A goodness-of-fit test for semi-parametric copula models for bivariate censored data

Jimin Shin

Follow this and additional works at: https://scholarsjunction.msstate.edu/td

Recommended Citation
https://scholarsjunction.msstate.edu/td/152

This Graduate Thesis is brought to you for free and open access by the Theses and Dissertations at Scholars Junction. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholars Junction. For more information, please contact scholcomm@msstate.libanswers.com.
A goodness-of-fit test for semi-parametric copula models for bivariate censored data

By

Jimin Shin

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Statistics
in the Department of Mathematics and Statistics

Mississippi State, Mississippi

August 2020
A goodness-of-fit test for semi-parametric copula models for bivariate censored data

By

Jimin Shin

Approved:

Qian Zhou  
(Major Professor)

Prakash N. Patil  
(Committee Member)

Tung-Lung Wu  
(Committee Member)

Mohammad Sepherifar  
(Graduate Coordinator)

Rick Travis  
Dean  
College of Arts & Sciences
In this thesis, we suggest a goodness-of-fit test for semi-parametric copula models. We extended the pseudo in-and-out-sample (PIOS) test proposed in [17], which is based on the PIOS test in [28]. The PIOS test is constructed by comparing the pseudo “in-sample” likelihood and pseudo “out-of-sample” likelihood. Our contribution is two-fold. First, we use the approximate test statistics instead of the exact test statistics to alleviate the computational burden of calculating the test statistics. Secondly, we propose a parametric bootstrap procedure to approximate the distribution of the test statistic. Unlike the nonparametric bootstrap which resamples from the original data, the parametric procedure resamples the data from the copula model under the null hypothesis. We conduct simulation studies to investigate the performance of the approximate test statistic and parametric bootstrap. The results show that the parametric bootstrap presents higher test power with a well-controlled type I error compared to the nonparametric bootstrap.
Key words: goodness-of-fit test, copula model, bivariate distribution, pseudo in-and-out-of-sample likelihood
ACKNOWLEDGEMENTS

First and foremost, I would like to deeply thank my advisor, Dr. Qian Zhou. Without her encouragement and inspiration, I would not be able to finish my Master’s program successfully.

I am grateful to Dr. Prakash Patil and Dr. Tung-Lung for serving on my examining committee and their kind suggestions and comments.

I would like to express my gratitude to the professors and graduate students with whom I have taken courses.

Last but not least, I would like to give my heartfelt appreciation to my family and friends for supporting me with kind advice and encouragement.
CONTENTS

ACKNOWLEDGEMENTS .............................................................. ii

LIST OF TABLES ........................................................................ v

LIST OF FIGURES ..................................................................... vi

CHAPTER

1. INTRODUCTION ................................................................. 1
   1.1 Copula Families For Modeling Multivariate Survival Time ...... 1
   1.2 Literature Review of Testing Copula ................................. 7
   1.3 Literature Review of In-and-Out-of-Sample Test ................. 8
   1.4 Contribution ................................................................. 8
   1.5 Organization ............................................................... 9

2. IN-AND-OUT-OF-SAMPLE PSEUDO LIKELIHOOD RATIO TEST .. 10
   2.1 Notation ..................................................................... 10
   2.2 Likelihood Function ...................................................... 11
   2.3 Two-Step Pseudo Maximum Likelihood Estimation on Dependence Parameter .................................................. 12
   2.4 Pseudo In-and-Out-of-Sample Likelihood Ratio Test .......... 12
      2.4.1 Approximation on PIOS Test Statistic ....................... 13
      2.4.2 Parametric Bootstrap on PIOS Test Statistic ............... 13

3. SIMULATION ........................................................................ 16
   3.1 Data Generation .......................................................... 16
   3.2 Exact and Approximate PIOS Test Statistics ..................... 17
   3.3 Nonparametric and Parametric Bootstrap ......................... 21
      3.3.1 Result .................................................................. 21

4. DISCUSSION ......................................................................... 24
LIST OF TABLES

3.1 Average of the exact PIOS test statistic $T$ and the approximate PIOS test statistic $\tilde{T}$ .......................................................... 18

3.2 Average computing time (in seconds) of calculating the exact PIOS test statistic $T$ and the approximate PIOS test statistic $\tilde{T}$ .......................................................... 18

3.3 Empirical standard deviation of $T$ and the average standard error obtained from nonparametric and parametric bootstrap ................................................. 23

3.4 The proportions of rejecting the null hypothesis at the significance level 5% with nonparametric and parametric bootstrap ................................................. 23
LIST OF FIGURES

1.1 Scatter plots of copula models with $\tau = \rho = 0.4$ ........................................ 5

1.2 Scatter plots of copula models with $\tau = \rho = 0.8$ ........................................ 6

3.1 Distributions of the exact PIOS test statistic $T$ and the approximate PIOS test statistic $\tilde{T}$ ................................................................. 19

3.2 Quantile-quantile plot of the exact PIOS test statistic $T$ and the approximate PIOS test statistic $\tilde{T}$ ................................................................. 20
CHAPTER 1

INTRODUCTION

Copula models have been widely used for analyzing the dependence among two or more random variables. Let $T = (T_1, \ldots, T_d)'$ be a vector of event times, and let $S_i$ be the marginal distribution of univariate event time, i.e. $S_i(t) = Pr(T_i > t)$, and let $S$ be the joint distribution of the multivariate event times vector, i.e. $S(t_1, \cdots, t_d) = Pr(T_1 > t_1, \cdots, T_d > t_d)$. According to Sklar’s theorem [25], there exists a copula function $C$ on $[0, 1]^d$, such that

$$S(t_1, \ldots, t_d) = C(S_1(t_1), \ldots, S_d(t_d); \theta),$$

where the parameter $\theta$ in the copula function $C$ is called the copula parameter or dependence parameter, which captures the dependence among the multivariate event times. In the thesis, we will focus on bivariate event times, i.e., $d = 2$.

1.1 Copula Families For Modeling Multivariate Survival Time

Archimedean copulas [19] are the most widely used family for modeling multivariate survival data. Archimedean copulas include several copula families, such as Clayton, Frank, and Gumbel, which exhibit different dependence properties. Each of these families is derived from a so-called generator function, which leads to an explicit expression. This allows the likelihood-based estimation procedure to be feasible. Also, it has been shown
that Archimedean copulas are closely related to another popular approach, proportional frailty models. This approach uses a latent variable, called frailty, to induce the dependence among the multivariate event times.

In the following paragraphs, we describe the expression of Clayton, Frank, and Gumbel copulas. As mentioned earlier, each of these families is derived from a so-called generator function which is denoted by \( \varphi \). The \( \varphi \) function is a continuous, strictly decreasing function from \([0, 1]\) to \([0, \infty)\) with \( \varphi(1) = 0 \). The pseudo-inverse of \( \varphi \) is defined as \( \varphi^{-1} \): \([0, \infty) \rightarrow [0, 1] \). It is defined as

\[
\varphi^{-1}(t) = \begin{cases} 
\varphi^{-1}(t), & 0 \leq t \leq \varphi(0) \\
0, & \varphi(0) \leq t \leq \infty.
\end{cases}
\]

The Archimedean copula function \( C \) is defined as

\[
C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)). \tag{1.2}
\]

Another popular dependence measure for bivariate random variables is Kendall’s rank coefficient, also called Kendall’s \( \tau \). Let \((T_{i1}, T_{i2})\) and \((T_{j1}, T_{j2})\) be the bivariate event times of two randomly selected subjects. The Kendall’s \( \tau \) is defined as

\[
\tau = Pr[(T_{i1} - T_{j1})(T_{i2} - T_{j2}) > 0] - Pr[(T_{i1} - T_{j1})(T_{i2} - T_{j2}) < 0], \tag{1.3}
\]

which is a measure of concordance for these two event times. For each of the following copula families, Kendall’s \( \tau \) can be expressed as a function of the copula parameter \( \theta \).

**Clayton copula** [2] The Clayton copula is given by

\[
C_{\theta}(u, v) = \left[ \max \left( u^{-\theta} + v^{-\theta} - 1, 0 \right) \right]^{-1/\theta}, \tag{1.4}
\]
where the generator function is \( \varphi_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1) \) with \( \theta \in [-1, \infty) \setminus \{0\} \). For \( \theta > 0 \), the copula is a strict Archimedean copula. As \( \theta \to 0 \), \( T_1 \) and \( T_2 \) are independent, and as \( \theta \to \infty \), they are co-monotonic. Therefore, the Clayton copula interpolates between independence and perfectly positive dependence and represents lower tail dependency. Finally, the Kendall’s \( \tau \) for the Clayton copula is given as \( \tau = \theta/(\theta + 2) \).

**Frank copula** [7] The Frank copula is given by

\[
C_\theta(u, v) = -\frac{1}{\theta} \ln \left[ 1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right],
\]

where the generator function is \( \varphi_\theta(t) = -\ln(e^{-\theta t} - 1)/(e^{-\theta} - 1) \) with \( \theta \neq 0 \). The Frank copula is a strict Archimedean copula for all \( \theta \). Like the Clayton copula, the Frank copula is independent as \( \theta \to 0 \) and co-monotonic as \( \theta \to \infty \). When \( \theta \to -\infty \), the Frank copula is so-called counter-monotonic. Therefore, the Frank copula interpolates between perfectly positive dependence and perfectly negative dependence. Lastly, the Kendall’s \( \tau \) for the Frank copula is \( \tau = 1 - 4\theta^{-1}(1 - D(\theta)) \), where \( D(\theta) = \frac{1}{\theta} \int_0^\infty \frac{t}{e^t - 1} \, dt \).

**Gumbel copula** [11] The Gumbel copula is given by

\[
C_\theta(u, v) = \exp \left( - \left[ (\ln u)^\theta + (\ln v)^\theta \right]^{-1/\theta} \right),
\]

where the generator function is \( \varphi_\theta(t) = (\ln t)^\theta \) with \( \theta \in [1, \infty) \). Like the Frank copula, the Gumbel copula is a strict Archimedean copula for all \( \theta \). When \( \theta = 1 \), the Gumbel copula is independent. As \( \theta \to \infty \), the copula is co-monotonic. Therefore, the Gumbel copula
interpolates between independence and perfectly positive dependence and represents upper tail dependency. The Kendall’s $\tau$ for the Gumbel copula is given as $\tau = 1 - 1/\theta$.

While the Archimedean copula is the most commonly used for modeling the multivariate survival data, the Gaussian copula, also called Normal copula, is considered in the literature as well, such as [14] and [20]. The Gaussian copula is defined as

$$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)), \quad (1.7)$$

where $\Phi(x)$ is the standard univariate Gaussian distribution, and $\Phi_\rho$ is the joint cdf of bivariate normal random variables with mean $(0, 0)$ and covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ where the correlation parameter $\rho \in [-1, 1]$. When $\rho = 0$, it corresponds to the independence structure. It can be expressed as

$$\Phi_\rho(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \frac{1}{2\pi \sqrt{1-\rho^2}} \exp \left\{ \frac{2\rho st - s^2 - t^2}{2(1-\rho^2)} \right\} d\!s d\!t. \quad (1.8)$$

The Kendall’s tau for the Gaussian copula is given as $\tau = \frac{2}{\pi} \arcsin \rho$.

Figure 1.1 and Figure 1.2 are the scatter plots of bivariate random variables generated from each of the four copula families with two different dependence strength. For the Clayton, Frank, and Gumbel copulas, the Kendall’s $\tau$ is 0.4 in Figure 1.1 and 0.8 in Figure 1.2; for Gaussian copula, the correlation parameter $\rho$ is 0.4 in Figure 1.1 and 0.8 in Figure 1.2. We can see that Clayton copula shows lower tail dependence whereas Gumbel copula has upper tail dependence. On the other hand, Frank copula has no dependence on both tails, but Gaussian copula shows the symmetric tail dependence. Furthermore, the tail dependence gets stronger with a higher value of the copula parameter.
Figure 1.1

Scatter plots of copula models with $\tau = \rho = 0.4$
Figure 1.2

Scatter plots of copula models with $\tau = \rho = 0.8$. 
1.2 Literature Review of Testing Copula

Many researchers conducted goodness-of-fit tests for copulas so far. For example, [5] suggested the bias-corrected goodness-of-fit test with a fixed smoothing parameter. [16] tested for Gaussian copula using $\chi^2$-distribution. [6] proposed two distribution-free goodness-of-fit test statistics for copulas. [22] extended the goodness-of-fit test in [6] by holding the smoothing parameters fixed. [3] proposed the goodness-of-fit test for copulas based on Rosenblatt’s transformation. However, the test only performs well when the marginal distribution functions are known. [18] proposed the goodness-of-fit test based on the theoretical and sample versions of Spearman’s dependence function. [8] suggested the goodness-of-fit test based on a Cramér-von Mises statistics. Lastly, [12] state the rank-based goodness-of-fit test for copulas using the information matrix. Finally, other works on the goodness-of-fit test for copula models have reviewed in [10].

The literature above are based on the fully observed data. However, the event times are often subject to censoring. Various researchers tested the goodness-of-fit tests for copulas with the censored event data. [23] proposed the goodness-of-fit test for bivariate survival model with Clayton copula. [27] studied the model specification procedure for bivariate survival models for right-censored data generated by the Archimedean copula. [9] extended the procedure in [27] by using a non-truncated version of Kendall’s process. [1] proposed a class of goodness-of-fit tests for bivariate-right censored data with unspecified association parameters. [4] extended the idea in [23] to the general Archimedean copula.
1.3 Literature Review of In-and-Out-of-Sample Test

In this paper, we use the pseudo in-and-out-of-sample (PIOS) likelihood ratio for testing the model misspecification. The in-and-out-of-sample (IOS) test is first proposed in [21]. Before [21], only the out-of-sample likelihoods were used as model selection criteria in several studies, such as [26] and [15]. [28] extended the IOS test in [21] to the PIOS test. Also, [29] extended the test to the pseudo-in-and-out-of-likelihood (PIOL) test. The test enables the application to the univariate time series.

1.4 Contribution

A prominent challenge in implementing the PIOS test is computation time. When one sample is deleted each time to construct the out-of-sample likelihood, the copula parameter needs to be re-estimated via the maximization procedure, which is time-consuming, especially for a large sample size. To overcome this obstacle, we consider an approximation method that requires the estimation of the copula parameter only once, which significantly reduces computation time. The simulation results in [17] indicate somewhat low test power, which is another challenge in implementing the PIOS test. He used bootstrap, a resampling technique, to obtain the \( p \)-value. This procedure approximates the distribution of the test statistic by resampling the data. Mei’s method is a nonparametric bootstrap procedure which does not utilize any model assumption. Here, we investigate a parametric bootstrap procedure that resamples the data under the null hypothesis.
1.5 Organization

This paper is organized as follows. In Chapter 2, we will first describe the likelihood function of the copula model with right-censored bivariate event times. Then, we will describe the procedure of estimating the marginal distribution and copula parameter. Next, we will present the PIOS test statistic and the approximation method. In the end, we will describe the nonparametric and parametric bootstrap procedures for obtaining the empirical $p$-values. In Chapter 3, we will study the performance of the approximate PIOS test statistics and compare the nonparametric and parametric bootstrap procedures through the simulation study. In the last chapter, we will state the conclusion including the discussion and future work.
CHAPTER 2
IN-AND-OUT-OF-SAMPLE PSEUDO LIKELIHOOD RATIO TEST

2.1 Notation

Let $T_{i1}$ and $T_{i2}$ be the bivariate event times for the $i$th subject, and $C_i$ be the censoring time, where $i = 1, 2, \ldots, n$. The censoring time $C_i$ is independent of the bivariate event times $T_{i1}$ and $T_{i2}$. Also, $X_{i1}$ and $X_{i2}$ are the observed times and it can be expressed by $X_{ij} = \min \{T_{ij}, C_i\}$, where $j = 1, 2$. Lastly, let $\delta_{ij} = I(T_{ij} \leq C_i)$ be the censoring indicator variable, where $I(\cdot)$ is an indicator function. Let $\mathcal{D}_i = \{(X_{i1}, X_{i2}, \delta_{i1}, \delta_{i2})\}$ denote the available data for subject $i$.

The underlying true copula function is denoted as $C_0$, i.e.,

$$C_0(S_1(t_1), S_2(t_2); \theta) = S(t_1, t_2) = P(T_{i1} > t_1, T_{i2} > t_2).$$

In practice, the true copula function is rarely known, so we use a copula family $C(U_1, U_2; \theta)$ with the dependence parameter $\theta$ to model the bivariate event times.

Now we want to test whether the underlying true copula belongs to the “working” copula family, i.e.,

$$H_0 : C_0 \in C = \{C(\cdot; \theta), \theta \in \Theta\} \quad \text{versus} \quad H_1 : C_0 \notin C = \{C(\cdot; \theta), \theta \in \Theta\}$$

where $\Theta \subset \mathbb{R}$ is the parameter space. Here, we focus on copula families such as Archimedean and Gaussian with only one copula parameter.
2.2 Likelihood Function

- Neither $T_{i1}$ nor $T_{i2}$ are censored, i.e., $(\delta_{i1} = 1$ and $\delta_{i2} = 1)$. The corresponding likelihood component is $\partial^2 S(X_{i1}, X_{i2})/\partial X_{i1}\partial X_{i2}$.

- $T_{i1}$ is censored, but $T_{i2}$ is observed, i.e., $(\delta_{i1} = 0$ and $\delta_{i2} = 1)$. The corresponding likelihood component is $-\partial S(X_{i1}, X_{i2})/\partial X_{i2}$.

- $T_{i2}$ is censored, but $T_{i1}$ is observed, i.e., $(\delta_{i1} = 1$ and $\delta_{i2} = 0)$. The corresponding likelihood component is $-\partial S(X_{i1}, X_{i2})/\partial X_{i1}$.

- Both $T_{i1}$ and $T_{i2}$ are censored, i.e., $(\delta_{i1} = 0$ and $\delta_{i2} = 0)$. The corresponding likelihood component is $S(X_{i1}, X_{i2})$.

Therefore, the full log-likelihood function is expressed as $l(\theta) = \sum_{i=1}^{n} l(\theta; D_i)$. Thus, the $i$-th component of the log-likelihood function is expressed as

$$l(\theta; D_i) = \delta_{i1}\delta_{i2} \log \left\{ \frac{\partial^2 S(X_{i1}, X_{i2})}{\partial X_{i1}\partial X_{i2}} \right\} + (1 - \delta_{i1})\delta_{i2} \log \left\{ -\frac{\partial S(X_{i1}, X_{i2})}{\partial X_{i2}} \right\}$$

$$+ (1 - \delta_{i2})\delta_{i1} \log \left\{ -\frac{\partial S(X_{i1}, X_{i2})}{\partial X_{i1}} \right\} + (1 - \delta_{i1})(1 - \delta_{i2}) \log \left\{ S(X_{i1}, X_{i2}) \right\}.$$

(2.1)

It can be rewritten as

$$l(\theta; D_i) = l(\theta; U_{i1}, U_{i2})$$

$$= \delta_{i1}\delta_{i2} \log \left\{ c_{\theta}(U_{i1}, U_{i2}) f_1(X_{i1}) f_2(X_{i2}) \right\} + (1 - \delta_{i1})\delta_{i2} \log \left\{ \frac{\partial c_{\theta}(U_{i1}, U_{i2})}{\partial U_{i2}} f_2(X_{i2}) \right\}$$

$$+ (1 - \delta_{i2})\delta_{i1} \log \left\{ \frac{\partial^2 c_{\theta}(U_{i1}, U_{i2})}{\partial U_{i1}} f_1(X_{i1}) \right\} + (1 - \delta_{i1})(1 - \delta_{i2}) \log \left\{ c_{\theta}(U_{i1}, U_{i2}) \right\},$$

(2.2)

where $U_{i1} = S_1(X_{i1})$, $U_{i2} = S_2(X_{i2})$ for $i = 1, \ldots, n$, $c_{\theta}(u_1, u_2) = C(u_1, u_2; \theta)$, $c_{\theta}(u_1, u_2) = \partial^2 c_{\theta}(u_1, u_2)/\partial u_1 u_2$, $f_1(t_1) = -S_1'(t_1)$, and $f_2(t_2) = -S_2'(t_2)$. Note that the marginal probability density functions $f_1$ and $f_2$ do not contain the copula parameter, so we can exclude them in the likelihood function.
2.3 Two-Step Pseudo Maximum Likelihood Estimation on Dependence Parameter

The marginal distribution and copula parameter can be estimated separately following the two-step pseudo maximum likelihood estimation proposed in [24]. The first step is to estimate the survival functions of \( T_i^1 \) and \( T_i^2 \) nonparametrically by using Kaplan-Meier estimator [13]. The next step is to estimate \( \theta \) by maximizing pseudo likelihood function \( \sum_{i=1}^{n} l(\theta; \hat{U}_i^1, \hat{U}_i^2) \), where \( \hat{U}_i^1 \) and \( \hat{U}_i^2 \) are the estimated survival functions, i.e., \( \hat{U}_i^1 = \hat{S}(X_i^1) \) and \( \hat{U} = \hat{S}(X_i^2) \). The estimate of \( \theta \) is given by

\[
\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} l(\theta; \hat{U}_i^1, \hat{U}_i^2).
\]

Finally, \( \hat{\theta} \) is called as pseudo maximum likelihood estimates (PMLE).

2.4 Pseudo In-and-Out-of-Sample Likelihood Ratio Test

The pseudo in-and-out-of-sample (PIOS) likelihood ratio test statistic is defined as \( T = \hat{l}_\text{in} - \hat{l}_\text{out} \), where \( \hat{l}_\text{in} \) is the in-sample likelihood, and \( \hat{l}_\text{out} \) is the out-of-sample likelihood. The pseudo “in-sample” likelihood \( \hat{l}_\text{in} = \sum_{i=1}^{n} l(\hat{\theta}; \hat{U}_i^1, \hat{U}_i^2) \) is the log-likelihood function using \( \hat{\theta} \) with full data, where \( \hat{\theta} = \arg\max \sum_{i=1}^{n} l(\theta; \hat{U}_i^1, \hat{U}_i^2) \). The pseudo “out-of-sample” likelihood \( \hat{l}_\text{out} = \sum_{i=1}^{n} l(\hat{\theta}_{(-i)}; \hat{U}_i^1, \hat{U}_i^2) \) is the log-likelihood function using \( \hat{\theta}_{(-i)} \) obtained from the data with the \( i \)-th observation deleted for \( i = 1, \ldots, n \). Here, \( \hat{\theta}_{(-i)} = \arg\max \sum_{j \neq i} l(\theta; \hat{U}_j^1, \hat{U}_j^2) \). Finally, the PIOS test is defined as

\[
T = \hat{l}_\text{in} - \hat{l}_\text{out} = \sum_{i=1}^{n} \left\{ l(\hat{\theta}; \hat{U}_i^1, \hat{U}_i^2) - l(\hat{\theta}_{(-i)}; \hat{U}_i^1, \hat{U}_i^2) \right\}.
\] (2.3)
2.4.1 Approximation on PIOS Test Statistic

As seen in the previous section, each time a sample is deleted, the copula parameter needs to be re-estimated. The re-estimation process is time-consuming, especially when the sample size $n$ is large. Therefore, we implement the approximation method which is proposed in [29] to reduce the computing time of the PIOS by reducing the number of maximization step from $n$ times to one while obtaining $\hat{\theta}$. Let $\tilde{\theta}_{(-i)}$ be the approximated $\hat{\theta}_{(-i)}$, which is calculated by

$$\tilde{\theta}_{(-i)} = \hat{\theta} + \left[ \sum_{i=1}^{n} \frac{\partial^2 l(\hat{\theta}; \hat{U}_{i1}, \hat{U}_{i2})}{\partial \theta^2} \right]^{-1} \frac{\partial l(\hat{\theta}; \hat{U}_{i1}, \hat{U}_{i2})}{\partial \hat{\theta}}.$$ (2.4)

By using $\tilde{\theta}_{(-i)}$ instead of $\hat{\theta}_{(-i)}$ in the PIOS test (2.4), we have the approximated PIOS test statistic $\tilde{T}$, i.e.,

$$\tilde{T} = \sum_{i=1}^{n} \left\{ l(\hat{\theta}; \hat{U}_{i1}, \hat{U}_{i2}) - l(\tilde{\theta}_{(-i)}; \hat{U}_{i1}, \hat{U}_{i2}) \right\},$$ (2.5)

and in the remainder of the thesis, the PIOS test (2.4) is referred to as the exact PIOS test statistic.

2.4.2 Parametric Bootstrap on PIOS Test Statistic

According to [28], the PIOS test statistic $T$ converges in probability to the dimension of the parameter vector of $\theta$ when the null hypothesis is true. All copula models considered in this thesis have a dependence parameter $\theta$ with one dimension; i.e. $T \xrightarrow{p} 1$ when the null hypothesis is true. The asymptotic variance of the test statistic cannot be obtained or estimated analytically. [17] implemented a nonparametric bootstrap procedure to approximate the finite-sample distribution of the test statistic. The procedure is executed as follows.
1. Draw a sample with a size of $n$ from the data with replacement.

2. Calculate the approximate PIOS test statistic by using the re-sampled data. Denote the test statistic as $T^{(b)}$ for $b = 1, \ldots, B$.

3. Repeat steps 1 and 2 for $B$ times.

In this thesis, we consider a parametric bootstrap procedure to obtain the resamples of the test statistic.

1. Generate a bivariate random vector $(U_{i1}, U_{i2})$ from the copula model under the null hypothesis with the estimated $\hat{\theta}$.

2. Obtain $(T_{i1}^{(b)}, T_{i2}^{(b)})$, where $T_{i1}^{(b)} = \hat{S}_1^{-1}(U_{i1})$ and $T_{i2}^{(b)} = \hat{S}_2^{-1}(U_{i2})$. Note that $\hat{S}_1$ and $\hat{S}_2$ are the nonparametric estimates of the marginal survival functions.

3. Generate the random number $U_i^C$ from $U[0, 1]$, then $C_i^{(b)} = \hat{G}^{-1}(U_i^C)$, where $\hat{G}$ is the estimate of the survival function of $C_i$, $Pr(C_i > t)$. As mentioned earlier, the $\hat{G}$ is obtained using the Kaplan-Meier estimator based on $\{X_{Ci}, \delta_{Ci}\}$, where $X_{Ci} = \max \{X_{i1}, X_{i2}\}$ and $\delta_{Ci} = 1 - \delta_{i1}\delta_{i2}$.

4. Generate a sample $\mathcal{D}^{(b)} = \{(X_{i1}^{(b)}, X_{i2}^{(b)}, \delta_{i1}^{(b)}, \delta_{i2}^{(b)}), i = 1, \ldots, n\}$, where $X_{ij}^{(b)} = \min \{T_{ij}^{(b)}, C_i^{(b)}\}$ and $\delta_{ij}^{(b)} = I(T_{ij}^{(b)} < C_i^{(b)})$ for $j = 1, 2$.

5. Calculate the approximate PIOS test statistic by using the data $\mathcal{D}^{(b)}$, and denote the test statistic as $T^{(b)}$ for $b = 1, \ldots, B$.

6. Repeat the step 1, 2, and 3 for $B$ times.
7. Obtain the $p$-value by using the formula $\frac{1}{B} \sum_{i=1}^{B} I \left[ T_{c}^{(b)} > |\tilde{T} - 1| \right] \times 2$, where

$$T_{c}^{(b)} = T^{(b)} - \overline{T}^{b}.$$  

$\overline{T}^{b}$ is the average of $T^{(b)}$ for $b = 1, \ldots, B$.  


In this chapter, we first investigate how closely the approximated PIOS test statistic approximates the exact statistic. Secondly, we will compare the performance of the non-parametric and parametric bootstrap in terms of type I error and test power. The test will be conducted on three copula families, Clayton and Gumbel from the Archimedean copula and Normal copula. We set the value of the copula parameter $\theta$ for the Clayton and Gumbel copulas to correspond to the Kendall’s $\tau = 0.4$. We set the copula parameter $\theta$ for Normal copula to correspond to the correlation coefficient $\rho = 0.4$. Also, the censoring time is generated from an exponential distribution with the rate parameters of 0.2 and 0.9, which leads to 20% and 60% censoring rates (the percentage of subjects with at least one event time censored). Now, we will describe how to generate the data from a copula family.

### 3.1 Data Generation

We generate the simulated data set by the following steps for each $i = 1, \ldots, n$:

Step 1. Generate a bivariate random vector $(u_{i1}, u_{i2})$ from a copula model $\mathcal{C}_0(\cdot; \cdot; \theta_0)$.

When the copula family is Clayton or Gumbel, the value of $\theta_0$ is obtained from the relationship between $\theta_0$ and $\tau$ under each family with $\tau = 0.4$. When the cop-
ula family is Normal, \( \theta_0 = \rho = 0.4 \). The \texttt{R} package \texttt{copula} can be used for this step.

Step 2. Generate bivariate survival times \((T_{i1}, T_{i2})\) from \((u_{i1}, u_{i2})\) obtained in step 1, where \( T_{i1} = F_{1}^{-1}(u_{i1}) \) and \( T_{i2} = F_{2}^{-1}(u_{i2}) \). \( F_{j}^{-1}(\cdot) \) for \( j = 1, 2 \) is the quantile function of a Weibull distribution with the shape parameter being 2 and the scale parameter being 5.

Step 3. Generate the censoring time \( C_i \) from an exponential distribution with the rate parameter 0.2 and 0.9.

Step 4. Obtain the observed data \( \mathcal{D} = \{(X_{i1}, X_{i2}, \delta_{i1}, \delta_{i2}), i = 1, \ldots, n\} \), where \( X_{i1} = \min(T_{i1}, C_i) \), \( X_{i2} = \min(T_{i2}, C_i) \), \( \delta_{i1} = I(T_{i1} \leq C_i) \), and \( \delta_{i2} = I(T_{i2}, C_i) \).

### 3.2 Exact and Approximate PIOS Test Statistics

The section focuses on comparing the distribution of the exact and approximated PIOS test statistics under the null hypothesis. Figure 3.1 shows the histogram of the two test statistics under the three copula families with two different censoring rates. Figure 3.2 shows the quantile-quantile plot of the two test statistics. Overall, when the censoring rate is low, the distribution of the approximated PIOS test statistic is close to the distribution of the exact statistic. Table 3.1 reports the average of these two test statistics. The difference between the average of exact and approximate test statistics is small.

Table 3.2 reports the average computational time in seconds for calculation of the exact and approximate test statistics using a computer with 8 GB RAM and 8 processors.
Table 3.1

Average of the exact PIOS test statistic $T$ and the approximate PIOS test statistic $\tilde{T}$

<table>
<thead>
<tr>
<th>Copula Family</th>
<th>Censoring Rate</th>
<th>$T$</th>
<th>$\tilde{T}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>0.2</td>
<td>1.0352</td>
<td>1.0141</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>1.0297</td>
<td>0.9910</td>
</tr>
<tr>
<td>Gumbel</td>
<td>0.2</td>
<td>1.0189</td>
<td>1.0030</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>1.0336</td>
<td>1.0056</td>
</tr>
<tr>
<td>Normal</td>
<td>0.2</td>
<td>1.0397</td>
<td>1.0139</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>1.0472</td>
<td>1.0205</td>
</tr>
</tbody>
</table>

Table 3.2

Average computing time (in seconds) of calculating the exact PIOS test statistic $T$ and the approximate PIOS test statistic $\tilde{T}$

<table>
<thead>
<tr>
<th>Copula Family</th>
<th>Censoring Rate</th>
<th>$T$</th>
<th>$\tilde{T}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>0.2</td>
<td>1.36</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>1.32</td>
<td>0.08</td>
</tr>
<tr>
<td>Gumbel</td>
<td>0.2</td>
<td>2.89</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>2.24</td>
<td>0.09</td>
</tr>
<tr>
<td>Normal</td>
<td>0.2</td>
<td>26.66</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>92.53</td>
<td>1.07</td>
</tr>
</tbody>
</table>
Figure 3.1

Distributions of the exact PIOS test statistic $T$ and the approximate PIOS test statistic $\tilde{T}$
Figure 3.2

Quantile-quantile plot of the exact PIOS test statistic $T$ and the approximate PIOS test statistic $\hat{T}$
According to Table 3.2, using the approximation method takes much less time than calculating the exact statistic, especially with Normal copula. Therefore, we can conclude that the approximation method performs well and in the following simulation, only the approximated PIOS test statistics are calculated.

3.3 Nonparametric and Parametric Bootstrap

In this section, we compare the nonparametric and parametric bootstrap procedures by looking at the type I error and test power. The data are generated from each of the three copula families and tested against each other, for a total of nine tests. Table 3.3 reports the empirical standard deviation of the PIOS test statistic and the average standard error obtained from the nonparametric and parametric bootstrap. Table 3.4 reports the proportion of rejecting the null hypothesis under each true copula and the copula specified in the null hypothesis. The null hypothesis is rejected at significance level 0.05.

3.3.1 Result

The average standard error from nonparametric bootstrap is underestimated, which leads to the type I error to be inflated. On the other hand, the average standard error from the parametric bootstrap is larger than the empirical standard deviation, so the test using parametric bootstrap is more conservative, but the type I error is still controlled under the nominal level.

When the null hypothesis is not the true copula, the average standard error of the parametric bootstrap is smaller than that of the nonparametric bootstrap for most of the cases according to Table 3.3. It is because the nonparametric bootstrap approximates the dis-
tribution of the test statistic $T$ without considering the null hypothesis. On the contrary, the parametric bootstrap generates the data under the null hypothesis and approximates the distribution of the test statistic $T$ under the null hypothesis.

Table 3.4 reports that the test power with the parametric bootstrap is greater than that with the nonparametric bootstrap except for the case when the true copula is Normal and the null hypothesis is Clayton copula. A possible reasons for this case is that the Clayton copula is lower tail dependence, and the Normal copula is symmetric tail dependence. We test the right-censored data, so it might be hard to differentiate the two copulas.

The test power for the other cases is not high because of two reasons. First, the sample size is not large. With a larger sample size, the simulation could show a better power. Second, the dependence level is not high. When the dependence level, $\tau$ for Archimedean family and $\rho$ for the Normal copula, is small, the features are not distinguishable. We use $\tau = 0.4$ for Clayton and Gumbel copula and $\rho = 0.4$ for Normal copula. In the future, we could use $\tau = 0.8$ and $\rho = 0.8$ for more studies.

Lastly, Table 3.4 reports that the test power is lower with higher censoring rate, which is reasonable. If we increase the sample size, the test power would increase as well.
Table 3.3
Empirical standard deviation of $T$ and the average standard error obtained from nonparametric and parametric bootstrap

<table>
<thead>
<tr>
<th>Censoring Rate=0.2</th>
<th>Empirical Standard Deviation</th>
<th></th>
<th>Censoring Rate=0.6</th>
<th>Empirical Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Null</td>
<td>Clayton</td>
<td>Gumbel</td>
<td>Normal</td>
</tr>
<tr>
<td></td>
<td>Clayton</td>
<td>0.1787</td>
<td>0.1611</td>
<td>0.2344</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1618</td>
<td>0.1577</td>
<td>0.1970</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1973</td>
<td>0.1290</td>
<td>0.1861</td>
</tr>
<tr>
<td></td>
<td>Gumbel</td>
<td>0.2368</td>
<td>0.1325</td>
<td>0.2797</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2474</td>
<td>0.1280</td>
<td>0.2198</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1917</td>
<td>0.1339</td>
<td>0.1875</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>0.1868</td>
<td>0.1187</td>
<td>0.1628</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1897</td>
<td>0.1153</td>
<td>0.1450</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1900</td>
<td>0.1205</td>
<td>0.1680</td>
</tr>
</tbody>
</table>

Table 3.4
The proportions of rejecting the null hypothesis at the significance level 5% with nonparametric and parametric bootstrap

<table>
<thead>
<tr>
<th>Censoring Rate=0.2</th>
<th>Nonparametric</th>
<th></th>
<th>Censoring Rate=0.6</th>
<th>Nonparametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Null</td>
<td>Clayton</td>
<td>Gumbel</td>
<td>Normal</td>
</tr>
<tr>
<td></td>
<td>Clayton</td>
<td>0.080</td>
<td>0.310</td>
<td>0.178</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.042</td>
<td>0.456</td>
<td>0.278</td>
</tr>
<tr>
<td></td>
<td>Gumbel</td>
<td>0.114</td>
<td>0.058</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.358</td>
<td>0.042</td>
<td>0.352</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>0.046</td>
<td>0.114</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.046</td>
<td>0.358</td>
<td>0.040</td>
</tr>
</tbody>
</table>
[17] extended the PIOS test to the semi-parametric copula models for the right-censored bivariate event time. The PIOS test is a global goodness-of-fit test that compares “in-sample” and “out-of-sample” (pseudo) likelihood. Based on his work, we proposed an approximation method to alleviate the computation burden of calculating the exact PIOS statistics. We also designed a parametric bootstrap procedure to calculate the empirical $p$-value. Compared to nonparametric bootstrap, it leads to higher test power while having the type I error well controlled.

In the future work, more simulation studies should be conducted with a larger sample size, higher dependence level, and other copula families such as Frank. Also, we need to compare our method with existing tests, such as the goodness-of-fit tests in [23] and [4], which are targeted at only the Archimedean copula family. Our PIOS test is advantageous because it can be extended to all copula models, such as the $t$ copula model. Also, we can consider estimating the marginal distribution of the survival time with covariates for the goodness-of-fit test. Lastly, we will conduct more rigorous theoretical proof on asymptotic distribution of the PIOS test.
BIBLIOGRAPHY


