Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 228-238
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

# Dependence Model Selection for Semi-Competing Risks Data

**Jin-Jian Hsieh, Cheng-Fang Tsai**

Department of Mathematics, National Chung Cheng University, Taiwan
Email: jjhsieh@math.ccu.edu.tw

**How to cite this paper:** Hsieh, J.-J. and Tsai, C.-F. (2020) Dependence Model Selection for Semi-Competing Risks Data. *Open Journal of Statistics*, **10**, 228-238.
https://doi.org/10.4236/ojs.2020.102016

**Received:** February 17, 2020
**Accepted:** March 31, 2020
**Published:** April 3, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

## Abstract

We consider the model selection problem of the dependency between the terminal event and the non-terminal event under semi-competing risks data. When the relationship between the two events is unspecified, the inference on the non-terminal event is not identifiable. We cannot make inference on the non-terminal event without extra assumptions. Thus, an association model for semi-competing risks data is necessary, and it is important to select an appropriate dependence model for a data set. We construct the likelihood function for semi-competing risks data to select an appropriate dependence model. From simulation studies, it shows the performance of the proposed approach is well. Finally, we apply our method to a bone marrow transplant data set.

## Keywords

Copula Model, Likelihood Function, Model Selection, Semi-Competing Risks Data

## 1. Introduction

Semi-competing risks data [1] were often encountered in a biomedical study in which a terminal event censors a non-terminal event. A common example of the terminal event is death, and the non-terminal event usually is disease progression or relapse. When the relationship between the two events is unspecified, the inference on the non-terminal event is not identifiable. We cannot make inference on the non-terminal event without extra assumptions. Thus, an association model for semi-competing risks data is necessary for copula-based approaches [1]-[6], and it is important to select an appropriate dependence model for a data set. Hsieh *et al.* [4] used a distance measure between a non-parametric estimator and a model based estimator to select a proper dependence model. In this paper, we construct

the likelihood function under several candidate models for the semi-competing risks data and use the likelihood function to select a most fitted model. In simulations, we compare our proposed methods with Hsieh *et al.* [4]. This paper is organized as follows. In Section 2, we introduce the structure of semi-competing risks data and copula models. In Section 3, we derive the likelihood function for semi-competing risks data and introduce three model selection methods. We examine the finite sample performance of the proposed methods and compare them with Hsieh *et al.* [4] in Section 4. In Section 5, we use the Bone Marrow Transplant data from Klein and Moeschberger [7] to illustrate our suggested methods. Finally, we make some conclusions in Section 6.

## 2. Data and Model Assumption

Semi-competing risks data consist of a terminal event and a non-terminal event, which a terminal event may censor a non-terminal event. Let $T$ be the time from the initial event (e.g. disease diagnosis) to the non-terminal event (e.g. a status of disease progression), $D$ be the time from the initial event to the terminal event (e.g. death), and $C$ be the time from the initial event until lost to follow-up or the end of study. In general, we assume that $C$ is independent of $(T,D)$. Define $X=\min(T,D,C)$, $Y=\min(D,C)$, $\delta_1=I(T\le Y)$, and $\delta_2=I(D\le C)$. The observed data can be denoted as $\{(X_i,Y_i,\delta_{1i},\delta_{2i})\,|\,i=1,2,\cdots,n\}$.

With semi-competing risks data, we are interested in its dependence structure between $T$ and $D$ and require to ensure the validity for the inference of the non-terminal event time $T$. The most commonly used model for dependence is the copula model [8]. We assume $(T,D)$ follow a copula model as

$$Pr(T>t,D>d)=C_\alpha(S(t),G_1(d)),\ 0\le t,d\le\infty, \tag{1}$$

where $S(t)$ is the marginal survival function of $T$, $G_1(d)$ is the marginal survival function of $D$, $C_\alpha(\cdot,\cdot)$ is a parametric copula function defined on a unit square, and indexed by a single real parameter, $\alpha$, which is related to Kendall's tau [9]. To define Kendall's tau, suppose that $(T_1,D_1)$ and $(T_2,D_2)$ are two independent realizations of the joint distribution. Then, $\tau$ is the difference between the probability of concordance and the probability of discordance of these two observations, namely,

$$\tau=Pr\{(T_1-T_2)(D_1-D_2)\ge 0\}-Pr\{(T_1-T_2)(D_1-D_2)<0\}.$$

A copula $C_\alpha$ is said to be (strictly) Archimedean copula (AC) when

$$C_\alpha(u,v)=\varphi_\alpha^{-1}(\varphi_\alpha(u)+\varphi_\alpha(v)), \tag{2}$$

for all $0\le u,v\le 1$, where the $\varphi_\alpha:(0,1]\to R^+$ is a decreasing convex function satisfying $\varphi_\alpha(1)=0$ and $\varphi_\alpha(0^+)=\infty$. We take the following three examples for Archimedean copula, the Clayton, Frank and Gumbel. The Clayton copula is given by

$$C_\alpha(u,v)=\left(u^{-\alpha}+v^{-\alpha}-1\right)^{-1/\alpha},$$

and its generator is

$$\varphi_\alpha(t)=\left(t^{-\alpha}-1\right)/\alpha,$$

where $\alpha\in(0,\infty)\setminus\{0\}$. The relationship between Kendall's tau $\tau$ and the Clayton copula parameter $\alpha$ is given by $\tau=\alpha/(\alpha+2)$. The Frank copula is given by

$$C_\alpha(u,v)=-\frac{1}{\alpha}\ln\left(1+\frac{\left(\exp(-\alpha u)-1\right)\left(\exp(-\alpha v)-1\right)}{\exp(-\alpha)-1}\right),$$

and its generator is

$$\varphi_\alpha(t)=-\ln\left(\frac{\exp(-\alpha t)-1}{\exp(-\alpha)-1}\right),$$

where $\alpha\in(-\infty,\infty)\setminus\{0\}$. The relationship between Kendall's tau $\tau$ and the Frank copula parameter $\alpha$ is given by $1+4\{D_1(\alpha)-1\}/\alpha$, where $D_1(\alpha)=\int_0^\alpha\{t/\alpha(\mathrm{e}^t-1)\}\mathrm{d}t$. The Gumbel copula is given by

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

and its generator is

$$\varphi_\alpha(t)=\left[-\ln(t)\right]^{\alpha+1},$$

where $\alpha\in(0,\infty)$. The relationship between Kendall's tau $\tau$ and the Gumbel copula parameter $\alpha$ is given by $\tau=\alpha/(\alpha+1)$.

## 3. The Proposed Model Selection Methods

In statistical analysis, model selection is an important issue. Several candidate models are considered to fit data. Which model is the most appropriate for the considered data? Under semi-competing risks data, the observed data can be denoted as $\{X_i,Y_i,\delta_{1i},\delta_{2i}\,|\,i=1,2,\cdots,n\}$. To specify the dependency of $(T,D)$, we usually assume $(T,D)$ follows an AC model. Our goal is to choose a best copula model for the dependency of $T$ and $D$ among some candidate models, and the idea is to use the likelihood function information to choose the most fitted copula model from a candidate copula model set. Therefore, we need to derive the likelihood function under different copula models. We derive the likelihood function by considering the four possible situations for the values of $\delta_1$ and $\delta_2$. Let $S(t)$ be the survival function of $T$, $f_T(t)$ be the pdf of $T$, $G_1(d)$ be the survival function of $D$, $g_1(d)$ be the pdf of $D$, $G_2(y)$ be the survival function of $C$, and $g_2(y)$ be the pdf of $C$. For each case, we write a Randon-Nikodým derivative of the distribution of $(X,Y)$, and denote the $\frac{\partial^{i+j}}{\partial u^i\partial v^j}C(u,v)$ by $C^{(i,j)}(u,v)$.

- (Type A) If $\delta_1=\delta_2=0$, which means $X=Y=C$,

$$P(X=x,Y=y,\delta_1=0,\delta_2=0)=g_2(y)C_\alpha^{(0,0)}\left(S(y),G_1(y)\right).$$

- (Type B) If $\delta_1 = 0, \delta_2 = 1$, which means $X = Y = D$,

$$\begin{aligned} &P(X = x, Y = y, \delta_1 = 0, \delta_2 = 1) \\ &= G_2(y)\left[g_1(y) - g_1(y) - C_\alpha^{(0,1)}\left(S(y), G_1(y)\right)\right] \\ &= -G_2(y) C_\alpha^{(0,1)}\left(S(y), G_1(y)\right). \end{aligned}$$

- (Type C) If $\delta_1 = 1, \delta_2 = 0$, which means $X = T$ and $Y = C$,

$$\begin{aligned} &P(X = x, Y = y, \delta_1 = 1, \delta_2 = 0) \\ &= g_2(y)\left[f_T(x) - \frac{\partial}{\partial x}\left(1 - S(x) - G_1(y) + C_\alpha\left(S(x), G_1(y)\right)\right)\right] \\ &= g_2(y)\left[f_T(x) - f_T(x) - C_\alpha^{(1,0)}\left(S(x), G_1(y)\right)\right] \\ &= -g_2(y) C_\alpha^{(1,0)}\left(S(x), G_1(y)\right). \end{aligned}$$

- (Type D) If $\delta_1 = \delta_2 = 1$, which means $X = T$ and $Y = D$,

$$\begin{aligned} &P(X = x, Y = y, \delta_1 = 1, \delta_2 = 1) \\ &= G_2(y)\left[\frac{\partial^2}{\partial x \partial y} P(T \le x, D \le y)\right] \\ &= G_2(y)\left[\frac{\partial^2}{\partial x \partial y}\left(1 - S(x) - G_1(y) + C_\alpha\left(S(x), G_1(y)\right)\right)\right] \\ &= G_2(y) C_\alpha^{(1,1)}\left(S(x), G_1(y)\right). \end{aligned}$$

Summarizing the above situations, we can derive the likelihood function for one observation as

$$\begin{aligned} &L(\alpha, S \mid x, y, \delta_1, \delta_2) \propto (-1)^{\delta_1 + \delta_2} C_\alpha^{(\delta_1, \delta_2)}\left(S(x), G_1(y)\right) \\ &= \left(\varphi_\alpha^{-1}\right)^{\delta_1 + \delta_2}\left[\varphi_\alpha\left(S(x)\right) + \varphi_\alpha\left(G_1(y)\right)\right] f_T^{\delta_1}(x) g_1^{\delta_2}(y) \\ &= \left(\varphi_\alpha^{-1}\right)^{\delta_1 + \delta_2}\left[\varphi_\alpha\left(S(x)\right) + \varphi_\alpha\left(G_1(y)\right)\right]\left[S\left(x^-\right) - S(x)\right]^{\delta_1} g_1^{\delta_2}(y). \end{aligned} \tag{3}$$

We can use the Kaplan-Meier estimator $\hat{G}_1(\cdot)$ to estimate $G_1(\cdot)$ based on data $\{(Y_i, \delta_{2i}) \mid i = 1, 2, \cdots, n\}$ and define $\hat{g}_1(y)$ as $\hat{G}_1\left(y^-\right) - \hat{G}_1(y)$. Then we have a likelihood function for the $n$ observations as

$$\begin{aligned} &L\left(\alpha, S \mid \{x_i, y_i, \delta_{1i}, \delta_{2i} \mid i = 1, 2, \cdots, n\}\right) \\ &= \prod_{i=1}^{n}\left(\varphi_\alpha^{-1}\right)^{\delta_{1i} + \delta_{2i}}\left[\varphi_\alpha\left(S(x_i)\right) + \varphi_\alpha\left(\hat{G}_1(y_i)\right)\right] \times \left[S\left(x_i^-\right) - S(x_i)\right]^{\delta_{1i}} \hat{g}_1^{\delta_{2i}}(y_i). \end{aligned} \tag{4}$$

Based on the likelihood function, we consider three approaches to select an appropriate copula model, which are

$$\mathrm{A}_i^j = -2\ln\left(L_i^j\right), i = 1, 2, 3, \tag{5}$$

where $L_i^j$ is the corresponding maximum likelihood function described in (4) based on the $i$th approach $(i = 1, 2, 3)$ with $\varphi_{j,\alpha}(\cdot)$, which is the $j$th candidate copula model. Suppose that there are $M$ candidate copula models for consideration. For the $i$th method, we compute $\mathrm{A}_i^j$, $j = 1, 2, \cdots, M$. Then, select a copula model with the smallest $\mathrm{A}_i^j$.

For the first method, $L_1^j$, we use the estimator $\hat{S}(x)$ of $S(x)$ by Lakhal *et al.* [3], which is extended from Zheng and Klein [10], and we have the likelihood as

$$\begin{aligned} L_{1,j}^*(\alpha) &= L\left(\alpha \mid \left\{\left(x_i, y_i, \delta_{1i}, \delta_{2i}\right) \mid i = 1,2,\cdots,n\right\}\right) \\ &= \prod_{i=1}^{n} \left(\varphi_{j,\alpha}^{-1}\right)^{\delta_{1i}+\delta_{2i}} \left[\varphi_{j,\alpha}\left(\hat{S}(x_i)\right) + \varphi_{j,\alpha}\left(\hat{G}_1(y_i)\right)\right] \\ &\quad \times \left[\hat{S}\left(x_i^-\right) - \hat{S}(x_i)\right]^{\delta_{1i}} \hat{g}_1^{\delta_{2i}}(y_i). \end{aligned} \tag{6}$$

Now this function can be represented in the form of only one unknown parameter $\alpha$. Next, we apply the $optimize(\,)$ in R to obtain the maximum likelihood, and define $L_1^j = \max_{\alpha} L_{1,j}^*(\alpha)$.

For the second method, $L_2^j$, Kaplan-Meier [11] noted that the survival function can be written as $S(t) = \prod_{j:T_{(j)} \le t} \left(1 - h_j\right), t \ge 0$, where $h_j = 1 - S\left(T_{(j)}\right) \big/ S\left(T_{(j-1)}\right)$, and define $T_{(1)} < T_{(2)} < \cdots < T_{(k)}$ as the sort of $\left\{x_i \mid \delta_{1i} = 1, i = 1,2,\cdots,n\right\}$. So, there is a sequence $\boldsymbol{h} = \left(h_1,\cdots,h_k\right)$ corresponded to $\left(T_{(1)}, T_{(2)},\cdots,T_{(k)}\right)$. By Lakhal *et al.* [3], we can estimate the $\alpha$ parameter, which is also studied by Wang [2] and Heuchenne *et al.* [6]. From the above, we can write $L\left(S \mid \left\{\left(x_i, y_i, \delta_{1i}, \delta_{2i}\right) \mid i = 1,2,\cdots,n\right\}\right)$ as

$$\begin{aligned} L_{2,j}^*(\boldsymbol{h}) &= L\left(\boldsymbol{h} \mid \left\{\left(x_i, y_i, \delta_{1i}, \delta_{2i}\right) \mid i = 1,2,\cdots,n\right\}\right) \\ &= \prod_{i=1}^{n} \left(\varphi_{j,\hat{\alpha}}^{-1}\right)^{\delta_{1i}+\delta_{2i}} \left(\varphi_{j,\hat{\alpha}}\left(\prod_{j:T_j \le x_i} \left(1 - h_j\right)\right) + \varphi_{j,\hat{\alpha}}\left(\hat{G}_1(y_i)\right)\right) \\ &\quad \times \left[\prod_{j:T_j \le x_i^-} \left(1 - h_j\right) - \prod_{j:T_j \le x_i} \left(1 - h_j\right)\right]^{\delta_{1i}} \hat{g}_1^{\delta_{2i}}(y_i). \end{aligned} \tag{7}$$

Next, we use the PSO (Particle Swarm Optimization, Kennedy and Eberhart [12]) algorithm, which is a computational approach that optimizes the corresponding likelihood function by iteratively trying to improve a candidate solution, to obtain the *mle* of *h*, which is denoted as $\hat{\boldsymbol{h}}^{mle}$, and define $L_2^j = \max_{\boldsymbol{h}} L_{2,j}^*(\boldsymbol{h})$.

The third method, $L_3^j$, is similar to the second method but the number of the maximizers in likelihood function are more than $L_{2,j}^*(\boldsymbol{h})$. We maximize the corresponding likelihood with respective to $\alpha$ and $\boldsymbol{h}$ simultaneously, and the corresponding likelihood function is

$$\begin{aligned} L_{3,j}^*(\alpha, \boldsymbol{h}) &= L\left(\alpha, \boldsymbol{h} \mid \left\{\left(x_i, y_i, \delta_{1i}, \delta_{2i}\right) \mid i = 1,2,\cdots,n\right\}\right) \\ &= \prod_{i=1}^{n} \left(\varphi_{\alpha}^{-1}\right)^{\delta_{1i}+\delta_{2i}} \left(\varphi_{\alpha}\left(\prod_{j:T_j \le x_i} \left(1 - h_j\right)\right) + \varphi_{\alpha}\left(\hat{G}_1(y_i)\right)\right) \\ &\quad \times \left[\prod_{j:T_j \le x_i^-} \left(1 - h_j\right) - \prod_{j:T_j \le x_i} \left(1 - h_j\right)\right]^{\delta_{1i}} \hat{g}_1^{\delta_{2i}}(y_i). \end{aligned} \tag{8}$$

Then, use the PSO (Particle Swarm Optimization) algorithm to obtain the maximum likelihood, and define $L_3^j = \max_{\alpha,\boldsymbol{h}} L_{3,j}^*(\alpha, \boldsymbol{h})$. The difference for the three methods is the maximizer number in the corresponding likelihood function. In simulations, we would compare the correct selection probability for the three methods.

## 4. Simulation Studies

This section examines the performance of the proposed model selection methods and compares it with Hsieh *et al.* [4] through several simulation settings. Simulated data are generated from three copula models which are the Clayton model, Frank model, and Gumbel model. Based on simulated data from one copula among the three copulas in the above, three candidate models are considered to fit the simulated data. There are two different settings under two different censoring rates:

**High censoring rate:**

**Case 1:** $T \sim \exp(0.8)$, $D \sim \exp(1)$, and $C \sim U(0,6)$.

(The censoring rate is about 42% for $T$ and about 16% for $D$.)

**Case 2:** $T \sim \text{Weibull}(1,2)$, $D \sim \text{Weibull}(0.5,2)$, and $C \sim U(0,4)$.

(The censoring rate is about 33% for $T$ and about 28% for $D$.)

**Low censoring rate:**

**Case 3:** $T \sim \exp(0.5)$, $D \sim \exp(1)$, and $C \sim U(0,8)$.

(The censoring rate is about 23% for $T$ and about 12% for $D$.)

**Case 4:** $T \sim \text{Weibull}(5,2)$, $D \sim \text{Weibull}(4,3)$, and $C \sim U(0,4)$.

(The censoring rate is about 22% for $T$ and about 14% for $D$.)

In the above situations, we also set three different Kendall's tau, $\tau = 0.2, 0.5$, and 0.8 to determine which model is the most fitted candidate for simulated data. The sample size is 100 with 500 replications.

Tables 1-4 summarize the simulation results, and it presents the model selected percentage under different simulation data. Note that $\text{A}_i$ is the method based on the $\text{A}_i^j$ approach, $i = 1,2,3$, and $D^k$ is the method by Hsieh *et al.* [4]. From the results, the performance of the three proposed selection methods is better than Hsieh *et al.* [4], especially for Frank and Gumbel models. Thus, our proposed methods are more stable than Hsieh *et al.* [4]. From the Tables, we can find that the probability of choosing a correct model rises with increasing Kendall's tau, and also find that the performance under low censoring rate is better than high censoring rate. Based on the comparisons, we recommend using the first method, $\text{A}_1$, because it takes less computer running time than $\text{A}_2$ and $\text{A}_3$.

## 5. Data Analysis

In this section, we apply our proposed methods to analyze the bone marrow transplant data from Klein and Moeschberger [7]. There were 137 leukemia patients receiving bone marrow transplants. The data can be divided into three different groups, acute lymphoblastic leukemia (ALL) with 38 patients, acute myelocytic leukemia (AML) low-risk with 54 patients, and AML high-risk with 45 patients. Let $T$ be the time to relapse of leukemia, $D$ be the time to death, and $C$ be the censoring time. The observed variables are $X = \min(T,D,C)$, $Y = \min(D,C)$, $\delta_1 = I(T \le Y)$, and $\delta_2 = I(D \le C)$. For each group, we choose the most fitted

**Table 1.** Model selection probabilities under $T \sim \exp(0.8), D \sim \exp(1), C \sim U(0,6)$.

| | | Selected Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\tau = 0.2$ | | | $\tau = 0.5$ | | | $\tau = 0.8$ | | |
| Data | Method | C | F | G | C | F | G | C | F | G |
| Clayton | $A_1$ | **0.68** | 0.18 | 0.14 | **0.92** | 0.07 | 0.01 | **0.91** | 0.08 | 0.00 |
| | $A_2$ | **0.72** | 0.15 | 0.13 | **0.89** | 0.09 | 0.02 | **0.93** | 0.07 | 0.00 |
| | $A_3$ | **0.73** | 0.15 | 0.13 | **0.92** | 0.06 | 0.02 | **0.93** | 0.06 | 0.00 |
| | $D^k$ | **0.77** | 0.14 | 0.09 | **0.94** | 0.05 | 0.01 | **0.97** | 0.03 | 0.00 |
| Frank | $A_1$ | 0.25 | **0.42** | 0.33 | 0.10 | **0.67** | 0.23 | 0.02 | **0.84** | 0.14 |
| | $A_2$ | 0.32 | **0.35** | 0.33 | 0.11 | **0.64** | 0.25 | 0.03 | **0.81** | 0.16 |
| | $A_3$ | **0.34** | **0.34** | 0.32 | 0.11 | **0.63** | 0.26 | 0.04 | **0.80** | 0.16 |
| | $D^k$ | **0.60** | 0.24 | 0.16 | 0.29 | **0.52** | 0.19 | 0.14 | **0.71** | 0.15 |
| Gumbel | $A_1$ | 0.11 | 0.14 | **0.75** | 0.02 | 0.08 | **0.90** | 0.00 | 0.03 | **0.97** |
| | $A_2$ | 0.15 | 0.13 | **0.72** | 0.03 | 0.07 | **0.91** | 0.00 | 0.02 | **0.98** |
| | $A_3$ | 0.15 | 0.12 | **0.73** | 0.02 | 0.07 | **0.91** | 0.00 | 0.06 | **0.94** |
| | $D^k$ | **0.46** | 0.24 | 0.30 | 0.11 | 0.37 | **0.52** | 0.01 | 0.18 | **0.81** |

C: Clayton; F: Frank; G: Gumbel.

**Table 2.** Model selection probabilities under $T \sim W(1,2), D \sim W(0.5,2), C \sim U(0,4)$.

| | | Selected Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\tau = 0.2$ | | | $\tau = 0.5$ | | | $\tau = 0.8$ | | |
| Data | Method | C | F | G | C | F | G | C | F | G |
| Clayton | $A_1$ | **0.65** | 0.20 | 0.15 | **0.87** | 0.11 | 0.02 | **0.81** | 0.19 | 0.00 |
| | $A_2$ | **0.70** | 0.17 | 0.13 | **0.86** | 0.12 | 0.02 | **0.89** | 0.11 | 0.00 |
| | $A_3$ | **0.71** | 0.16 | 0.13 | **0.87** | 0.11 | 0.02 | **0.91** | 0.09 | 0.00 |
| | $D^k$ | **0.73** | 0.12 | 0.15 | **0.86** | 0.12 | 0.02 | **0.77** | 0.20 | 0.03 |
| Frank | $A_1$ | 0.22 | **0.46** | 0.32 | 0.10 | **0.69** | 0.21 | 0.02 | **0.91** | 0.07 |
| | $A_2$ | 0.25 | **0.44** | 0.32 | 0.09 | **0.68** | 0.23 | 0.02 | **0.89** | 0.08 |
| | $A_3$ | 0.27 | **0.41** | 0.32 | 0.10 | **0.66** | 0.23 | 0.04 | **0.88** | 0.09 |
| | $D^k$ | **0.58** | 0.18 | 0.24 | **0.45** | 0.42 | 0.13 | 0.37 | **0.48** | 0.16 |
| Gumbel | $A_1$ | 0.10 | 0.16 | **0.78** | 0.00 | 0.08 | **0.92** | 0.00 | 0.04 | **0.96** |
| | $A_2$ | 0.12 | 0.17 | **0.75** | 0.01 | 0.07 | **0.92** | 0.00 | 0.02 | **0.98** |
| | $A_3$ | 0.13 | 0.17 | **0.75** | 0.01 | 0.07 | **0.92** | 0.00 | 0.11 | **0.89** |
| | $D^k$ | 0.51 | **0.53** | 0.32 | 0.12 | 0.37 | **0.51** | 0.03 | 0.16 | **0.80** |

C: Clayton; F: Frank; G: Gumbel.

**Table 3.** Model selection probabilities under $T \sim \exp(0.5), D \sim \exp(1), C \sim U(0,8)$.

| | | Selected Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\tau = 0.2$ | | | $\tau = 0.5$ | | | $\tau = 0.8$ | | |
| Data | Method | C | F | G | C | F | G | C | F | G |
| Clayton | $A_1$ | **0.73** | 0.16 | 0.11 | **0.91** | 0.08 | 0.01 | **0.90** | 0.10 | 0.00 |
| | $A_2$ | **0.73** | 0.16 | 0.11 | **0.89** | 0.10 | 0.01 | **0.93** | 0.07 | 0.00 |
| | $A_3$ | **0.75** | 0.14 | 0.11 | **0.91** | 0.08 | 0.01 | **0.93** | 0.07 | 0.00 |
| | $D^k$ | **0.75** | 0.14 | 0.11 | **0.91** | 0.09 | 0.00 | **0.84** | 0.14 | 0.02 |
| Frank | $A_1$ | 0.24 | **0.47** | 0.29 | 0.04 | **0.76** | 0.20 | 0.00 | **0.93** | 0.07 |
| | $A_2$ | 0.25 | **0.45** | 0.30 | 0.04 | **0.76** | 0.20 | 0.00 | **0.91** | 0.09 |
| | $A_3$ | 0.27 | **0.44** | 0.29 | 0.06 | **0.70** | 0.25 | 0.00 | **0.91** | 0.09 |
| | $D^k$ | **0.59** | 0.25 | 0.16 | 0.38 | **0.51** | 0.12 | 0.19 | **0.63** | 0.18 |
| Gumbel | $A_1$ | 0.08 | 0.15 | **0.77** | 0.00 | 0.06 | **0.94** | 0.00 | 0.04 | **0.96** |
| | $A_2$ | 0.09 | 0.13 | **0.77** | 0.00 | 0.06 | **0.94** | 0.00 | 0.03 | **0.97** |
| | $A_3$ | 0.10 | 0.13 | **0.78** | 0.00 | 0.05 | **0.95** | 0.00 | 0.06 | **0.94** |
| | $D^k$ | **0.49** | 0.24 | 0.27 | 0.10 | 0.35 | **0.55** | 0.01 | 0.14 | **0.85** |

C: Clayton; F: Frank; G: Gumbel.

**Table 4.** Model selection probabilities under $T \sim W(5,2), D \sim W(4,3), C \sim U(0,4)$.

| | | Selected Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\tau = 0.2$ | | | $\tau = 0.5$ | | | $\tau = 0.8$ | | |
| Data | Method | C | F | G | C | F | G | C | F | G |
| Clayton | $A_1$ | **0.68** | 0.21 | 0.11 | **0.68** | 0.23 | 0.09 | **0.93** | 0.07 | 0.00 |
| | $A_2$ | **0.70** | 0.18 | 0.11 | **0.73** | 0.18 | 0.09 | **0.94** | 0.06 | 0.00 |
| | $A_3$ | **0.73** | 0.16 | 0.11 | **0.75** | 0.17 | 0.08 | **0.96** | 0.04 | 0.00 |
| | $D^k$ | **0.76** | 0.15 | 0.10 | **0.75** | 0.14 | 0.11 | **0.92** | 0.05 | 0.03 |
| Frank | $A_1$ | 0.23 | **0.50** | 0.27 | 0.04 | **0.82** | 0.13 | 0.00 | **0.96** | 0.03 |
| | $A_2$ | 0.24 | **0.48** | 0.28 | 0.06 | **0.79** | 0.15 | 0.00 | **0.95** | 0.05 |
| | $A_3$ | 0.26 | **0.45** | 0.29 | 0.07 | **0.78** | 0.15 | 0.01 | **0.93** | 0.06 |
| | $D^k$ | **0.60** | 0.25 | 0.15 | 0.38 | **0.52** | 0.10 | 0.23 | **0.59** | 0.18 |
| Gumbel | $A_1$ | 0.08 | 0.15 | **0.77** | 0.00 | 0.08 | **0.92** | 0.00 | 0.04 | **0.96** |
| | $A_2$ | 0.10 | 0.12 | **0.77** | 0.00 | 0.07 | **0.93** | 0.00 | 0.02 | **0.98** |
| | $A_3$ | 0.10 | 0.11 | **0.78** | 0.00 | 0.06 | **0.93** | 0.00 | 0.04 | **0.96** |
| | $D^k$ | **0.50** | 0.23 | 0.27 | 0.13 | **0.46** | 0.41 | 0.04 | 0.34 | **0.63** |

C: Clayton; F: Frank; G: Gumbel.

model among the three models, Clayton, Frank, and Gumbel copula, by the four methods, and present the results in Tables 5-8. From the results, our methods choose Clayton copula for ALL group, AML high risk group, and all patients; select Gumbel model for AML low risk group. Hsieh *et al.* [4] choose Gumbel copula for ALL group, AML high risk group, and all patients; selects Frank copula for AML low risk group.

Table 5. The selected model for each method under ALL group.

| Group | ALL group | | |
|---|---|---|---|
| | Clayton | Frank | Gumbel |
| $A_1$ | **138.92** | 139.98 | 141.78 |
| $A_2$ | **138.97** | 140.00 | 141.66 |
| $A_3$ | **138.86** | 139.87 | 141.57 |
| $D^k$ | 0.412 | 0.386 | **0.385** |

Table 6. The selected model for each method under AML low risk group.

| Group | AML low risk group | | |
|---|---|---|---|
| | Clayton | Frank | Gumbel |
| $A_1$ | 142.58 | 142.41 | **141.89** |
| $A_2$ | 142.65 | 142.35 | **141.80** |
| $A_3$ | 142.52 | 142.33 | **141.73** |
| $D^k$ | 0.521 | **0.488** | 0.495 |

Table 7. The selected model for each method under AML high risk group.

| Group | AML high risk group | | |
|---|---|---|---|
| | Clayton | Frank | Gumbel |
| $A_1$ | **213.10** | 215.12 | 217.59 |
| $A_2$ | **212.84** | 214.49 | 216.90 |
| $A_3$ | **212.74** | 214.48 | 216.65 |
| $D^k$ | 0.310 | 0.292 | **0.284** |

Table 8. The selected model for each method under all patients.

| Group | all patients | | |
|---|---|---|---|
| | Clayton | Frank | Gumbel |
| $A_1$ | **627.82** | 628.05 | 633.14 |
| $A_2$ | **627.68** | 627.86 | 633.23 |
| $A_3$ | **627.63** | 627.77 | 632.65 |
| $D^k$ | 0.208 | 0.197 | **0.184** |

## 6. Concluding Remarks

In this paper, we study the model selection problem under semi-competing risks data. Because the non-terminal event is dependently censored by the terminal event, we cannot make inference on the non-terminal event without extra assumptions. Thus, an association model for semi-competing risks data is necessary, and a model selection method is necessary for the association model. We construct the likelihood function for semi-competing risks data under a copula model and propose three approaches based on the likelihood function to select a fitted model. The simulation analysis shows the performance of the proposed methods are more stable than Hsieh *et al.* [4], and $A_1$ takes less time than $A_2$ and $A_3$. With covariates, we can stratify the data according to the covariates and apply the model selection approach for each stratum. For the continuous covariates, we can group it as a categorical variable. Finally, we apply our proposed methods to analyze the Bone Marrow Transplant data. Base on the selected model, an interesting problem is to consider the goodness-of-fit test, which is treated as future work.

## Acknowledgements

This paper was financially supported by the Ministry of Science and Technology of Taiwan (MOST 108-2118-M-194-001-MY2).

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Fine, J.P., Jiang, H. and Chappell, R. (2001) On Semi-Competing Risks Data. *Biometrika*, **88**, 907-919. https://doi.org/10.1093/biomet/88.4.907

[2] Wang, W. (2003) Estimating the Association Parameter for Copula Models under Dependent Censoring. *Journal of the Royal Statistical Society*: *Series B*, **65**, 257-273. https://doi.org/10.1111/1467-9868.00385

[3] Lakhal, L., Rivest, L.P. and Abdous, B. (2008) Estimating Survival and Association in a Semi-Competing Risk Model. *Biometrics*, **64**, 180-188. https://doi.org/10.1111/j.1541-0420.2007.00872.x

[4] Hsieh, J.J., Wang, W. and Ding, A.A. (2008) Regression Analysis Based on Semi-Competing Risks Data. *Journal of the Royal Statistical Society*: *Series B*, **70**, 3-20.

[5] Xu, J., Kalbfleisch, J.D. and Tai, B. (2010) Statistical Analysis of Illness-Death Processes and Semicompeting Risks Data. *Biometrics*, **66**, 716-725. https://doi.org/10.1111/j.1541-0420.2009.01340.x

[6] Heuchenne, C., Laurent, S., Legrand, C. and Keilegom, I.V. (2014) Likelihood Based Inference for Semi-Competing Risks. *Communications in Statistics—Simulations and Computations*, **43**, 1112-1132. https://doi.org/10.1080/03610918.2012.725495

[7] Klein, J.P. and Moeschberger, M.L. (2003) Survival Analysis Techniques for Censored and Truncated Data. Springer, New York. https://doi.org/10.1007/b97377

[8] Genest, C. and MacKay, J. (1986) The Joy of Copulas: Bivariate Distributions with

Uniform Marginals. *The American Statistician*, **40**, 280-283.
https://doi.org/10.1080/00031305.1986.10475414

[9] Oakes, D. (1989) Bivariate Survival Models Induced by Frailties. *Journal of the American Statistical Association*, **74**, 487-493.
https://doi.org/10.1080/01621459.1989.10478795

[10] Zheng, M. and Klein, J. (1995) Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula. *Biometrika*, **82**, 127-138.
https://doi.org/10.1093/biomet/82.1.127

[11] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457-481.
https://doi.org/10.1080/01621459.1958.10501452

[12] Kennedy, J. and Eberhart, R. (1995) Particle Swarm Optimization. *Proceedings of IEEE International Conference on Neural Networks*, Vol. 4, 1942-1948.
https://doi.org/10.1109/ICNN.1995.488968