CCE	GCE
Normal	0.0	0.014	0.020	0.009	0.014	0.025	0.013
	0.3	0.011	0.019	0.005	0.009	0.015	0.010
	0.6	0.008	0.007	0.003	0.006	0.004	0.009
Frank	0.0	0.007	0.009	0.005	0.007	0.013	0.007
	0.3	0.004	0.008	0.007	0.005	0.010	0.005
	0.6	0.002	0.004	0.009	0.003	0.006	0.003
Clayton	0.0	0.007	0.005	0.005	0.007	0.005	0.006
	0.3	0.004	0.004	0.007	0.005	0.004	0.006
	0.6	0.002	0.002	0.010	0.003	0.002	0.004
Gumbel	0.0	0.005	0.007	0.005	0.006	0.010	0.006
	0.3	0.005	0.008	0.004	0.006	0.011	0.005
	0.6	0.003	0.005	0.002	0.004	0.010	0.003

5 Discussion

In the real application, it would not be appropriate to use our four suggested copula based estimation procedures to get the estimate with knowning that the data comes from another copula distribution rather than Normal, Clayton, Gumbel and Frank. Thus a nonparametric indentification of the copula structure is in demand [22].

Moreover, when the sample size is small, discordances between two pairs of time period will appear even though the true relationship is concordant because of the variability of the observations obtained from a continuous distribution. Under this condition, only Pearson´s correlation coefficient can fully make use of this information. Relative statistical power got from permutation test can be used to illustrate the ability to show the degree of discordances [20].

For the simulation part, we only considered the setting of low-moderate percentage of censoring (i. e. 0%, 30%) of the total survival times. And we didn´t study in details whether the first part of the bivariate survival times contains more censored observations. It would be valuable to test other patterns of censoring percentage and to study further and test with different distribution of the censored observations among the two time periods as well as the performace of those procedures under relatively high percentage of censoring.

All the data used in our simulation study were generated from a bivariate distribution with unit exponential margins and four types of copula respectively. In further studies, we could examine data generated from various other types of marginal distributions such as the Weibull, log-logistic, gamma and log-normal in order to get a more generalized conclusion.

For the normal based copula, estimation takes much more time than it does for the other three. Finding ways to increase the efficiency of the calculation and estimation would be an important topic to explore in future research.

Furthermore, it would be highly interesting to study whether (T_1, Q_1) could be used to predict (T_2, Q_2) instead of just knowing the correlation between the two.

6 Reference

[1] Haddad R I, Shin D M. Recent advances in head and neck cancer [J]. New England Journal of Medicine, 2008, 359(11): 1143-1154.

[2] Kendall M G, Gibbons J D. Rank correlation methods [M]. Oxford: Oxford University Press, 1990.

[3] Schemper M, Kaider A, Wakounig S, et al. Estimating the correlation of bivariate failure times under censoring [J]. Stat Med, 2013, 32(27): 4781-4790.

[4] Shih J H,Louis T A. Inferences on the association parameter in copula models for bivariate survival data [J]. Biometrics,1995,51(4):1384-1399.
[5] Liebetrau A M. Measures of association [M]. Newbury Park,London,New Belhi: Sage Publications,Inc,1983: 44.
[6] Lai C D,Balakrishnan N. Continuous bivariate distributions [M]. 2nd edition. New York: Springer,2009.
[7] Sklar M. Fonctions de repartition an dimensions et leurs marges [J]. Publ Inst Statist Univ Paris,1959(8):229-231.
[8] Nelsen R B. An introduction to copulas [M]. New York: Springer-Verlag,2006: 2.
[9] Aas,Kjersti. Modelling the dependence structure of financial assets: a survey of four copulas [J]. Samba,2004,22(4):1-18.
[10] Huster W J,Brookmeyer R,Self S G. Modelling paired survival data with covariates [J]. Biometrics,1989,45(1): 145-156.
[11] McGilchrist C A,Aisbett C W. Regression with frailty in survival analysis [J]. Biometrics,1991,47(2):461-466.
[12] Collett D. Modelling survival data in medical research [M]. Ohio: Chapman and Hall/CRC,2003: 12.
[13] Bélisle C J P. Convergence theorems for a class of simulated annealing algorithms on $\mathbf{R}^d$ [J]. J Appl Probab,1992,29(4):885-895.
[14] Broyden C G. The convergence of a class of double-rank minimization algorithms: 1. general considerations [J]. IMA J Appl Math,1970,6(1):76-90.
[15] Fletcher R. A new approach to variable metric algorithms [J]. Comput J,1970,13(3):317-322.
[16] Goldfarb D. A family of variable-metric methods derived by variational means [J]. Math Comput,1970,24(109):23-26.
[17] Efron B,Tibshirani R J. An introduction to the bootstrap [M]. Ohio: Chapman and Hall/CRC,1998: 153.
[18] Wagenmakers E J,Farrell S. AIC model selection using Akaike weights [J]. Psychon Bull Rev,2004,11(1):192-196.
[19] Akaike H. A new look at the statistical model identification [J]. IEEE Trans Autom Control,1974,19(6):716-723.
[20] Yan Jun. Enjoy the joy of copulas: with a package copula [J]. Journal of Statistical Software,2007,21(4):1-21.
[21] Casella G,Berger R L. Statistical inference [M]. Mason,OH: Cengage Learning,2001: 417.
[22] Li Bo,Genton M G. Nonparametric identification of copula structures [J]. J Am Stat Assoc,2013,108(502):666-675.

双变量生存数据之间的关联评估
——一种基于 Copula 的方法

韩晓珍[1],付平福[2,3*]

(1. 克利夫兰诊所定量健康科学系,俄亥俄　44195,美国; 2. 凯斯西储大学人口与定量健康科学系,俄亥俄　44106,美国; 3. 凯斯西储大学综合癌症中心,俄亥俄　44106,美国)

摘要: 生存时间之间的关联分析引起许多从事生物与医学领域研究者的兴趣. 这种分析的目的是利用 Coupla 方法调查在蒙特卡罗适度审查背景下,双变量生存数据之间的关联. 该文利用皮尔森关联系数去估计双变量失败数据之间的关联. 研究结果表明: 当审查百分比低时,基于 Gubel 的估计方法更为鲁棒,而且正关联越强,分别用审查百分比是 0% 和 30% 所估计的结果越精确. 这对基于 Frank,Gumbel 和 Clayton 的估计方法是正确的,甚至在 Copula 假设条件下真实情形也成立.

关键词: 双变量生存数据; Coupla 方法; 关联分析; 关联系数; 审查百分比

(责任编辑: 王金莲)