Thesis for the Degree of Doctor of Science

One-Stage Copula Modeling Approaches for Clustered Multivariate Survival Data

by

Sook Hee Kwon

Department of Statistics

The Graduate School

Pukyong National University

August 2021

One-Stage Copula Modeling Approaches for Clustered Multivariate Survival Data (군집된 다변량 생존 자료에 대한 1단계 코플라 모형 접근법)

Advisor: Prof. Il Do Ha

by
Sook Hee Kwon

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Science

in Department of Statistics, The Graduate School,
Pukyong National University

August 2021

One-Stage Copula Modeling Approaches for Clustered Multivariate Survival Data

A dissertation

by

Sook Hee Kwon

Approved by:

_____

(Daeheung Jang)

_____ _____

(Seongbaek Yi) (Saang Yoon Hyun)

_____ _____

(Ki Mun Jung) (Il Do Ha)

August 27, 2021

Contents

List of Tables

List of Figures

One-Stage Copula Modeling Approaches
for Clustered Multivariate Survival Data

Sook Hee Kwon

Department of Statistics, The Graduate School,
Pukyong National University

Abstract

Copula survival and frailty models have been widely used to analyze clustered multivariate survival data. The copula models consist of copula function with marginal distribution. The copula model is a marginal model, while the frailty model is a conditional model. In particular, the family of Archimedean copula functions, a broad class of copulas, is useful for modeling such dependency among survival data. However, the inference of copula survival models has been relatively less studied. In general, one- and two-stage estimation methods have been used for likelihood-based inference. The two-stage procedure can provide inefficient estimation results because it estimates the copula's marginal and association parameters separately. However, a more efficient one-stage procedure has been mainly developed under a restrictive parametric assumption of the marginal distribution due to the complexity of the likelihood with an unknown marginal baseline hazard function.

In this thesis, we propose a flexible M-spline Archimedes copula modeling approach using a one-stage likelihood procedure. To reduce the complexity of the likelihood, the unknown marginal baseline hazard is modeled based on the cubic M-spline basis function that does not require a specific parametric form. The estimation procedure of the proposed method is derived and theoretical properties are also studied. Simulation results show that the proposed one-stage estimation method gives a consistent estimator and also provides more efficient estimation results than

the existing one- and two-stage methods. The proposed method is illustrated with three practical data examples.

In this thesis, we also propose a variable selection procedure in the copula model using a one-stage estimation method based on a penalized likelihood. The performance of the proposed method is evaluated through simulation studies, and the usefulness of the new method is illustrated using clinical data sets.

Ⅰ. INTRODUCTION

In survival analysis, clustered survival time data are mainly obtained by a cluster, such as a center or a subject, and the dependence among the event times has often been modeled using copula model or frailty model (Hougaard, 2000; Duchateau and Janssen, 2000). Here, frailty refers to an unobserved random effect that multiplicatively affects each individual's hazard rate (Duchateau and Janssen, 2008; Ha et al., 2017). Copula is also one of the convenient ways to describe the dependence between random variables. According to Sklar's (1959) theorem, the joint distribution of random variables can be expressed as a copula function with the marginal distribution of each random variable. In particular, it is worth noting that the frailty model is a conditional model, and the copula model is a marginal model, and so the two models are different (Goethals et al., 2008; Prenen et al., 2017a). Joint survival function of the copula model is easily constructed by specifying both copula and marginal survival functions, while that of the frailty model is generally obtained by computing difficult integrations for frailty term. A basic comparative study between copula and frailty modeling approaches has been conducted for correlated or multivariate survival data with the same or different cluster sizes (Duchateau and Janssen, 2008; Goethals et al., 2008; Prenen et al., 2017a). However, the estimation for the copula models was relatively less studied.

In this thesis, we study an efficient estimation method of copula survival models. For the inference of a copula-based survival model, one-stage or two-stage estimation method has been generally used. In previous studies, for copula-based models where maximum likelihood inference is computationally difficult, the two-stage

estimation procedure has been widely used for survival data with the same cluster size (Shih and Louis, 1995; Andersen et al., 2005). However, the two-stage estimation procedure is based on a separate estimation between parameters in marginal distribution and association parameters in copula function. Thus, the resulting two-stage estimates may not effectively reflect the dependence information among survival times, which may be statistically less efficient. In different types of copula models with bivariate survival data, Marra and Radice (2020), through simulation studies, pointed out that the two-stage estimation is inefficient especially with a strong dependence, but that the one-stage estimation shows a good performance (Cheng et al,, 2014; Romeo et al., 2018). In particular, Chen et al. (2006) studied a one-stage estimation procedure without covariates in multivariate survival data. Recently, Prenen et al. (2017a) proposed a new method of Archimedean copula model for multivariate survival data with different cluster sizes, as well as several estimation procedures including one-stage and two-stage estimation methods.

In the Archimedean copula survival model, the two-stage procedure provides both parametric and non-parametric (i.e., Breslow's (1972) method) estimates of marginal baseline hazards. However, for the one-stage procedure, Prenen et al.'s (2017a) method provides only parametric approaches (e.g. Weibull, piecewise exponential) for marginal baseline hazards as the derivation of Breslow's (1972) estimator is difficult due to the complexity of the likelihood formulation under Archimedean copula survival model with an unknown baseline hazard function. The use of piecewise exponential for the baseline hazard may give flexible estimation results, but it requires choosing a suitable partition (i.e. cut-point) of the follow-up

time. Furthermore, the piecewise exponential hazard function is discontinuous at chosen locations of the partitions. Under these situations, a spline-based method can be a better alternative owing to its computational efficiency and flexibility of the model. Overall, the one-stage approach is preferred and it especially leads to less biased estimates in small sample cases (Prenen et al., 2017a).

Therefore, in this thesis, we propose a flexible parametric Archimedean copula survival regression modeling approach using a one-stage likelihood procedure. To reduce the complexity of the full likelihood, the unknown marginal baseline hazards are modeled based on a cubic M-spline basis function (Ramsay, 1988). The estimation procedure of the proposed method is also derived. The simulation results demonstrate that the proposed one-stage estimation method gives a consistent estimator and also provides more efficient estimation results over existing one- and two-stage methods (Emura et al., 2017, 2019, 2020). In addition, we study the sensitivity of the proposed method against misspecification of the Archimedes copula regression model for correlated survival data with different cluster sizes. The usefulness of this new method is illustrated using three well-known clinical data sets, i.e. kidney infection data (McGilchrist and Aisbett, 1991), chronic granulomatous disease (CGD) recurrence data (Fleming and Harrington, 1991) and bladder cancer recurrence data (Oddens et al., 2013) from a multicenter clinical trial conducted by the European Organization for Research and Treatment of Cancer (EORTC). In addition, our results are compared with existing one- and two-stage results using the three data sets.

In this thesis, we also propose a variable selection method in a

copula survival regression model with a parametric marginal distribution using a one-stage estimation method based on penalized likelihood. Here we also study four penalty functions, i.e. least absolute shrinkage and selection operator (LASSO; Tibshiran, 1996), adaptive LASSO (ALASSO; Zou, 2006), smoothly clipped absolute deviation (SCAD; Fan and Li, 2001) and h-likelihood (HL; Lee and Oh, 2014). The new variable selection procedures are derived. Thus the performance of the proposed methods is evaluated using simulation studies. The usefulness of the proposed method is illustrated using two clinical data sets: the kidney infection data and the CGD recurrence data.

This thesis is organized as follows. In Chapter 2, we review copula and frailty survival models and provide background knowledge of this thesis. In Chapter 3, we propose a one-stage procedure for estimating regression and association parameters using the M-spline method. The results of simulation studies are presented in Chapter 4. In Chapter 5, the proposed one-stage method is illustrated with three real-data examples. In Chapter 6, we propose a variable selection method using a penalized likelihood for the copula model, and also present the simulation results and real data examples. Discussion is given in Chapter 7. Finally, technical details including R codes and further simulation results are given in the Appendix.

Ⅱ. COPULA AND FRAILTY MODELS

We first review the basic quantities (e.g. survival and hazard functions) for survival analysis. Let T be survival time. We assume that T is an absolute continuous random variable taking on non-negative value. Therefore, T has a cumulative distribution function $F(t)$, defined as

$$F(t)=P(T\le t)=\int_0^t f(\kappa)d\kappa,\ \ t\ge 0.$$

By the continuity of T, the probability density function $f(t)$ of T is given by

$$f(t)=\frac{dF(t)}{dt}.$$

Thus, the survival function $S(t)$ of T is defined as

$$S(t)=P(T>t)=1-F(t)=\int_t^\infty f(\kappa)d\kappa\ ,\ \ t\ge 0,$$

which measures the probability that the event does not occur until time t. For examples, $S(t)$ means the probability that the patient survives beyond time t or the machine does not fail until time t.

The hazard function $\lambda(t)$ of T at time t is defined by

$$\lambda(t)=\lim_{\Delta t\to 0}\frac{P(t\le T<t+\Delta t|\, T\ge t)}{\Delta t}=\lim_{\Delta t\to 0}\frac{P(t\le T<t+\Delta t)}{\Delta t\cdot P(T\ge t)}$$

$$=\frac{1}{P(T\ge t)}\left[\lim_{\Delta t\to 0}\frac{P(t\le T<t+\Delta t)}{\Delta t}\right]=\frac{f(t)}{S(t)},$$

and

$$\lambda(t)=\frac{f(t)}{S(t)}=-\frac{dS(t)}{dt}\cdot\frac{1}{S(t)}=-\frac{d\log S(t)}{dt}.$$

The hazard function shows the instantaneous failure rate at time t if the event has not yet occurred at that moment. However, care is necessary in that the hazard function is not a probability. The cumulative hazard function $\Lambda(t)$ is defined as

$$\Lambda(t)=\int_0^t\lambda(\kappa)d\kappa=-\log S(t),$$

and

$$S(t)=\exp\{-\Lambda(t)\}.$$

2.1. Correlated survival models

In this section, we review the basic concepts and differences between the copula model and the frailty model which are widely used in the analysis of correlated survival data.

Let T_{ij} be event time (time-to-event) for the jth $(j=1,\cdots,n_i)$ observation of the ith $(i=1,\cdots,q)$ cluster (or subject) and let C_{ij} be

the censoring time corresponding to T_{ij}. Here, n_i is the cluster size, q is the number of clusters, and $N = \sum_{i=1}^{q} n_i$ is the total sample size. Then the observable random variables are as follows:

$$Y_{ij} = \min(T_{ij}, C_{ij}) \text{ and } \delta_{ij} = I(T_{ij} \leq C_{ij}),$$

where δ_{ij} is an event indicator function, indicating whether censoring is occurred or not.

2.1.1. Copula model

Copula models assume that the joint survival function of the individuals within a cluster is given by a copula function with the marginal survival function of each individual (Sklar, 1959). There are many copula functions describing rich patterns of tail dependence, ranging from tail independence to tail dependence, and different kinds of asymmetry. Among all types of copulas, frequently used copulas include Gaussian copula and T copula from elliptical copula family, and Gumbel copula, Clayton copula and Frank copula from Archimedean copula family (Nelson, 1999; Cherubini et al., 2004; Skoglund, 2010). The forms of these five major copulas and their bivariate copula functions are shown in Figure 2.1.1.

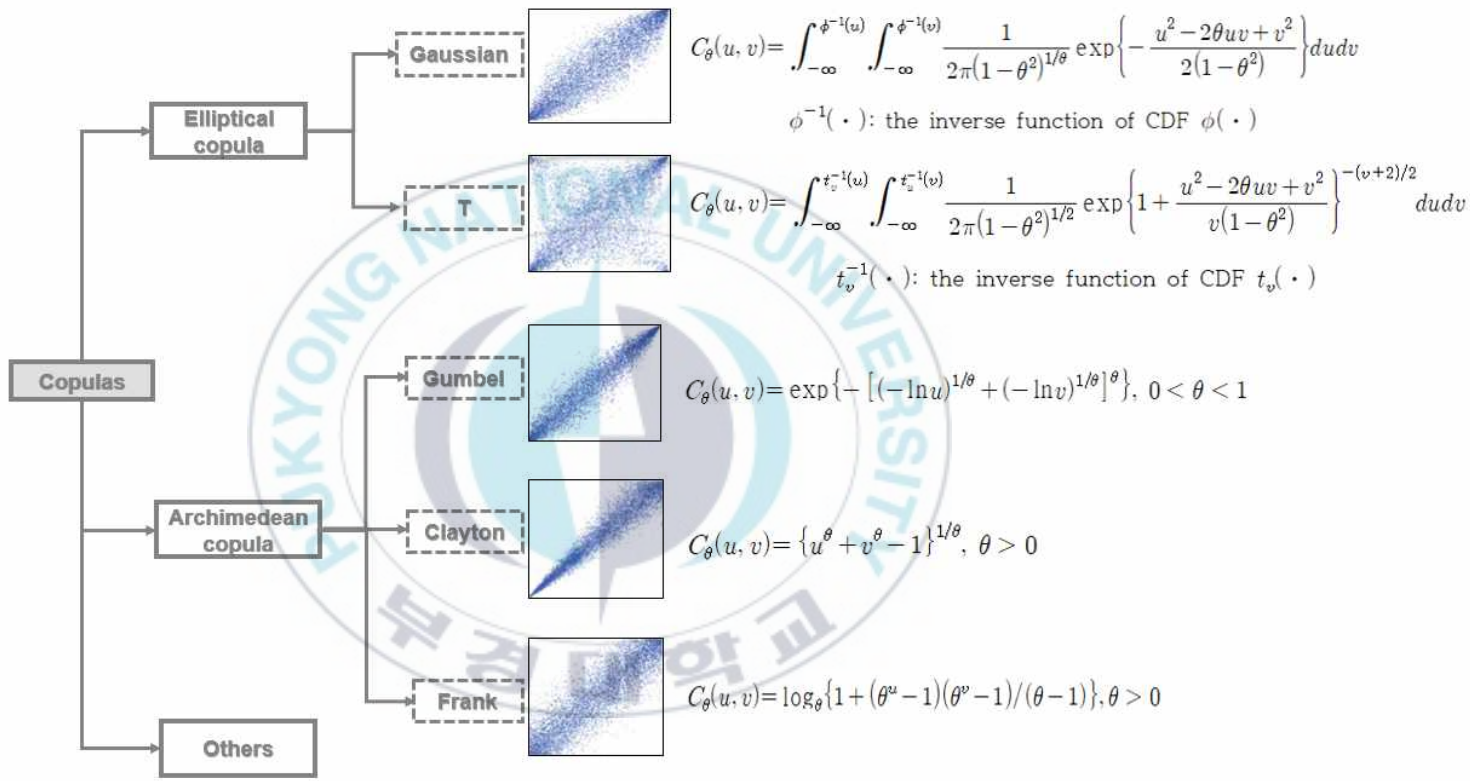


Figure 2.1.1. The five primary copulas and their bivariate copulas

https://www.assetinsights.net/Glossary/G_Gumbel_Copula.html

Below we begin with the definition of a copula survival model.

Definition 2.1. Let $S_j(t_j)=P(T_j>t_j)$ be a marginal survival function for the jth event time T_j $(j=1,\cdots,J)$ within a cluster. Then the joint survival function for J−variable event times $T_1, \cdots, T_J$ can be expressed as a copula survival function with each marginal function as follows:

$$\begin{aligned} S(t_1, \cdots, t_J) &= P(T_1 > t_1, \cdots, T_J > t_J) \\ &= C_\theta\{S_1(t_1), \cdots, S_J(t_J)\}, \end{aligned} \tag{2.1.1}$$

where $C_\theta(\cdot)$ is a J−variate copula function which is a distribution function on $[0,1]^J \rightarrow [0,1]$, and θ is an association parameter that explains the dependency among survival data.

Definition 2.2. The Archimedean copula model considered in this thesis is defined as follows (Joe, 1997; Prenen et al., 2017a):

$$C_\theta(w_1, \cdots, w_J) = \psi_\theta\{\psi_\theta^{-1}(w_1) + \cdots + \psi_\theta^{-1}(w_J)\}, \tag{2.1.2}$$

where the generator $\psi_\theta : [0,\infty) \rightarrow [0,1]$ is a continuous strictly monotonic decreasing function, and $\psi_\theta(0)=1$, $\psi_\theta(\infty)=0$ and ψ_θ^{-1} is the inverse function of ψ_θ. The generator ψ_θ of the Archimedean copula depends on the association parameter θ.

Consider a vector of p−dimensional covariates, denoted by $x_{ij}=(x_{ij1}, \cdots, x_{ijp})^T$, corresponding to survival time T_{ij}.

Definition 2.3. For the Archimedean copula family (Joe, 1997; Prenen et al., 2017a), the joint survival function of $T_{i1}, \cdots, T_{in_i}$ for cluster i given x_{ij} $(j=1, \cdots, n_i)$ is expressed as

$$\begin{aligned} S(t_{i1}, \cdots, t_{in_i}|x_{ij}, \forall j) &= P(T_{i1} > t_{i1}, \cdots, T_{in_i} > t_{in_i}|x_{ij}, \forall j) \\ &= \psi_\theta\left[\psi_\theta^{-1}\{S_1(t_{i1}|x_{i1})\} + \cdots + \psi_\theta^{-1}\{S_{n_i}(t_{in_i}|x_{in_i})\}\right], \end{aligned} \tag{2.1.3}$$

where $S_j(t_{ij}|x_{ij})$ is a marginal survival function for T_{ij} given x_{ij} $(j=1, \cdots, n_i)$, and the generator ψ_θ of the Archimedean copula can be expressed as a Laplace transform of the positive distribution function $G_\theta(\cdot)$ with $G_\theta(0)=0$:

$$\psi_\theta(\kappa) = \int_0^\infty \exp(-\kappa y) dG_\theta(y), \ \kappa \geq 0. \tag{2.1.4}$$

Thus, the joint survival function above for cluster i can be rewritten as

$$\begin{aligned} S(t_{i1}, \cdots, t_{in_i}|x_{ij}, \forall j) &= \int \exp\left[-y\sum_{j=1}^{n_i}\psi_\theta^{-1}\{S_j(t_{ij}|x_{ij})\}\right] dG_\theta(y) \\ &= \int \prod_{j=1}^{n_i} \exp\left[-y\psi_\theta^{-1}\{S_j(t_{ij}|x_{ij})\}\right] dG_\theta(y). \end{aligned} \tag{2.1.5}$$

In this thesis, we assume that the marginal survival function $S_j(t_{ij}|x_{ij})$ is obtained from the proportional hazard (PH) model:

$$\lambda_{ij}(t|x_{ij}) = \lambda_0(t)\exp(x_{ij}^T\beta), \tag{2.1.6}$$

where $\lambda_0(\cdot)$ is a baseline hazard function, which can be a parametric or non-parametric form and $\beta=(\beta_1, \cdots, \beta_p)^T$ is a $p\times 1$ vector of the regression parameters corresponding to covariates x_{ij}. For a parametric case of $\lambda_0(t)$, let $\zeta_0>0$ be the scale parameter and $\phi>0$ be the shape parameter. Then, for example, $\lambda_0(t)=\zeta_0\phi t^{\phi-1}$ for the Weibull distribution and $\lambda_0(t)=\zeta_0\exp(\phi t)$ for the Gompertz distribution.

Below we consider two popular members of the Archimedean copula family, i.e., the Clayton and the GH copulas.

• Clayton copula model

For the Clayton copula model, the generator having a gamma distribution function $G_\theta(\cdot)$ with mean 1 and variance θ is as follows:

$$\psi_\theta(s)=(1+\theta s)^{-1/\theta} \text{ for } \theta>0, \qquad (2.1.7)$$

with $\psi_\theta^{-1}(s)=(s^{-\theta}-1)/\theta$.

The Clayton copula model has a lower tail dependence and its Kendall's tau is given by $\tau=\theta/(\theta+2)$. This means a positive association among event times when $\theta>0$ and independence when $\theta\to 0$. From (2.1.5)-(2.1.7), the joint survival function under the Clayton copula model is given by an explicit form:

$$S(t_{i1}, \cdots, t_{in_i}|x_{ij}, \forall j)=\left\{\sum_{j=1}^{n_i} S_j(t_{ij}|x_{ij})^{-\theta}-n_i+1\right\}^{-1/\theta}, \qquad (2.1.8)$$

where the marginal survival function is given by

$$S_j(t_{ij}|x_{ij}) = \exp\{-\Lambda(t_{ij}|x_{ij})\}. \quad (2.1.9)$$

Here, the corresponding marginal cumulative hazard function from (2.1.6) is given by

$$\Lambda_j(t_{ij}|x_{ij}) = \Lambda_0(t_{ij})\exp(x_{ij}^T\beta), \quad (2.1.10)$$

and the baseline cumulative hazard function:

$$\Lambda_0(t) = \int_0^t \lambda_0(\kappa)d\kappa, \quad (2.1.11)$$

• Gumbel Hougaard (GH) copula

The GH copula's generator with a positive stable distribution function $G_\theta(\cdot)$ (Hougaard, 2000; Prenen et al., 2017a; Gumbel, 1960) is given by

$$\psi_\theta(s) = \exp(-s^\theta) \text{ for } 0 < \theta < 1, \quad (2.1.12)$$

with $\psi_\theta^{-1}(s) = (-\log s)^{1/\theta}$.

The GH copula has an upper tail dependence and its Kendall's tau is given by $\tau = 1-\theta$, which means a positive association among event times when $\theta \to 0$ and independence when $\theta \to 1$. Thus, the joint survival function under the GH copula model has also an explicit form:

$$S(t_{i1}, \cdots, t_{in_i}|x_{ij}, \forall j) = \exp\left[-\left\{\sum_{j=1}^{n_i}(-\log S_j(t_{ij}|x_{ij}))^{1/\theta}\right\}^\theta\right]. \quad (2.1.13)$$

2.1.2. Frailty model

Vaupel et al. (1979) introduction the concept of frailty to describe the impact of individual heterogeneity in univariate (independent) survival data. Furthermore, Oakes (1989) provided one way to account for a dependence among survival times within a cluster (or subject). Generally, frailty is a common unobserved random effect that affects multiplicative on the hazard function of survival time.

Definition 2.4. Denote by U_i the unobserved random effect of the ith cluster. The frailty model (Duchateau and Janssen, 2008; Ha et al., 2017) is defined as follows. Given $U_i = u_i$, the conditional hazard function for the survival time is of the form:

$$\lambda_{ij}(t|u_i, x_{ij}) = \lambda_0(t)\exp(x_{ij}^T\beta)u_i, \qquad (2.1.14)$$

where $\lambda_0(\cdot)$ is a specified or unspecified baseline hazard function, and the frailty U_i is assumed to be independently and identically distributed.

Traditionally it is assumed that $E(U_i)=1$ and $\mathrm{var}(U_i)=\theta$ for the gamma frailty model and $V_i = \log U_i \sim N(0,\theta)$ for the lognormal frailty model. Note that $\theta \in [0,\infty)$ means the strength of association among survival times within a cluster. The θ in (2.1.8) is an association parameter of the copula model, whereas the θ in (2.1.14) indicates the frailty variance. Afterwards, it can be seen that the two results are different in (2.1.17). However, when $\theta \to 0$, there is no correlation between survival times, so that the two models give almost the same

results (Duchateau and Janssen, 2008; Ha et al., 2017).

Definition 2.5. The marginal joint survival function of $T_{i1}, \cdots, T_{in_i}$, denoted by $S^*(t_{i1}, \cdots, t_{in_i} | x_{ij}, \forall j)$, can be derived by integrating out the frailty from the conditional survival function, $S(t_{i1}, \cdots, t_{in_i} | x_{ij}, u_i, \forall j)$. Under the conditional independence of $T_{i1}, \cdots, T_{in_i}$ given u_i (Nielsen et al., 1992), we have

$$\begin{aligned} S^*(t_{i1}, \cdots, t_{in_i} | x_{ij}, \forall j) &= \int S(t_{i1}, \cdots, t_{in_i} | x_{ij}, u_i, \forall j) dG_\theta(u_i) \\ &= \int \exp\left[-u_i \sum_{j=1}^{n_i} \psi_\theta^{-1}\{S_j^*(t_{ij} | x_{ij})\} \right] dG_\theta(u_i), \end{aligned} \tag{2.1.15}$$

where $G_\theta(u_i)$ is the distribution function of the frailty U_i, and the jth marginal survival function of the cluster i is as follows:

$$\begin{aligned} S_j^*(t_{ij} | x_{ij}) &= \int S(t_{ij} | x_{ij}, u_i) dG_\theta(u_i) \\ &= \psi_\theta\{\Lambda(t_{ij} | x_{ij})\}. \end{aligned} \tag{2.1.16}$$

Here, $\Lambda(t_{ij} | x_{ij}) = \Lambda_0(t_{ij}) \exp(x_{ij}^T \beta)$ is the cumulative hazard function. Note that the generator ψ_θ of the Archimedean copula is expressed as the Laplace transform of distribution function G_θ of frailty.

The two joint survival functions (2.1.5) and (2.1.15) are similar (Goethals et al., 2008; Prenen et al., 2017a) in that both joint survival functions take the same copula form. However, in the following marginal survival functions, it can be seen that the copula

and frailty models have major differences;

$$S_j(t_{ij}|x_{ij}) \neq S_j^*(t_{ij}|x_{ij}). \qquad (2.1.17)$$

Note also that the association parameter θ shows up in $S_j^*(\cdot|x_{ij})$ of (2.1.16), but not in $S_j(\cdot|x_{ij})$ of (2.1.9). Under the gamma frailty model with mean 1 and variance θ, the marginal joint survival function (2.1.15) has an explicit form:

$$S^*(t_{i1}, \cdots, t_{in_i}|x_{ij}, \forall j) = \left\{ \sum_{j=1}^{n_i} S_j^*(t_{ij}|x_{ij})^{-\theta} - n_i + 1 \right\}^{-1/\theta}, \qquad (2.1.18)$$

where $S_j^*(t_{ij}|x_{ij}) = \{1+\theta\Lambda(t_{ij}|x_{ij})\}^{-1/\theta}$. Therefore, we can clearly confirm the difference in (2.1.17) from the two joint survival functions (2.1.8) and (2.1.16).

2.2. The estimation procedures of copula models

In this section, we study one-stage and two-stage estimation methods in copula survival model. Let $\varphi = (\beta^T, \Lambda_0, \theta)^T$ be the unknown parameters depending on the two methods, where $\Lambda_0 = \Lambda_0(\alpha)$ is the known or unknown baseline cumulative hazard function inherent in the marginal hazard function, and α is the unknown baseline parameter dependent on the Λ_0.

(ⅰ) One-stage estimation procedure

First, the one-stage estimation procedure was proposed by Chen et

al. (2006) and Prenen et al. (2017a) to find the maximum likelihood estimators (MLEs) of φ that maximizes the copula-based log-likelihood function $\ell_c(\varphi)$, as defined in (3.1.3).

Definition 2.6 Under the copula model, the MLEs of $\hat{\varphi}$ is defined as

$$\hat{\varphi} = \arg\max_{\varphi} \ell_c(\varphi), \tag{2.2.1}$$

where argmax denotes the arguments of the maximum. $\hat{\varphi}=\left(\hat{\beta}^T, \hat{\Lambda}_0, \hat{\theta}\right)^T$ and a more detailed procedure for finding the MLEs of (2.2.1) can be found in Prenen et al. (2017a).

Note that $\varphi=\left(\beta^T, \Lambda_0(\alpha), \theta\right)^T$ is obtained by solving the following estimating equations:

$$U(\varphi) = \frac{\partial \ell_c(\varphi)}{\partial \varphi} = 0$$

$$\text{i.e.} \begin{cases} U_\beta(\beta, \Lambda_0, \theta) = \dfrac{\partial \ell_c(\beta, \Lambda_0, \theta)}{\partial \beta} = 0, \\ U_\alpha(\beta, \Lambda_0, \theta) = \dfrac{\partial \ell_c(\beta, \Lambda_0, \theta)}{\partial \alpha} = 0, \\ U_\theta(\beta, \Lambda_0, \theta) = \dfrac{\partial \ell_c(\beta, \Lambda_0, \theta)}{\partial \theta} = 0. \end{cases}$$

(ii) Two-stage estimation procedure

Unlike the one-stage procedure, the two-stage procedure proposed by Shih and Louis (1995) and Andersen (2005) is an approach for estimating the unknown parameter φ by proceeding the following two

steps.

Note that in first step, (β, Λ_0) are estimated by maximizing the classical log-likelihood of (β, Λ_0), denoted by $\ell(\beta, \Lambda_0)$.

In second step, the association parameter θ is estimated by plugging the first-step estimates $(\hat{\beta}, \hat{\Lambda}_0)$ under the marginal hazard in (2.1.14) into the following pseudo likelihood:

$$\ell_c^*(\theta) = \ell_c(\hat{\beta}, \hat{\Lambda}_0, \theta) = \ell(\beta, \Lambda_0, \theta)|_{\beta=\hat{\beta},\ \Lambda_0=\hat{\Lambda}_0},$$

which is then maximized for the association parameter θ. Thus two-stage estimator of θ is obtained by solving

$$U_\theta(\tilde{\beta}, \tilde{\Lambda}_0, \theta) = \frac{\partial \ell_c(\tilde{\beta}, \tilde{\Lambda}_0, \theta)}{\partial \theta} = 0.$$

Note that the main difference between the two estimation methods above is to estimate (β, Λ_0). That is, $(\hat{\beta}, \hat{\Lambda}_0)$ is updated when φ is estimated in the one-stage method of (2.2.1), whereas $(\tilde{\beta}, \tilde{\Lambda}_0)$ is not updated at all in the two-stage estimation method because of the use of $\ell_c^*(\theta)$ in the second step.

2.3. Comparison of Copula and Frailty Models

In this thesis, both one-stage and two-stage estimation methods are used to estimate the copula model. To estimate the frailty model, we can use marginal likelihood (Nielsen et al., 1992) and hierarchical likelihood (h-likelihood; Lee and Nelder, 1996; Ha et al., 2017). The

estimation method's are applied to the simulation in Section 2.3.3 and the illustration in Section 2.3.4.

Let $\varphi = \left(\beta^T, \Lambda_0, \theta\right)^T$ be the unknown parameters dependent on the two models. Here $\Lambda_0 = \Lambda_0(\alpha)$ is the known or unknown baseline cumulative hazard function in the marginal hazard function, and α is an unknown baseline parameter that depends on the parametric function Λ_0.

The marginal likelihood method (Nielsen et al., 1992), which is obtained by integrating the unobserved frailty with the frailty model inference method, has been commonly used. However, if the frailty distribution is not a gamma distribution or if the frailty model is complex, the marginal likelihood method requires a difficult integration. In order to overcome this problem, the h-likelihood method (Lee and Nelder, 1996; Ha et al., 2017) that does not require integration itself has also been proposed. For detailed explanations of these two likelihood approaches to the frailty model, we recommend three books, Hougaard (2000), Duchateau, Janssen (2008), and Ha et al. (2017).

2.3.1. Comparison of R packages

For the fit of the copula and frailty models for analyzing correlated survival data, this section considers three recently developed R packages. For the copula models, we use the **Sunclarco** R package (Prenen et al., 2017b), and for the frailty model, we use the **frailtyEM** R package based on marginal likelihood (Balan and Putter, 2017) and the **frailtyHL** R package based on h-likelihood (Ha et al., 2017, 2018).

The characteristics of the three R packages used in this thesis are

summarized in Table 2.3.1. First, the **Sunclarco** provides one-stage and two-stage procedure methods allowing for parametric or non-parametric (NP) distributions for the baseline hazard function in the marginal hazard function (2.1.6). For the parametric basis distribution, we can use the PE (Piecewise Exponential) which is an exponential distribution with constant hazards within each time interval and the Weibull distribution. For the non-parametric basis distribution, the Cox PH model, where the estimated baseline cumulative hazard is assumed to be a discrete step function, can be used (Breslow, 1972). For the parameteric estimation methods, we use the classical likelihood and for non-parametric methods we use partial likelihood (PL) obtained by eliminating Λ_0.

Secondly, in the case of **frailtyEM**, the marginal likelihood function and EM method are used. Here, the non-parametric method is used for an unknown baseline hazard function, but various parametric distributions, such as gamma distribution and positive stable distribution are allowed for the frailty distribution.

Finally, **frailtyHL** uses a h-likelihood procedure. For the baseline hazard function, a non-parametric method is used as in **frailtyEM**, and a gamma distribution and a log-normal distribution are allowed for the frailty distribution.

Table 2.3.1. R packages for fitting copula and frailty models (Kwon and Ha, 2019)

Model	R package / R function (Literature)	Estimation procedure	Distributions
Copula model	Sunclarco / SunclarcoModel() (Prenen et al., 2017b)	partial likelihood	PE, Weibull, NP
Frailty model	frailtyEM / emfrail() (Balan and Putter, 2017)	marginal likelihood	gamma, PS, IG, CP, PVF
Frailty model	frailtyHL / frailtyHL() (Ha et al., 2018)	h-likelihood	gamma, LN

PE=piecewise exponential ; NP=non-parametric; PS=positive stable; IG=inverse gaussian; CP=compound Poisson; PVF=power variance function; LN=log-normal.

2.3.2. Data description

(1) Kidney infection data

McGilchrist and Aisbett (1991) presented a bivariate survival data set which consists of times to the first and second infections (i.e. $n_i = 2$ for all i) on the same patient among 38 kidney patients using a portable dialysis machine. Infections can occur at the location of insertion of the catheter. The catheter is later removed if any infections occur, and it can also be removed for other reasons, which is regarded as censoring. Here, each survival time is the time to infection since insertion of the catheter. The survival time from the same patient may be correlated due to a common patient effect.

Table 2.3.2 describes the kidney data and provides their basic statistics. Here, in the case of continuous variables, the mean, median and range were summarized, and in the case of categorical variables, the frequency was used. In particular, the censoring rate for infection

time is about 23.7%.

Table 2.3.2. Description and basic statistics of variables for kidney infection data

Variables	Description	Basic statistics
time	time to infection since insertion of the catheter	Mean: 101.6 Median: 39.5 Range (2.0～562.0)
disease	disease type (0=GN, 1=AN, 2=PKD, 3=Other)	GN: 18, AN: 24 PKD: 8, Other: 26
age	age (in years)	Mean: 43.7 Median: 45.5 Range (10.0～69.0)
sex	sex type (1=male, 2=female)	Male: 10, Female: 28
id	Subject' s identification number ($q=38,\ n_i=2$)	
status	event status (0; No infection, 1; infection; censoring rate; about 23.7%)	

(2) Recurrent CGD data

The chronic granulomatous disease (CGD; Fleming and Harrington, 1991) data set is from a placebo-controlled randomized trial of gamma interferon (γ-IFN) in CGD. The trial is aimed to investigate the effectiveness of γ-IFN in reducing the rate of serious infections in the CGD patients. In total, 135 patients from 13 centers (hospitals) were observed for about 1 year. This data set shows that there are recurrences of different cluster sizes (i.e. recurrences were 1 to 8 per patient).

Table 2.3.3 describes the recurrent CGD data and provides their basic statistics as in Table 2.3.2. The censoring rate is about 63%.

Table 2.3.3. Description and basic statistics of variables for the recurrent CGD data

Variables	Description	Basic statistics
treatment	placebo or gamma interferon	placebo: 120 γIFN-g: 83
sex	sex type (male, female)	male : 168 female: 35
age	age (in years) at study entry	Mean: 13.7 Median: 12.0 Range (1.0～44.0)
height	height in cm at study entry	Mean: 138.1 Median: 140.0 Range (76.3～189.0)
weight	weight in kg at study entry	Mean: 39.34 Median: 33.40 Range (10.40～101.50)
inherit	pattern of inheritance	X-linked: 131 autosomal: 72
steroids	use of steroids at study entry,1=yes	0.03448
propylac	use of prophylactic antibiotics at study entry	0.8473
tstart, tstop	start and end of each time interval (tstart, tstop)	Mean: (69.5, 254.1) Median: (140.0, 273.0) Range (76.3～189.0)
id	Subject's identification number (135)	
center	enrolling center (NIH:41, Scripps Institute:36, Amsterdam:28, Univ. of Zurich:21, Mott Children's Hosp:20, L.A. Children's Hosp:13, Other:44)	
status	the event status (0; No infection, 1; infection; censoring; about 63%)	

(3) Bladder cancer data

The bladder cancer data set is the multicenter bladder cancer clinical trial which was conducted by the EORTC (Oddens et al., 2013). The survival data set used in this study was the duration of the disease-free interval (DFI): the time (days) to the first recurrence after surgery (transurethral resection) of 1,066 patients having bladder cancer from 46 centers in 13 European countries. The Bacillus Calmette-Guerin (BCG) was given after surgery to try for

reducing the risk of recurrence. In order to reduce its toxicity, a disadvantage of BCG, the two different doses (1/3 dose and full dose), and the durations of maintenance BCG therapy (1 year and 3 years) were assessed. In this thesis, we aim to evaluate the risk factors for the time to recurrence.

Table 2.3.4. Description and basic statistics variables for the Bladder cancer data

Variables	Description	Basic statistics
timeDFI	Time to the first recurrence after surgery (days)	Mean: 1314.39 Median: 93.5 Range (2.0～4743.0)
statusDFI	Indicator of the recurrence of the bladder cancer (0: No, 1: Yes)	No: 264 Yes: 202
Trtdose	Amount of yhe dose of BCG (1: 1/3 dose BCG, 2: full-dose BCG)	1/3 dose BCG: 245 full-dose BCG: 221
Trtduration	Duration of maintenance (0: 1 year, 1: 3 years)	1 year: 221 3 years: 245
Age	Years	Mean: 75.50 Median: 75 Range (70～85)
Gender	0: Male, 1: Female	Male: 382 Female: 84
TypeBC	Type of the bladder cancer (0: Primary, 1: Recurrent)	Primary: 260 Recurrent: 206
Tumsize	Largest tumor diameter (mm)	Mean: 18.14 Median: 15 Range (2～98)
Nbtum	No. of tumors	Mean: 2.95 Median: 2 Range (1～10)
Tstage	T category of the bladder cancer (0: pTa, 1: pT1)	pTa: 279 pT1: 187
Ggrade	WHO grade of the bladder cancer (1: G1, 2: G2, 3: G3)	G1: 122 G2: 202 G3: 142
patientid	subject patients number (1,066)	
institution	46 institution in 13 European countries.	
status	1=the recurrence of the patients (44.3%)	

Table 2.3.4 provides a description of the variables used in this

analysis; the censoring variable indicates whether a recurrence was observed or not, with a recurrence being noted in 44.3% of the patients, leading to censoring rate 54.7%.

2.3.3. Simulation study

Below are the simulation design and estimation results for comparing copula model and frailty model for correlated survival data.

First, the method of Prenen et al. (2017a) based on the sampling algorithm of Marshall and Olkin (1988) was used to generate survival time data under the Clayton copula model. The correlation parameter of Clayton copula model is set with a small correlation strength $\theta = 0.1$ (i.e. Kendall's tau $\tau \fallingdotseq 0.048$) and a slightly larger $\theta = 1.0$ (ie $\tau \fallingdotseq 0.333$) respectively. The standard exponential distribution was assumed for the baseline distribution of the marginal hazard function, and the standard normal distribution was used for one single covariate x. The corresponding regression parameter was fixed as $\beta = 1$, the censoring time was generated from a uniform distribution with about 20% censoring rate, and the sample size was considered as $(q, n_i) = (100, 4)$ for all $i = 1, \cdots, q$.

Next, in the case of the frailty model, survival time data were generated under the gamma frailty model using the simulation scheme by Ha et al. (2019). The frailty U_i $(i = 1, \cdots, 100)$ was generated from the gamma distribution considering the mean 1 and the variance $\theta = 0.1$ and 1.0. The rest of the design is the same as the design of Clayton copula model presented above, and the number of replications of the simulation was 200 times.

The mean, mean of estimated standard errors(SE), and standard deviation (SD) were calculated for each of $\hat{\beta}$ and $\hat{\theta}$, respectively.

The one-stage and two-stage copula modeling methods using the Sunclarco package were performed as shown in Table 2.3.5. As the baseline distribution of the marginal hazard function (2.1.6), the PE and Weibull in one-stage and Weibull and NP (Cox) distributions in two-stage were applied, respectively. For the fit and comparison of the models, the Clayton copula model was first used as a true model and two models (Clayton copula model, gamma frailty model) were fitted. Similarly, the gamma frailty model was taken as a true model and then the two models were fitted.

The abbreviations and notations used in the tables of this thesis are as follows:

- Est: estimator
- Mean and SD: mean and standard deviation for estimates
- SE: mean of estimated standard error

The simulation results are shown in Table 2.3.5, and the results are shown below.

(i) $\theta = 0.1$: When the Clayton copula model is assumed as a true model, β is estimated well in terms of small bias of $\hat{\beta}$ in both Clayton copula and the gamma frailty models. However, θ is overestimated in the two-stage method (Weibull and Cox) of Clayton copula model. Assuming that the gamma frailty is a true model, both β and θ are well estimated for fitting the gamma frailty model, but fitting Clayton copula model seems to be relatively slightly underestimated for β. When the correlation strength is small (i.e. $\theta = 0.1$), it is observed that the estimation results of the two models (i.e. Clayton and gamma

frailty models) are overall well estimated.

(ii) $\theta = 1.0$: When the Clayton copula model is assumed as the true model, β and θ could be estimated well in terms of small biases of $\hat{\beta}$ and $\hat{\theta}$ in fitting Clayton copula model, but for the gamma frailty model, β is overestimated and θ is underestimated. Assuming that the gamma frailty is a true model, fitting the gamma frailty model is well estimated, whereas fitting the Clayton copula model tends to estimate both the one-stage and two-stage estimation for β erroneously, and leads to an underestimated of θ.

Therefore, according to Table 2.3.5 when the correlation strength is small as $\theta = 0.1$, it was observed that the estimation results of the two models are generally good in estimating parameters even if each model is incorrectly fitted. However, when the correlation strength is high as $\theta = 1.0$, the estimation results of the two models are relatively sensitive to the true model. That is, when $\theta = 1.0$, if both models fit the true model, they fit well, but if they are incorrectly specified, both models show a large bias for the estimator. From these results, it can be seen that when the correlation between the survival data is small, the two models give similar results. However, when the correlation is large, the two models give different results; this means that fitting a proper model for the clustered survival data is important in model-based data analysis.

Table 2.3.5. $(q,\ n_i)=(100,\ 4)$: Simulation results on the estimation for correctly or incorrectly fitted model when the true model is the Clayton copula model or the gamma frailty model, respectively; $\beta=1$

Baseline hazard function		Fitted model											
		Clayton								Gamma frailty			
		One-stage				Two-stage							
		Weibull		PE		Weibull		Cox		frailtyEM		frailtyHL	
True model	Est	Mean	SE SD	Mean	SE SD	Mean	SE SD	Mean	SE SD	Mean	SE SD	Mean	SE SD
$\theta=0.1$													
Clayton	$\hat{\beta}$	1.013	0.070 0.068	1.024	0.067 0.071	1.014	0.070 0.069	1.012	0.069 0.072	1.058	0.084 0.079	1.064	0.076 0.080
	$\hat{\theta}$	0.102	0.066 0.065	0.102	0.066 0.066	0.188	0.068 0.058	0.109	0.068 0.058	0.075	0.061 0.056	0.087	0.061 0.058
Gamma frailty	$\hat{\beta}$	0.965	0.070 0.073	0.962	0.072 0.075	0.964	0.072 0.075	0.956	0.073 0.077	1.010	0.084 0.084	1.015	0.077 0.084
	$\hat{\theta}$	0.095	0.065 0.067	0.098	0.067 0.069	0.099	0.066 0.060	0.102	0.058 0.061	0.089	0.067 0.062	0.102	0.066 0.064
$\theta=1$													
Clayton	$\hat{\beta}$	1.016	0.184 0.197	1.024	0.187 0.200	1.008	0.075 0.073	1.004	0.078 0.078	1.422	0.100 0.097	1.428	0.093 0.097
	$\hat{\theta}$	1.009	0.054 0.064	1.010	0.067 0.064	1.001	0.186 0.189	0.994	0.179 0.190	0.742	0.153 0.156	0.762	0.147 0.158
Gamma frailty	$\hat{\beta}$	0.631	0.054 0.064	0.599	0.055 0.060	0.620	0.076 0.080	0.636	0.076 0.078	0.999	0.091 0.087	1.004	0.086 0.088
	$\hat{\theta}$	0.859	0.175 0.218	0.943	0.190 0.214	0.835	0.202 0.196	0.922	0.203 0.195	0.995	0.189 0.183	1.022	0.185 0.187

2.3.4. Illustration

In this section, we consider three real data sets for the correlated survival data described in Section 2.3.2. The first data set is kidney infection data (McGilchrist et al., 1991) with the same cluster size, and second one is CGD recurrence data (Fleming et al., 1991) with different cluster sizes. The third one is bladder cancer data (Oddens et al. 2013; Park and Ha, 2019) from a multicenter clinical trial. As

shown in the simulation in Section 2.3.3, the Clayton copula model uses the Sunclarco package, and the gamma frailty model uses two packages (frailtyHL, frailtyEM). We fit the two models (Clayton copula and gamma frailty models) on three real data sets and compare the results. In particular, the estimated regression coefficients of the two models and the estimated associated parameters are compared in terms of the validity and sensitivity to the fitted results of model.

(1) Kidney infection data

Table 2.3.6 shows the results of fitting the Clayton copula model and the gamma frailty model with Age (age) and Sex (1 = male, 2 = female) as covariates. Following the results of the Wald test statistic, $(\text{Estimate}/\text{SE})^2$, the Age effect is not significant at the 5% significance level in both models (Clayton copula model, gamma frailty model). However, the Sex effect is crucial in one-stage estimation and gamma frailty model, whereas it is not significant in two-stage estimation (Weibull and Cox). The estimated value of the correlation coefficient θ is about 0.2 for the Clayton copula between the two models, and about twice or three times of the gamma frailty model (i.e. 0.397 in frailtyEM and 0.561 in frailtyHL). The estimate of the correlation parameter θ is reflected to the estimates of the regression parameters (effects of age and sex) in one-stage estimation and the estimation of gamma frailty model, as shown in the simulation results in Table 2.3.5 for both models. As a result, there seems to be a large difference in the absolute values of the regression estimates between the two models.

Table 2.3.6. Estimation results of fitting Clayton copula and gamma frailty models for kidney data

Baseline hazard function	Clayton								Gamma frailty			
	One-stage				Two-stage							
	Weibull		PE		Weibull		Cox		frailtyEM		frailtyHL	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Age	0.003	0.010	0.002	0.010	0.004	0.009	0.002	0.008	0.005	0.012	0.007	0.013
Sex: F	-0.937	0.301	-0.947	0.309	-0.875	0.510	-0.829	0.483	-1.553	0.445	-1.691	0.483
θ	0.207	0.196	0.205	0.212	0.211	0.473	0.209	0.110	0.397	0.235	0.561	0.280

(2) Recurrent CGD data

Table 2.3.7 shows the results of fitting the two models with Age (age) and treatment (Treat; 0=false drug, 1 = γ-IFN) as covariates. The estimated correlation parameter $\hat{\theta}$ is around 0.1 in both models, which is relatively small as compared to the results in Table 2.3.5. In the simulation results in Table 2.3.5, the estimated regression parameters of both models are similar. In particular, according to the results of the Wald test statistic, the treatment (i.e. γ-IFN) is very significant at the 5% significance level in both models.

Table 2.3.7. Estimation results of fitting Clayton copula and gamma frailty models for CGD data

Baseline hazard function	Clayton								Gamma frailty			
	One-stage				Two-stage							
	Weibull		PE		Weibull		Cox		frailtyEM		frailtyHL	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Age	-0.026	0.013	-0.025	0.013	-0.029	0.004	-0.029	0.011	-0.029	0.011	-0.027	0.014
Treat	-1.088	0.261	-1.049	0.258	-1.092	0.049	-1.052	0.138	-1.114	0.168	-1.130	0.270
θ	0.093	0.107	0.110	0.120	0.069	0.000	0.184	0.147	0.184	0.000	0.090	0.109

(3) Multicenter bladder cancer data

Table 2.3.8 shows the results of fitting the Clayton copula model and the gamma frailty model for bladder cancer data. The estimated correlation parameter $\hat{\theta}$ is 0.093 in the one-stage PE of the Clayton copula model, which is less than 0.114 in **frailtyEM** and 0.129 in **frailtyHL** under the gamma frailty model. Following penalized variable selection by Park and Ha (2019), significant variables in bladder cancer data were known as Trtduration, TypeBC, Nbtum, and G1. In the Table 2.3.8, according to the results of the Wald test statistic, Trtduration, TypeBC, and Nbtum are very significant at the 5% significance level for both models. However, the G1 is all significant except for one-stage Weibull. In addition, Trtdose is significant in one- and two-stage Weibull and two-stage Cox.

Table 2.3.8. Estimation results of fitting Clayton copula and gamma frailty models for bladder cancer data

Baseline hazard function	Clayton								Gamma frailty			
	One-stage				Two-stage							
	Weibull		PE		Weibull		Cox		frailtyEM		frailtyHL	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Trtdose	-0.181	0.089	-0.144	0.090	-0.148	0.064	-0.153	0.060	-0.153	0.093	-0.153	0.093
Trtdu -ration	-0.339	0.090	-0.193	0.091	-0.229	0.083	-0.204	0.079	-0.198	0.094	-0.197	0.094
Age	-0.003	0.004	-0.002	0.004	-0.001	0.005	-0.003	0.004	-0.002	0.005	-0.002	0.005
Cender	0.182	0.110	0.167	0.114	0.147	0.152	0.162	0.146	0.180	0.118	0.182	0.118
TypeBC	0.383	0.101	0.375	0.102	0.386	0.080	0.386	0.076	0.391	0.106	0.392	0.106
Tumsize	0.000	0.004	0.002	0.004	-0.001	00004	0.000	0.004	0.003	0.004	0.003	0.004
Nbtum	0.147	0.024	0.129	0.025	0.136	0.032	0.135	0.030	0.137	0.026	0.137	0.026
Tstage	0.206	0.122	0.069	0.125	0.021	0.146	0.038	0.137	0.076	0.128	0.078	0.129
G1	-0.196	0.159	-0.318	0.159	-0.344	0.167	-0.302	0.152	-0.326	0.164	-0.326	0.164
G2	-0.128	0.137	-0.215	0.137	-0.259	0.134	-0.213	0.125	-0.217	0.805	-0.217	0.142
θ	0.168	0.057	0.093	0.052	0.184	0.053	0.183	0.000	0.114	0.064	0.129	0.069

Ⅲ. ESTIMATION OF COPULA SURVIVAL MODELS

In this chapter, we propose a one−stage procedure for estimating the regression and association parameter under copula survival models in Section 2.1.1. Specifically, we use the M−spline method for estimating the baseline hazards.

3.1. Copula−based likelihood

Let T_{ij} and C_{ij} be the survival time and censoring time for the j −th observation of the i−th cluster(or subject) $(i=1, \cdots, q;\ j=1, \cdots, n_i)$, respectively. Here, q is the number of clusters, and n_i is the number of individuals in cluster i and $N=\sum_{i=1}^{q} n_i$ is the total sample size.
The observable random variables in the clustered survival data are;

$$Y_{ij}=\min(T_{ij},\ C_{ij}) \text{ and } \delta_{ij}=I(T_{ij} \le C_{ij}), \tag{3.1.1}$$

where C_{ij} is the censoring time corresponding to event time T_{ij}. In this thesis, we assume the following two usual assumptions in survival analysis (Prenen et al., 2017a; Ha et al., 2017).

Assumption (A1): Given covariates x_{ij}, T_{ij} and C_{ij} are conditionally independent, and pairs $(T_{ij},\ C_{ij})$ are also conditionally independent $(i=1, \cdots, q;\ j=1, \cdots, n_i)$.

Assumption (A2): Given covariates x_{ij}, C_{ij} are conditionally

noninformative of T_{ij}.

Let y_{ij} be the observed value of Y_{ij}. According to Prenen et al. (2017a), the contribution of cluster i to the likelihood function, denoted by L_i, is obtained from the derivative of the n_i-dimensional joint survival function over all uncensored individuals in the cluster i;

$$L_i = (-1)^{d_i} \frac{\partial^{d_i}}{\partial\{\delta_{ij}=1\}} S\left(y_{i1}, \cdots, y_{in_i} | x_{ij}, \ \forall j\right), \ j=1, 2, \cdots, n_i \tag{3.1.2}$$

where $\{\delta_{ij}=1\}$ is the set of uncensored individuals in the cluster i, and $d_i = \sum_{j=1}^{n_i} \delta_{ij}$ is the size of this set. We denote the marginal survival and density functions given x_{ij} as $S_{ij} = S_j(y_{ij}|x_{ij})$ and $f_{ij} = f_j(y_{ij}|x_{ij}) = -S_{ij}^{'}$, respectively. Since the generator of Archimedean ψ_θ is the Laplace transform of G_θ, from (2.1.5) and (3.1.2) the copula-based log-likelihood ℓ_c, for $\forall n$, individuals is as follows:

$$\begin{aligned} \ell_c &= \sum_{i=1}^{q} \ell_i \\ &= \sum_{i=1}^{q} \left[\sum_{j=1}^{n_i} \delta_{ij} \left\{ \log f_{ij} - \log \psi'_\theta \left\{ \psi_\theta^{-1}(S_{ij}) \right\} \right\} + \log \psi_\theta^{(d_i)} \left\{ \sum_{j=1}^{n_i} \psi_\theta^{-1}(S_{ij}) \right\} \right], \end{aligned} \tag{3.1.3}$$

where $\ell_i = \log L_i$. Here, under the marginal hazard model (2.1.6) $\lambda_{ij}(t|x_{ij}) = \lambda_0(t)\exp\left(x_{ij}^T \beta\right)$ we have that

$$f_{ij} = \lambda_0(y_{ij})\exp\left(x_{ij}^T \beta\right) S_{ij} \text{ and } S_{ij} = \exp\left\{ -\Lambda_0(y_{ij}) e^{x_{ij}^T \beta} \right\}. \tag{3.1.4}$$

Definition 3.1. The $k-$th derivative of the Clayton copula $\psi_\theta(s)=(1+\theta s)^{-1/\theta}$ is given by

$$\psi_\theta^{(k)}(s)=(-1)^k(1+\theta s)^{-(k+1/\theta)}\prod_{a=0}^{k-1}(1+a\theta). \tag{3.1.5}$$

In the Clayton copula model with the Weibull marginal hazard model (2.1.6), the log-likelihood (3.1.3) has a closed form;

$$\begin{aligned}\ell_c &= \sum_{ij}\delta_{ij}\left\{\log\lambda_0(y_{ij})+x_{ij}^T\beta-\theta\log S_{ij}\right\}\\ &-\sum_{i=1}^{q}\left[(d_i+\theta^{-1})\log\left(\sum_j S_{ij}^{-\theta}-n_i+1\right)-\sum_{l=0}^{d_i-1}\log(1+l\theta)\right],\end{aligned} \tag{3.1.6}$$

Definition 3.2. The $k-$th derivative of the GH copula $\psi_\theta(s)=\exp(-s^\theta)$ $(0<\theta<1)$ is given by

$$\psi_\theta^{(k)}(s)=\left(\sum_{a=1}^{d_i}\left[\sum_{b=1}^{a}\frac{d_i!}{a!}(-1)^{d_i-b}\binom{a}{b}\binom{b\theta}{d_i}\right]s^{a\theta-d_i}\right)\psi_\theta(s). \tag{3.1.7}$$

We can show that under the GH copula model, the corresponding likelihood also has an explicit form:

$$\begin{aligned}\ell_c &= \sum_{i=1}^{q}\sum_{j=1}^{n_i}\delta_{ij}\left(\log f_{ij}-\log\left\{\theta S_{ij}(-\log S_{ij})^{1-1/\theta}\right\}\right)-\left(\sum_{j=1}^{n_i}(-\log S_{ij})^{1/\theta}\right)^\theta\\ &+\sum_{i=1}^{q}\log\left(\sum_{a=1}^{d_i}\left[\sum_{b=1}^{a}\frac{d_i!}{a!}(-1)^{d_i-b}\binom{a}{b}\binom{b\theta}{d_i}\right]\left[\sum_{j=1}^{n_i}(-\log S_{ij})^{1/\theta}\right]^{a\theta-d_i}\right).\end{aligned} \tag{3.1.8}$$

Definition 3.3. Under the Clayton copula model with the Weibull marginal hazard model (2.1.6), the log−likelihood (3.1.6) has a closed form:

$$\ell_c = \sum_{ij}\delta_{ij}\{\log\lambda_{ij} + \theta\Lambda_{ij}\} - \sum_{i=1}^{q}\left[(d_i + \theta^{-1})\log(1 + S_{i+}^{*}) - \sum_{a=0}^{d_i-1}\log(1 + a\theta)\right], \quad (3.1.9)$$

where $\lambda_{ij} = \lambda_0\phi y_{ij}^{\phi-1}\exp(x_{ij}^T\beta)$, $\Lambda_{ij} = \Lambda_0(y_{ij})\exp(x_{ij}^T\beta) = \lambda_0 y_{ij}^{\phi}\exp(x_{ij}^T\beta)$ and $S_{i+}^{*} = \sum_{j=1}^{n_i}(S_{ij}^{-\theta} - 1)$.

In the copula survival models, two types of estimation methods, one− and two−stage methods, have commonly been used. Let $\varphi = (\beta^T, \lambda_0, \theta)^T$ be unknown parameters in the copula survival models. Here, λ_0 is the parameter in the baseline hazard $\lambda_0(t)$. For example, $\lambda_0 = (\theta_0, \phi)^T$ in the Weibull case. The one−stage estimation procedure (Prenen et al., 2017a) is performed by maximizing $\ell_c(\varphi)$ in (3.1.3). In other words, one−stage MLEs $\hat{\varphi}$ are defined by

$$\hat{\varphi} = \arg\max_{\varphi}\ell_c(\varphi). \quad (3.1.10)$$

Here $\hat{\varphi} = (\hat{\beta}^T, \hat{\lambda}_0, \hat{\theta})^T$ reflects the dependence among the survival data. The two−stage estimation procedure (Shih and Louis, 1995; Duchateau and Janssen, 2008) consists of two steps. As mentioned in Section 2.2, the two−stage estimation procedure estimates (β, λ_0) easily in the first step, but it estimates the association parameter θ without updating (β, λ_0) in the second step. The details of one− and

two-stage estimation procedures will be described in later sections.

3.2. M-spline modeling for baseline hazards

When the functional form of marginal baseline hazard $\lambda_0(t)$ in (2.1.6) is unknown, $\lambda_0(t)$ (or $\Lambda_0(t)$) has originally infinite dimensional parameters. The estimation of the parameters of interest (β, θ) is in the presence of the nuisance $\lambda_0(\cdot)$ under the copula model (2.1.5) with (2.1.6). However, the estimation of $\lambda_0(t)$ by direct maximization of ℓ_c in (3.1.3) is difficult because of its dimensional issue. To overcome this problem, we consider a M-spline function with finite dimensional parameters via a computationally efficient M-spline method for $\lambda_0(t)$ (Ramsay, 1988; Emura et al., 2017).

Definition 3.4. The cubic M-spline for the marginal baseline hazard $\lambda_0(t)$ is specified as

$$\lambda_0(t;h)=\sum_{l=1}^{L} h_l M_l(t), \tag{3.2.1}$$

where $h=(h_1, \cdots, h_L)^T$ and h_l's are unknown positive parameters, and $M_l(t)$'s are called the M-spline basis functions (Ramsay, 1988). Here, the number of bases L also represents the number of free variables.

The M-spline introduced by Ramsay(1988) can be considered as a normalized of B-spline (basis spline: $B_l(t)$) with unite integral within boundary knots. That is, an M-spline polynomial with order k can be expressed as

$$M_l(t) = k \cdot \frac{B_l(t)}{(t_{l+k} - t_l)}, \ (l = 1, \cdots, L).$$

Definition 3.5. The corresponding baseline cumulative hazard function and survival functions are, respectively given by

$$\Lambda_0(t; h) = \sum_{l=1}^{L} h_l I_l(t), \tag{3.2.2}$$

and

$$S_0(t; h) = \exp\left\{ - \sum_{l=1}^{L} h_l I_l(t) \right\}, \tag{3.2.3}$$

where $I_l(t)$ is the integration of $M_l(t)$, called the I-spline (or integrated spline) basis functions (Ramsay, 1988).

Following the suggestions of Emura et al. (2017, 2019a) and Commenges and Jacqmin-Gadda (2016), we use the number $L = 5$, giving a five-parameter spline model with flexible functional forms (Appendix A). This number allows flexible shapes, including increasing, decreasing, constant, convex, and unimodal hazard functions, keeping the over-fitting phenomenon. The characteristics of the flexible shape's parameter specifications are shown in the following Example 3.6 (Shih and Emura, 2020).

Example 3.6. For a five-parameter spline model with flexible functional forms, let

$$\Theta_1 = \left\{ h : 0 \le \frac{-3(4h_1 - 3h_2 + h_3)}{a_1(h)} < 1 \right\}, \ \Theta_2 = \left\{ h : 0 \le \frac{-3(h_2 - 3h_3 + h_4)}{2a_2(h)} \le 1 \right\}.$$

Then, for all $h_l > 0$, $l = 1, 2, \cdots, 5$, the proposed spline model yields the following shapes of hazard functions;

(i) Increasing hazard:

$$2h_1 \le h_2 \le h_4 \le 2h_5,\ \left\{6(h_2 - 2h_1) - 9(4h_1 - 3h_2 + h_3)^2 / a_1(h)\right\} I(h \in \Theta_1) \ge 0$$
$$\text{and } \left[3(h_4 - h_2)/2 - 9(h_2 - 2h_3 + h_4)^2 / \{4a_2(h)\}\right] I(h \in \Theta_2) \ge 0,$$

(ii) Decreasing hazard:

$$2h_1 \ge h_2 \ge h_4 \ge 2h_5,\ \left\{6(h_2 - 2h_1) - 9(4h_1 - 3h_2 + h_3)^2 / a_1(h)\right\} I(h \in \Theta_1) \le 0$$
$$\text{and } \left[3(h_4 - h_2)/2 - 9(h_2 - 2h_3 + h_4)^2 / \{4a_2(h)\}\right] I(h \in \Theta_2) \le 0,$$

(iii) Convex hazard:

$$4h_1 - 3h_2 + h_3 \ge 0,\ h_2 - 2h_3 + h_4 \ge 0 \text{ and } h_3 - 3h_4 + 4h_5 \ge 0,$$

(iv) Concave hazard:

$$4h_1 - 3h_2 + h_3 \le 0,\ h_2 - 2h_3 + h_4 \le 0 \text{ and } h_3 - 3h_4 + 4h_5 \le 0,$$

(v) Constant hazard:

$$2h_1 = h_2 = h_3 = h_4 = 2h_5,$$

where $a_1(h) = -12h_1 + 10.5h_2 - 6h_3 + 1.5h_4$ and $a_2(h) = -1.5h_2 + 6h_3 - 10.5h_4 + 12h_5$.

Definition 3.7. We define the basis functions $M_l(t)$'s and $I_l(t)$'s on the support $t \in [\xi_1, \xi_3]$, where ξ_1 is the lower limit, ξ_3 is the upper limit, and $\xi_2 = (\xi_1 + \xi_3)/2$ is the midpoint. In data analysis, one can choose $\xi_1 = \min(y_{ij})$ and $\xi_3 = \max(y_{ij})$.

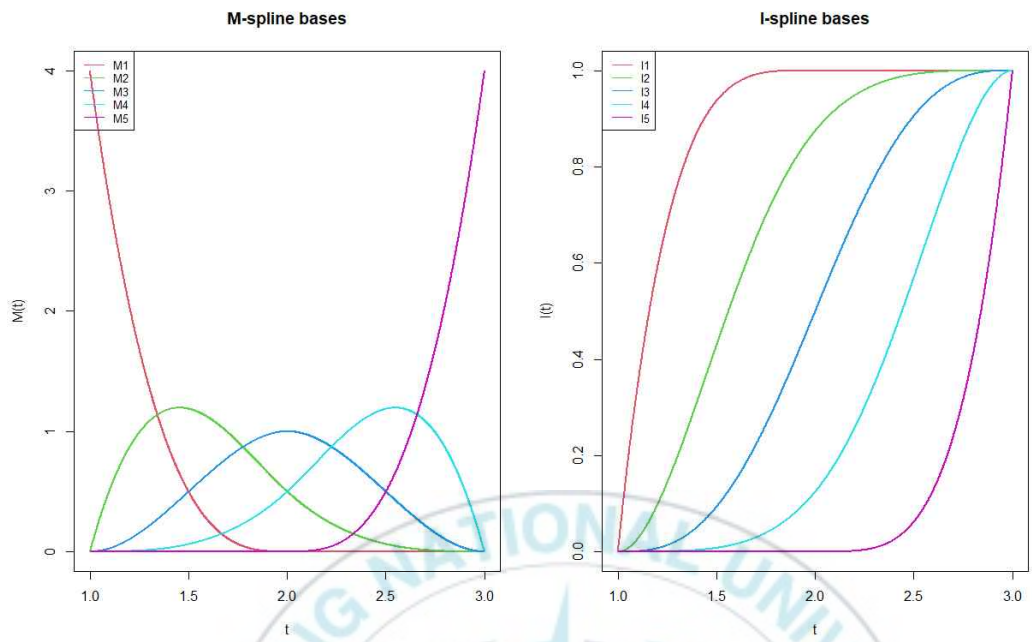


Figure 3.2.1. M-spline (left) and I-spline basis functions (right) with $L=5$

Figure 3.2.1 displays the M- and I-spline basis functions with $L=5$ and the knots $\xi_1=1$, $\xi_2=2$, and $\xi_3=3$ (Emura et al., 2017). The joint.Cox R package (Emura, 2019b) provides M.spline() and I.spline() functions and allows the calculation of $M_l(t)$ and $I_l(t)$. In the univariate PH regression model for independent survival data, the M-spline method provides essentially the same estimation results for regression parameters as the semi-parametric Cox's (1972) regression method using the partial likelihood (Shih and Emura, 2020).

3.3. One-stage estimation procedure

Now, we propose a one-stage estimation procedure for semi-parametric

copula model (2.1.5) with unknown marginal baseline hazard using the M-spline method proposed by Emura et al. (2017, 2019a). Let $\varphi=(\beta^T, h^T, \theta)^T$ be unknown parameters in the copula model. Here, $h=(h_1, \cdots, h_5)^T$ is a vector of unknown positive parameters in (3.2.1).

For the existence of MLEs, we add the following assumption.

Assumption (A3): $\ell_c(\varphi)$ in (3.1.3) is continuous on Ω.

Note here that $\Omega=\{(\beta^T, h^T, \theta)^T | \beta\in\mathbb{R}^p, h_i>0\ (i=1,2,3,4,5),\ \theta>0\}$ is the parameter space which has finite dimension.

Theorem 3.8. Under the two assumptions (A1) and (A2), the M-spline-based log-likelihood $\ell_c(\varphi)$ of $\varphi=(\beta^T, h^T, \theta)^T$ under the Clayton copula model can be expressed as

$$\ell_c(\varphi)=\sum_{ij}\delta_{ij}\left\{\log\lambda_0(y_{ij};\boldsymbol{h})+x_{ij}^T\beta-\theta\Lambda_0(y_{ij};\boldsymbol{h})e^{x_{ij}^T\beta}\right\}$$
$$-\sum_{i=1}^{q}\left[(d_i+\theta^{-1})\log\left(\sum_j\exp\left\{\theta\Lambda_0(y_{ij};\boldsymbol{h})e^{x_{ij}^T\beta}\right\}-n_i+1\right)-\sum_{l=0}^{d_i-1}\log(1+l\theta)\right], \quad (3.3.1)$$

where the M-spline based hazards $\lambda_0(y_{ij};\boldsymbol{h})$ and $\Lambda_0(y_{ij};\boldsymbol{h})$ are given in (3.2.1) and (3.2.2), respectively. Here $h=(h_1, \cdots, h_5)^T$ is a vector of the unknown positive parameters and φ $(\in\Omega)$ is an unknown true parameter.

Under (A1)-(A3), the one-stage MLEs $\hat{\varphi}$ of φ are obtained by

$$\hat{\varphi}=\arg\max_{\varphi\in\Omega}\ell_c(\varphi). \quad (3.3.2)$$

Proof. Under (A1), and (A2), the log-likelihood (3.3.1) based on the M-spline is constructed by replacing marginal baseline hazard $\lambda_0(t)$ in (3.1.3) and marginal cumulative baseline hazard $\Lambda_0(t)$ in (3.1.4) with the M-spline hazard (3.2.1) and I-spline hazard (3.2.2), respectively. Under (A3), the MLEs of $\varphi=(\beta^T,\ h^T,\ \theta)^T$ exists (Cox and Hinkley, 1974). Thus, we can find the MLEs $\hat{\varphi}=(\hat{\beta}^T,\ \hat{h}^T,\ \hat{\theta})^T$ by solving the following three estimating equations simultaneously:

$$\begin{cases} \dfrac{\partial \ell_c(\varphi)}{\partial \beta}=0, \\ \dfrac{\partial \ell_c(\varphi)}{\partial h}=0, \\ \dfrac{\partial \ell_c(\varphi)}{\partial \theta}=0. \end{cases}$$

These three estimating equations are non-linear with respect to $\varphi=(\beta^T,\ h^T,\ \theta)^T$. □

Property 3.9.

(i) The estimated SEs of $\hat{\varphi}$, denoted by $SE(\hat{\varphi})$, can be obtained directly because the estimated asymptotic variances of $(\hat{\beta},\hat{\theta})$ are obtained from the inverse of the observed information matrix $-\partial^2\ell_c/\partial\varphi\partial\varphi^T|_{\varphi=\hat{\varphi}}$ (Cox and Hinkley, 1974).

(ii) The one-stage MLEs $\hat{\varphi}=(\hat{\beta}^T,\hat{h}^T,\hat{\theta})^T$ reflects the dependence among the survival times. Here, (β,θ) are parameters of interest, but

h is a vector of nuisance parameters.

For the implementation of (3.3.2), we use the optim() R function, including the computation of the estimated SEs from the asymptotic variance above. Here, $\lambda_0(t;h)$ and $\Lambda_0(t;h)$ in (3.3.1) are easily calculated using the **joint.Cox** R package. These facts indicate that the proposed procedure is easily implemented with existing algorithms such as, the **joint.Cox** R packages. We have found through simulation studies in Chapter 4 and illustrations in Chapter 5 that our one-stage procedure provides a very fast fitting algorithm with the number of bases $L=5$ in (3.2.1).

In fact, the proposed one-stage semi-parametric procedure can also be viewed as computationally efficient sieve maximum likelihood (ML) approach (Grenander, 1981; Geman and Hwang, 1982). In this aspect, the proposed procedure may also be constructed by replacing an infinite-dimensional parameter space for the unknown baseline hazard function $\lambda_0(t)$ with a finite-dimensional parameter space (i.e. $L=5$) through the M-spline function $\lambda_0(t;h)$ in (3.2.1) (Ma et al., 2015; Chen et al., 2017).

3.4. Two-stage estimation procedures

Definition 3.10. The two-stage estimation procedure consists of the following two steps.

First step. The parameters (β, λ_0), where $\lambda_0 = \lambda_0(\cdot)$, are easily estimated using the classical ordinary censored likelihood, assuming that all individuals as independent according to the marginal hazard model (2.1.6), where the baseline hazard λ_0 can be parametric or

non-parametric.

Second step. The copula association parameter θ is estimated by maximizing the pseudo log-likelihood, defined by $\ell_c^*(\theta)$,

$$\ell_c^*(\theta) = \ell_c\left(\tilde{\beta}, \tilde{\lambda}_0, \theta\right), \tag{3.4.1}$$

where $\ell_c(\beta, \lambda_0, \theta)$ is given by (3.1.3) and

$$\ell_c\left(\tilde{\beta}, \tilde{\lambda}_0, \theta\right) = \ell_c(\beta, \lambda_0, \theta)|_{\beta=\tilde{\beta}, \lambda_0=\tilde{\lambda}_0}. \tag{3.4.2}$$

Here $\tilde{\beta}$ and $\tilde{\lambda}_0$ are MLEs obtained in the first step, but they are fixed in the second step for estimating θ, so that they are not updated. In this regard, in the two-stage estimation, the estimates of the regression and baseline parameters (β, λ_0) are equal to the estimates arising from the marginal hazard model (2.1.6) which leads to an independence model with

$$S\left(t_{i1}, \cdots, t_{in_i} | x_{ij}, \forall j\right) = S(t_{i1}|x_{i1}) \cdots S\left(t_{in_i}|x_{in_1}\right). \tag{3.4.3}$$

Property 3.11.

(i) The two-stage estimation of (β, λ_0) is based on the marginal hazard model (2.1.6) with independent event times, rather than the copula model (2.1.5) allowing for dependency among event times.

(ii) The resulting two-stage estimates $\left(\tilde{\beta}, \tilde{\lambda}_0, \tilde{\theta}\right)$ may not effectively reflect the dependence among the survival data. The likelihood function produces a direct variance estimate that is not valid, especially when there is a strong dependency among the data.

In particular, in two-stage semi-parametric estimation (Prenen et al., 2017a), the regression parameters and baseline hazards are estimated under the PH model of the marginal Cox in the first step, giving the same estimates as the generalized estimation equation (GEE; Liang and Zeger, 1986).

Using the robust sandwich estimator of the GEE approach introduced by Liang and Zeger (1986), the two-stage method gives a variance-covariance matrix of estimated regression parameters that account for the dependency due to clustering. Even the two-stage method considers within-cluster correlation through robust variance estimates by the GEE approach, it cannot explicitly explain the strength of these correlations.

3.5. Comparison of one-stage and two-stage procedures

In this section, we compare the proposed one-stage and the two-stage estimation procedures.

As mentioned in Section 3.4, the two-stage estimation procedure consists of two steps. In the first step, (β, λ_0), where the baseline hazard λ_0 can be parametric or non-parametric, are easily estimated using the classical right-censored likelihood by assuming all individuals as independent under the marginal hazard model (2.1.6). In the second step, the copula association parameter θ is estimated by maximizing a pseudo log-likelihood $\ell_c^*(\theta)$ in (3.4.1) which is used to estimate the dependence parameter θ, but when implementing $\ell_c^*(\theta)$, it is not updated (i.e. fixed) the estimates of regression parameter and baseline hazard from the first step. Thus, the two-stage estimation of

(β, λ_0) is based on the marginal hazard model (2.1.6) with independent event times, not in the copula survival model (2.1.5) allowing for dependency among event times, so that this pseudo-likelihood approach using $\ell_c^*(\theta)$ may not be effective to the estimation of unknown parameters. Since our one-stage procedure estimates simultaneously marginal and dependence parameters using $\ell_c(\varphi)$ in (3.3.1), the proposed one-stage estimates including our SEs properly reflect such dependence.

To evaluate the performance of our proposed method, the simulation study in Chapter IV is conducted, with the illustration in Chapter V. As can be seen in these two chapters, we recommend using an M-spline-based one-step procedure that avoids the potential loss of efficiency of the two-step estimation.

Ⅳ. SIMULATION STUDY FOR COPULA SURVIVAL MODELS

In this chapter, the simulation study is conducted to evaluate the performance of the proposed one-stage estimation method under the copula survival model with unknown marginal baseline hazard. We present simulation results for two classes, depending on whether the assumed Clayton copula model is correctly specified or misspecified.

4.1. Correctly specified copula models

We first consider the case where the assumed Clayton copula model is correctly specified. In order to evaluate the performance of the proposed one-stage estimation method, the simulation schemes are as follows. It was performed on a copula model with an unknown marginal baseline hazard using 500 replications of simulation data. Event times are simulated from a Clayton copula survival model (2.1.8) with association parameter at $\theta = 0.5$, 2 and 8 which give corresponding Kendall's tau $\tau = \theta/(\theta+2) = 0.2$, 0.5 and 0.8, respectively. Here, we consider the marginal PH model with the Gompertz distribution as the true baseline hazard:

$$\lambda(t|x) = \exp(\phi t + \beta_0 + \beta x), \tag{4.1.1}$$

where we set the shape parameter $\phi = 0.2$, 1 and 3; its hazard exponentially increases with ϕ. Data are generated using the sampling algorithm of Marshall and Olkin (1988) as follows (Prenen et al., 2017a; Ha et al., 2019). For $i = 1, \cdots, q$ and $j = 1, \cdots, n_i$, define

$$A_{ij} = \psi(-\log U_{ij}/Z_i), \tag{4.1.2}$$

where $\psi(s) = (1+\theta s)^{-1/\theta}$ for $\theta > 0$, U_{ij}'s follow independent and identically $U(0,1)$ and Z_i's follow independent and identically gamma distribution with mean 1 and variance θ. Then survival times T_{ij}'s are generated from

$$T_{ij} = (1/\phi)\log\{1-\phi\log A_{ij}/\exp(\beta_0+\beta x_{ij})\}. \tag{4.1.3}$$

Here, we set a log-scale parameter $\beta_0 = 0$ and a regression parameter $\beta = 1$, and a single covariate x_{ij} is generated from $N(0, 1)$. We consider the three cases of multivariate cluster sizes:

Case A. The same cluster size: $n_i = 2$; $(q, n_i) = (50, 2)$, $(200, 2)$ for all i.
Case B. The same cluster size: $n_i = 4$; $(50, 4)$, $(100, 4)$, $(200, 4)$ for all i.
Case C. The different cluster size, based on multicenter bladder cancer data in Section 5.3.

The corresponding censoring times C_{ij}'s are generated from an exponential distribution having a parameter empirically determined to achieve approximately about 20% censoring rates.

In this chapter, we first fit simulation model above (i.e. Clayton model with Gompertz marginal hazard) using the one-stage Clayton copula method based on a cubic M-spline. We investigate the behaviors for the estimates of parameters of interest (β, θ). For 500 replications of simulated data, we calculate the mean, standard

deviation (SD), the mean of the estimated standard errors (SE), and mean squared error (MSE) for each $(\hat{\beta}, \hat{\theta})$. We also compute the empirical coverage probability (CP) for a nominal 95% confidence interval (CI) for β and θ, respectively. In addition, we compare the performance of the proposed method with that of the five existing methods: one-stage Weibull and partial exponential (PE), two-stage Weibull, PE, and Cox of Prenen et al. (2017a). Here, the existing five methods are implemented with the Sunclarco R package by Prenen et al. (2017b); these two-stage procedures use a sandwich variance estimator for $\hat{\beta}$ that takes the dependence into account among survival times. Now we present the simulation results of three cases according to the cluster size.

Case A. The same cluster size: $n_i = 2$

The simulation results for Case A are summarized in Tables 4.1.1 and 4.1.2.

(i) Table 4.1.1 with a small sample $n = (50, 2)$:

Overall the proposed method works well in terms of biases of $\hat{\beta}$ and $\hat{\theta}$. The estimated SEs of $\hat{\beta}$ and $\hat{\theta}$ are also very close to the corresponding empirical SDs, which are the estimates of the true $\{\mathrm{var}(\hat{\beta})\}^{1/2}$ and $\{\mathrm{var}(\hat{\theta})\}^{1/2}$, respectively. As a result, the CPs of the 95% CI match well with the nominal value of 0.95. On the other hand, all the two-stage estimators show some underestimation for θ, leading to substantially lower CPs. The one-stage Weibull method is very sensitive against misspecification of the marginal hazard distribution, leading that it gives seriously lower performances in terms of bias, MSE, and CP for $(\hat{\beta}, \hat{\theta})$, particularly when ϕ or θ

increases. The results of the one-stage PE method are compared with those of the proposed method under a small $\theta = 0.5$ (i.e. Kendall's tau $\tau = 0.2$). However, when θ is large as $\theta = 8$ (i.e. $\tau = 0.8$), the one-stage PE gives larger variations for $\hat{\beta}$ (i.e. SD and MSE), leading to very lower CPs. The trends of very lower performances from the two-stage Weibull method are similar to the one-stage Weibull method. The two-stage PE method gives fewer biases for $\hat{\beta}$, but its SE is underestimated as compared to the SD, leading to very lower CPs for θ. Moreover, $\hat{\theta}$ is seriously downward biased, giving very lower CPs for θ. As expected from the GEE-based marginal Cox's modeling approach (Spiekerman and Lin, 1998), the results of the two-stage Cox method are comparable to those of the proposed method in terms of the bias of $\hat{\beta}$ and CP for β, but the two-stage Cox method gives larger variations (i.e. SE, SD, and MSE) for $\hat{\beta}$, especially when θ increases. In the two-stage Cox method, $\hat{\theta}$ is again seriously biased downward, leading to very lower CPs.

(ii) Table 4.1.2 with sample $n = (200, 2)$:

As the sample size increases from $n = (50, 2)$ to $n = (200, 2)$, we observe that our estimators for (β, θ) are consistent and the SE estimates perform well as judged by the very good agreement between SE and SD. In particular, the proposed method's CPs for (β, θ) are in the 93%~96% range in almost all cases.

As shown in the box plots in Figures 4.1.1 and 4.1.2, it is observed that the performance of the proposed method performs well. In the case of $\theta = 2$, the simulation results for $n = 100$ and 400 are presented in Tables D1 and D2 in the Appendix D which are similar to the previous results.

Table 4.1.1. (q, n_i) = (50, 2): Simulation results on one-stage and two-stage estimation methods over 500 replications under Clayton copula models with Gompertz marginal hazard; 20% censoring rate; $\beta = 1$

θ	ϕ	Est	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
Baseline hazard function							One-stage							
			Weibull				PE				Proposed			
0.5	0.2	$\hat{\beta}$	0.986	0.139 0.131	0.017	0.968	1.040	0.146 0.148	0.023	0.946	1.032	0.146 0.151	0.024	0.934
		$\hat{\theta}$	0.532	0.319 0.307	0.095	0.954	0.545	0.320 0.370	0.139	0.923	0.544	0.324 0.358	0.130	0.904
	3	$\hat{\beta}$	0.863	0.135 0.136	0.037	0.792	1.037	0.147 0.152	0.024	0.952	1.038	0.148 0.141	0.021	0.956
		$\hat{\theta}$	0.587	0.347 0.352	0.131	0.946	0.557	0.328 0.378	0.145	0.925	0.541	0.327 0.327	0.108	0.950
8	0.2	$\hat{\beta}$	0.973	0.098 0.116	0.014	0.868	1.075	0.114 0.134	0.024	0.886	1.060	0.104 0.123	0.019	0.894
		$\hat{\theta}$	7.550	1.794 1.801	3.441	0.916	9.885	2.445 3.249	14.087	0.936	8.595	2.032 2.188	5.132	0.958
	3	$\hat{\beta}$	0.803	0.087 0.116	0.052	0.402	1.068	0.112 0.140	0.024	0.794	1.027	0.045 0.085	0.008	0.928
		$\hat{\theta}$	5.263	1.299 1.288	9.145	0.388	9.967	2.470 3.240	14.348	0.916	8.723	2.056 2.262	5.631	0.954
Baseline hazard function							Two-stage							
			Weibull				PE				Cox			
0.5	0.2	$\hat{\beta}$	0.990	0.139 0.133	0.018	0.962	1.037	0.126 0.152	0.024	0.866	1.023	0.142 0.144	0.021	0.952
		$\hat{\theta}$	0.454	0.321 0.201	0.042	0.940	0.433	0.153 0.215	0.050	0.642	0.449	0.197 0.212	0.047	0.680
	3	$\hat{\beta}$	0.883	0.126 0.132	0.025	0.660	1.034	0.127 0.157	0.026	0.882	1.025	0.144 0.147	0.022	0.940
		$\hat{\theta}$	0.473	0.404 0.213	0.031	0.800	0.435	0.157 0.214	0.050	0.666	0.443	0.196 0.214	0.049	0.676
8	0.2	$\hat{\beta}$	1.000	0.151 0.165	0.027	0.938	1.049	0.140 0.179	0.034	0.862	1.018	0.158 0.168	0.028	0.936
		$\hat{\theta}$	6.135	1.862 1.796	6.698	0.674	5.285	1.664 1.496	9.604	0.518	5.151	2.030 1.539	10.482	0.610
	3	$\hat{\beta}$	0.883	0.141 0.153	0.037	0.808	1.049	0.141 0.182	0.035	0.860	1.018	0.159 0.169	0.029	0.934
		$\hat{\theta}$	4.466	1.225 1.256	14.065	0.234	4.862	1.583 1.393	11.783	0.440	4.732	1.950 1.440	12.753	0.524

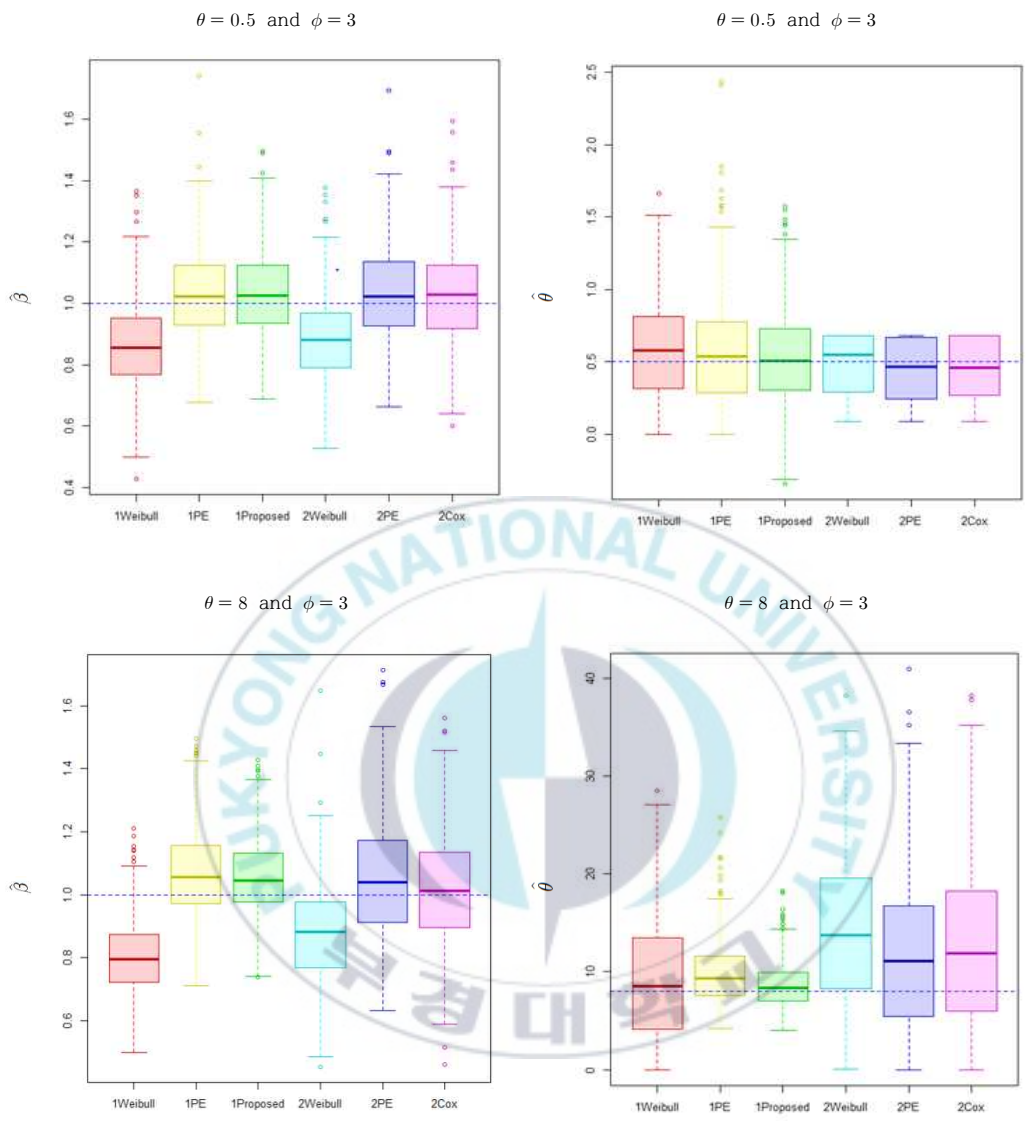


Figure 4.1.1. $(q, n_i) = (50, 2)$: Simulation result of Copula M-spline over 500 replications; 20% censoring rate; dotted line, true values of β and θ, respectively

Table 4.1.2. (q, n_i) = (200, 2): Simulation results on one-stage and two-stage estimation methods over 500 replications under Clayton copula models with Gompertz marginal hazard; 20% censoring rate; $\beta = 1$

θ	ϕ	Est	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
Baseline hazard function							One-stage							
			Weibull				PE				Proposed			
0.5	0.2	$\hat{\beta}$	0.967	0.067 0.066	0.005	0.922	1.002	0.070 0.071	0.005	0.942	1.007	0.070 0.069	0.005	0.952
		$\hat{\theta}$	0.521	0.155 0.153	0.024	0.964	0.521	0.155 0.157	0.025	0.952	0.512	0.153 0.156	0.024	0.946
	3	$\hat{\beta}$	0.838	0.066 0.062	0.030	0.308	0.996	0.070 0.069	0.005	0.946	1.006	0.071 0.069	0.005	0.952
		$\hat{\theta}$	0.573	0.173 0.174	0.035	0.960	0.513	0.156 0.161	0.026	0.950	0.511	0.155 0.160	0.026	0.950
8	0.2	$\hat{\beta}$	0.944	0.047 0.054	0.006	0.716	1.002	0.051 0.056	0.003	0.922	1.011	0.047 0.051	0.003	0.932
		$\hat{\theta}$	7.170	0.845 0.800	1.328	0.800	8.221	0.965 0.994	1.035	0.942	8.147	0.943 0.943	0.909	0.966
	3	$\hat{\beta}$	0.773	0.042 0.053	0.055	0.010	0.974	0.049 0.062	0.005	0.842	1.012	0.050 0.052	0.003	0.938
		$\hat{\theta}$	4.902	0.605 0.573	9.924	0.010	8.049	0.941 0.961	0.930	0.946	8.136	0.946 0.955	0.929	0.946
Baseline hazard function							Two-stage							
			Weibull				PE				Cox			
0.5	0.2	$\hat{\beta}$	0.970	0.069 0.069	0.006	0.926	1.002	0.071 0.073	0.005	0.934	1.000	0.073 0.073	0.005	0.938
		$\hat{\theta}$	0.501	0.155 0.131	0.017	0.972	0.488	0.125 0.132	0.018	0.828	0.489	0.132 0.133	0.018	0.834
	3	$\hat{\beta}$	0.883	0.126 0.132	0.025	0.660	1.034	0.127 0.157	0.026	0.882	1.025	0.144 0.147	0.022	0.940
		$\hat{\theta}$	0.473	0.404 0.213	0.031	0.800	0.435	0.157 0.214	0.050	0.666	0.443	0.196 0.214	0.049	0.676
8	0.2	$\hat{\beta}$	0.977	0.075 0.078	0.007	0.920	1.008	0.078 0.083	0.007	0.928	1.007	0.080 0.083	0.007	0.942
		$\hat{\theta}$	6.655	0.968 0.833	2.501	0.690	6.667	1.112 0.917	2.616	0.752	6.493	1.152 0.886	3.053	0.740
	3	$\hat{\beta}$	0.862	0.070 0.070	0.024	0.448	1.004	0.077 0.083	0.007	0.936	1.002	0.080 0.080	0.006	0.944
		$\hat{\theta}$	4.541	0.612 0.620	12.347	0.004	6.261	1.106 0.939	3.903	0.616	6.319	1.189 0.894	3.623	0.686

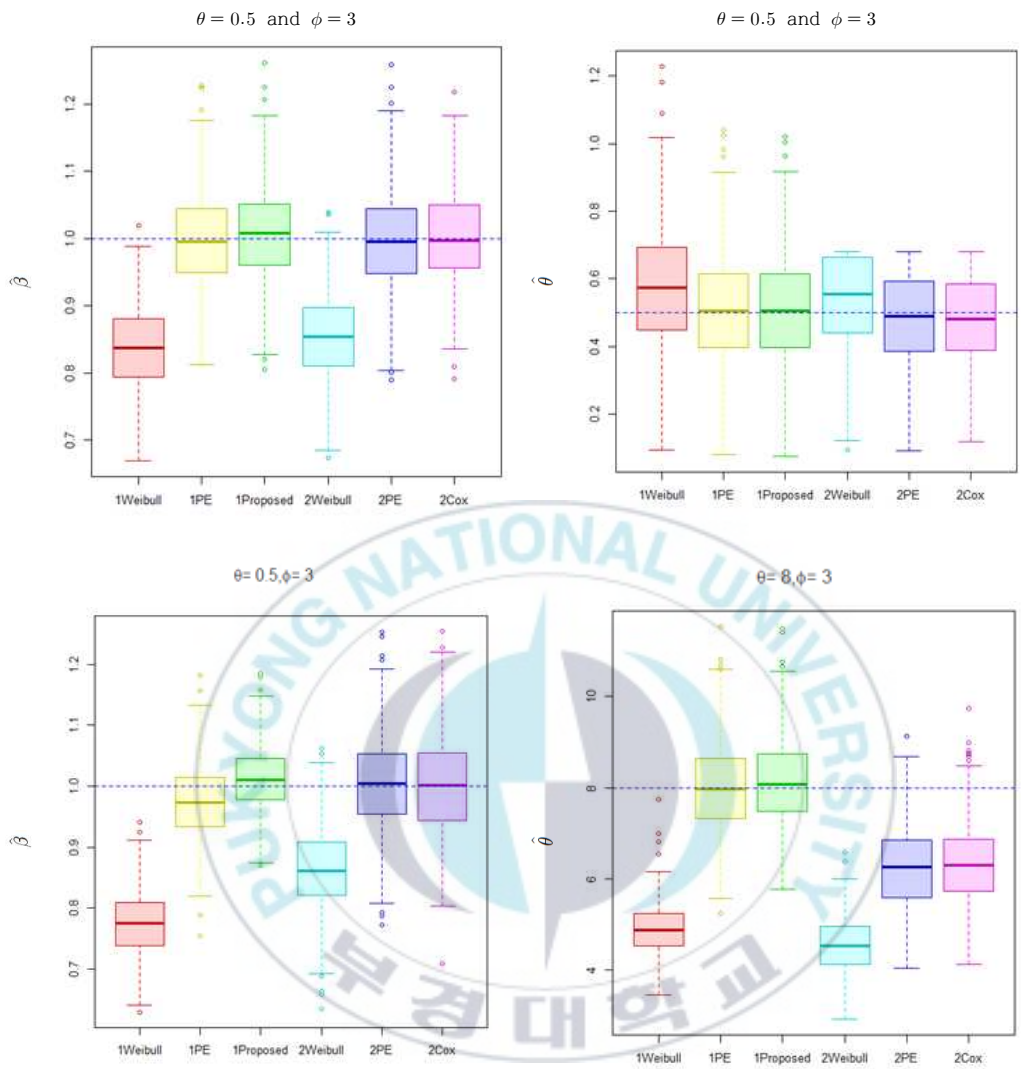


Figure 4.1.2. $(q, n_i)=(200,2)$: Simulation result of Copula M-spline over 500 replications; 20% censoring rate; dotted line, true values of β and θ, respectively

Case B. The same cluster size: $n_i = 4$

(i) Table 4.1.3 with a small sample $n=(50,4)$:

As expected, the proposed method works well overall. As the

association parameter θ and shape parameter ϕ increase in the one-stage and two-stage Weibull methods, the estimated marginal hazard parameters are sensitive to the specification of Gompertz marginal hazard.

• The result of the one-stage PE method gives lower performance for $\hat{\beta}$ as θ increases, and the two-stage PE method is severely biased downward toward $\hat{\theta}$, giving very low CPs. The two-stage Cox method, as θ and ϕ increase, results in very low CPs for $\hat{\theta}$.

(ii) Tables 4.1.4 with sample $n=(200,4)$:

As the sample size increases, the estimates for $(\hat{\beta},\ \hat{\theta})$ are consistent, and in particular, it is observed that the proposed method performs well with a good agreement between SE and SD. When the sample size is (200, 4), the CP is in the 95% range in almost all cases.

As shown in the box plot of Figure 4.1.3-4.1.4, it is observed that the performance of the proposed method performs better than the other methods. When $\theta=2$, the simulation results for $n=200$ and 800 are presented in Tables D3 and D4 of the Appendix D, and are similar to the previous results.

Summarizing the same cluster size, as compared to the proposed method, the performances of the existing five methods are generally lower, especially when θ increases. In particular, the existing one-stage methods (Weibull and PE) are sensitive to the estimation of β. However, the conventional two-stage methods (Weibull, PE, and Cox) are sensitive to the estimation of both β and θ. However, the proposed method performs well, and it is also more efficient in terms of MSE, especially when the intensity of dependence increases.

Table 4.1.3. (q, n_i) = (50, 4): Simulation results on one-stage and two-stage estimation methods over 500 replications under Clayton copula models with Gompertz marginal hazard; 20% censoring rate; $\beta = 1$

θ	ϕ	Est	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
Baseline hazard function				Weibull				One-stage PE				Proposed		
0.5	0.2	$\hat{\beta}$	0.977	0.095 0.098	0.010	0.930	1.023	0.100 0.104	0.011	0.942	1.026	0.100 0.104	0.011	0.938
		$\hat{\theta}$	0.519	0.179 0.199	0.040	0.916	0.508	0.177 0.194	0.038	0.916	0.503	0.175 0.193	0.037	0.908
	3	$\hat{\beta}$	0.829	0.093 0.095	0.038	0.518	1.018	0.101 0.105	0.011	0.932	1.025	0.101 0.105	0.012	0.934
		$\hat{\theta}$	0.610	0.214 0.240	0.069	0.938	0.509	0.179 0.195	0.038	0.908	0.504	0.177 0.193	0.037	0.902
8	0.2	$\hat{\beta}$	0.930	0.074 0.087	0.013	0.799	1.004	0.080 0.097	0.009	0.904	1.031	0.068 0.081	0.007	0.918
		$\hat{\theta}$	7.227	1.281 1.157	1.934	0.884	8.433	1.480 1.541	2.559	0.952	8.138	1.408 1.432	2.067	0.956
	3	$\hat{\beta}$	0.753	0.059 0.083	0.068	0.090	0.963	0.076 0.111	0.014	0.794	1.027	0.045 0.085	0.008	0.928
		$\hat{\theta}$	5.050	0.922 0.751	9.264	0.218	8.358	1.474 1.498	2.369	0.952	8.152	1.419 1.451	2.124	0.950
Baseline hazard function				Weibull				Two-stage PE				Cox		
0.5	0.2	$\hat{\beta}$	0.991	0.098 0.105	0.011	0.920	1.031	0.094 0.120	0.015	0.866	1.019	0.104 0.111	0.013	0.920
		$\hat{\theta}$	0.480	0.175 0.156	0.025	0.920	0.450	0.120 0.170	0.031	0.710	0.465	0.146 0.153	0.025	0.816
	3	$\hat{\beta}$	0.874	0.091 0.096	0.025	0.660	1.023	0.087 0.112	0.013	0.864	1.020	0.105 0.113	0.013	0.924
		$\hat{\theta}$	0.515	0.199 0.156	0.025	0.960	0.456	0.112 0.152	0.025	0.740	0.463	0.147 0.151	0.024	0.822
8	0.2	$\hat{\beta}$	0.989	0.122 0.134	0.018	0.920	1.029	0.117 0.150	0.023	0.870	1.020	0.132 0.144	0.021	0.936
		$\hat{\theta}$	6.317	1.339 1.279	4.467	0.684	6.057	1.327 1.367	5.639	0.598	5.282	1.446 1.203	8.832	0.496
	3	$\hat{\beta}$	0.874	0.115 0.123	0.031	0.736	1.025	0.112 0.146	0.022	0.874	1.022	0.132 0.146	0.022	0.940
		$\hat{\theta}$	4.424	0.801 0.835	13.485	0.048	5.136	1.512 1.189	9.614	0.354	4.911	1.403 1.091	10.732	0.406

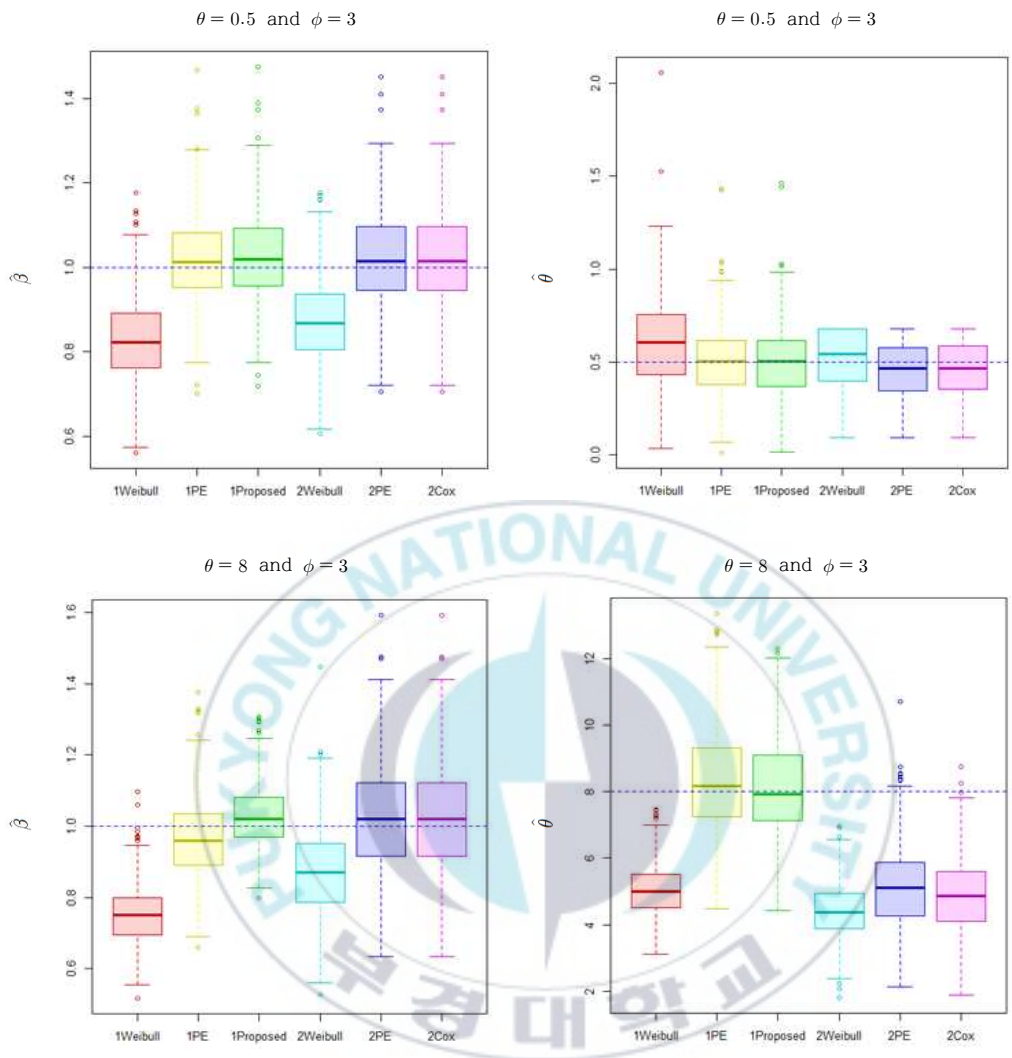


Figure 4.1.3. $(q, n_i) = (50, 4)$: Simulation result of copula M-spline over 500 replications; 20% censoring rate; dotted line, true values of β and θ, respectively

Table 4.1.4. (q, n_i) = (200, 4): Simulation results on one-stage and two-stage estimation methods over 500 replications under Clayton copula models with Gompertz marginal hazard; 20% censoring rate; $\beta = 1$

θ	ϕ	Est	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
Baseline hazard function								One-stage						
			Weibull				PE				Proposed			
0.5	0.2	$\hat{\beta}$	0.960	0.046 0.047	0.004	0.838	1.000	0.049 0.050	0.003	0.950	1.005	0.049 0.050	0.003	0.942
		$\hat{\theta}$	0.521	0.089 0.088	0.008	0.956	0.503	0.087 0.086	0.007	0.960	0.503	0.087 0.083	0.007	0.954
	3	$\hat{\beta}$	0.810	0.046 0.045	0.038	0.028	0.991	0.049 0.049	0.003	0.950	1.004	0.049 0.049	0.002	0.946
		$\hat{\theta}$	0.617	0.107 0.108	0.025	0.844	0.506	0.088 0.087	0.008	0.946	0.504	0.087 0.084	0.007	0.962
8	0.2	$\hat{\beta}$	0.913	0.036 0.042	0.009	0.368	0.971	0.038 0.043	0.003	0.824	1.005	0.030 0.031	0.001	0.944
		$\hat{\theta}$	7.172	0.633 0.582	1.024	0.704	8.091	0.705 0.705	0.512	0.950	8.019	0.689 0.714	0.509	0.948
	3	$\hat{\beta}$	0.733	0.029 0.039	0.073	0	0.929	0.036 0.050	0.008	0.794	1.007	0.038 0.037	0.001	0.962
		$\hat{\theta}$	4.976	0.454 0.356	9.269	0	8.008	0.703 0.691	0.477	0.930	8.009	0.695 0.715	0.510	0.948
Baseline hazard function								Two-stage						
			Weibull				PE				Cox			
0.5	0.2	$\hat{\beta}$	0.971	0.050 0.052	0.004	0.902	1.001	0.051 0.056	0.003	0.922	1.003	0.053 0.056	0.003	0.942
		$\hat{\theta}$	0.511	0.087 0.085	0.007	0.966	0.494	0.081 0.084	0.007	0.928	0.494	0.085 0.084	0.007	0.930
	3	$\hat{\beta}$	0.854	0.046 0.046	0.023	0.132	0.995	0.051 0.056	0.003	0.930	1.003	0.054 0.056	0.003	0.938
		$\hat{\theta}$	0.563	0.098 0.086	0.011	0.992	0.493	0.081 0.084	0.007	0.928	0.492	0.086 0.085	0.007	0.942
8	0.2	$\hat{\beta}$	0.967	0.063 0.065	0.005	0.920	1.000	0.065 0.065	0.004	0.944	1.002	0.067 0.066	0.004	0.956
		$\hat{\theta}$	6.799	0.702 0.615	1.818	0.600	7.026	0.851 0.749	1.507	0.776	6.820	0.874 0.706	1.889	0.734
	3	$\hat{\beta}$	0.855	0.058 0.057	0.024	0.284	0.994	0.065 0.065	0.004	0.942	1.002	0.067 0.066	0.004	0.956
		$\hat{\theta}$	4.525	0.426 0.410	12.247	0	6.625	0.828 0.717	2.403	0.594	6.528	0.901 0.698	2.654	0.602

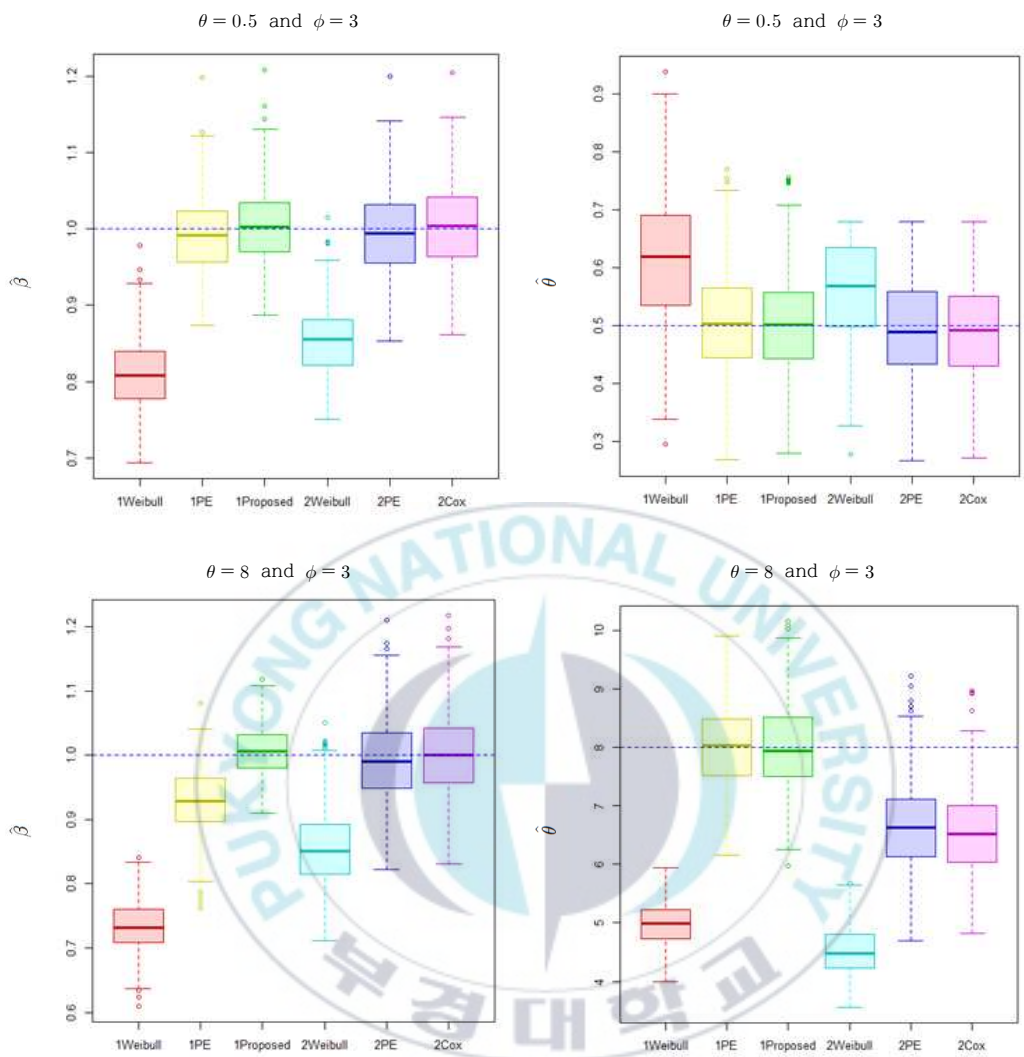


Figure. 4.1.4 $(q, n_i) = (200, 4)$: Simulation result of copula M-spline over 500 replications; 20% censoring rate; dotted line, true values of β and θ, respectively

Case C. The different cluster size

For the data generation of the different multivariate cluster size, we use the multicenter bladder cancer data structure (1,066 patients with $q = 46$ centers; 55.7% censoring) with different cluster (center) sizes.

For this purpose, the data are generated from the same simulated model above, i.e. the Clayton model with the Gompertz hazard

$$\lambda(t|x)=\exp\left(\phi t+\beta_0+\beta^T x\right),$$

where $\phi=1$, $\beta_0=0$ and $x=(x_1,\cdots,x_{10})^T$ is equal to the 10 covariates in Table 5.3.1 for all 500 replications. Here, we used the proposed estimates in Section 5.3 as the true parameters, i.e. the true regression parameters are

$$\begin{aligned}\beta&=(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7,\beta_8,\beta_9,\beta_{10})^T\\&=(-0.146,-0.203,0.002,0.162,0.382,0.002,0.135,0.056,-0.344,-0.233)^T,\end{aligned}$$

and the true association parameter is $\theta=0.083$. We also consider $\theta=2$ which gives a larger association. The remaining simulation schemes are the same as before.

For the simplicity of comparison, the simulated data are fitted using the four Clayton copula models (one-stage Weibull and proposed one-stage M-spline, and two-stage Weibull and Cox). The simulation results are summarized in Table 4.1.5, 4.1.6 and Figure 4.1.5. The trends of the estimation results are overall similar to those evident in the previous tables and figures in this Section. That is, the proposed method still performs efficiently, except for giving a low CP of θ under a small $\theta=0.083$. As expected, the one-stage and two-stage Weibull methods lead to lower performances for $(\hat{\beta},\hat{\theta})$, particularly for a larger association as in $\theta=2$. We again find that the estimates of β of the proposed method are similar to those of the two-stage Cox

method. However, for $(\hat{\beta}, \hat{\theta})$ the Cox method gives larger variations (i.e. SE, SD and MSE) as well as lower CPs under $\theta = 2$ as shown in the previous tables and figures in this Section.

Table 4.1.5. Different cluster size: Simulation results on one-stage and two-stage estimation methods with different cluster size over 500 replications under Clayton copula models with Gompertz marginal hazard having the multicenter bladder cancer data structures; $\theta = 0.083$

Baseline hazard function		One-stage								Two-stage							
		Weibull				Proposed				Weibull				Cox			
Param-eter	True	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
β_1	-0.14 6	-0.136	0.077 0.076	0.006	0.942	-0.149	0.080 0.079	0.006	0.952	-0.138	0.074 0.074	0.006	0.946	-0.150	0.078 0.081	0.007	0.934
β_2	-0.20 3	-0.182	0.078 0.076	0.006	0.954	-0.200	0.080 0.081	0.007	0.954	-0.195	0.073 0.074	0.006	0.942	-0.214	0.078 0.082	0.007	0.926
β_3	-0.00 2	-0.002	0.004 0.004	0.000	0.952	-0.002	0.004 0.004	0.000	0.952	-0.002	0.004 0.004	0.006	0.924	-0.002	0.004 0.004	0.000	0.932
β_4	0.162	0.162	0.099 0.094	0.009	0.944	0.164	0.102 0.104	0.011	0.944	0.157	0.095 0.094	0.006	0.944	0.167	0.100 0.103	0.011	0.936
β_5	0.382	0.344	0.090 0.085	0.009	0.960	0.390	0.092 0.088	0.008	0.960	0.362	0.087 0.091	0.006	0.912	0.399	0.093 0.094	0.010	0.924
β_6	0.002	0.002	0.003 0.003	0.000	0.938	0.002	0.003 0.003	0.000	0.938	0.002	0.003 0.003	0.006	0.894	0.002	0.003 0.004	0.000	0.906
β_7	0.135	0.126	0.024 0.024	0.001	0.946	0.137	0.024 0.025	0.001	0.946	0.131	0.023 0.025	0.006	0.930	0.139	0.024 0.026	0.001	0.918
β_8	0.056	0.056	0.108 0.104	0.011	0.936	0.057	0.111 0.118	0.014	0.936	0.053	0.106 0.116	0.006	0.912	0.060	0.113 0.118	0.015	0.942
β_9	-0.34 4	-0.314	0.140 0.130	0.018	0.936	-0.350	0.143 0.143	0.020	0.936	-0.328	0.137 0.147	0.006	0.922	-0.356	0.147 0.161	0.026	0.914
β_{10}	-0.23 3	-0.207	0.119 0.113	0.013	0.940	-0.233	0.122 0.125	0.016	0.940	-0.215	0.116 0.124	0.006	0.918	-0.231	0.123 0.136	0.018	0.916
θ	0.083	0.148	0.081 0.099	0.014	0.914	0.080	0.037 0.039	0.001	0.904	0.090	0.019 0.064	0.026	0.884	0.070	0.031 0.028	0.001	0.796

Table 4.1.6. Different cluster size: Simulation results on one-stage and two-stage estimation methods with different cluster size over 500 replications under Clayton copula models with Gompertz marginal hazard having the multicenter bladder cancer data structures; $\theta = 2$

Baseline hazard function		One-stage								Two-stage							
		Weibull				Proposed				Weibull				Cox			
Param-eter	True	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
β_1	-0.146	-0.118	0.035 0.040	0.002	0.824	-0.152	0.037 0.037	0.001	0.956	-0.144	0.061 0.060	0.004	0.914	-0.161	0.071 0.066	0.005	0.906
β_2	-0.203	-0.180	0.036 0.041	0.002	0.868	-0.215	0.040 0.039	0.002	0.936	-0.198	0.068 0.065	0.005	0.942	-0.220	0.072 0.072	0.005	0.944
β_3	-0.002	-0.002	0.002 0.003	0.000	0.678	-0.002	0.002 0.002	0.000	0.954	-0.002	0.004 0.004	0.000	0.932	-0.002	0.005 0.004	0.000	0.914
β_4	0.162	0.134	0.044 0.051	0.009	0.854	0.171	0.047 0.046	0.002	0.964	0.156	0.095 0.094	0.009	0.942	0.176	0.102 0.099	0.011	0.924
β_5	0.382	0.317	0.044 0.041	0.009	0.630	0.403	0.055 0.051	0.003	0.938	0.360	0.125 0.119	0.016	0.916	0.416	0.151 0.131	0.024	0.876
β_6	0.002	0.002	0.001 0.002	0.000	0.908	0.002	0.002 0.001	0.000	0.936	0.002	0.007 0.005	0.000	0.826	0.001	0.007 0.006	0.000	0.854
β_7	0.135	0.118	0.012 0.015	0.001	0.630	0.139	0.015 0.015	0.000	0.942	0.129	0.041 0.035	0.002	0.896	0.145	0.042 0.036	0.002	0.870
β_8	0.056	0.036	0.050 0.072	0.006	0.846	0.064	0.051 0.051	0.003	0.956	0.060	0.171 0.153	0.029	0.922	0.063	0.196 0.164	0.038	0.904
β_9	-0.344	-0.299	0.066 0.089	0.010	0.812	-0.358	0.071 0.071	0.005	0.954	-0.339	0.249 0.224	0.062	0.920	-0.371	0.281 0.240	0.079	0.908
β_{10}	-0.233	-0.209	0.055 0.076	0.006	0.838	-0.241	0.059 0.058	0.003	0.952	-0.231	0.196 0.170	0.038	0.916	-0.254	0.216 0.178	0.047	0.896
θ	0.083	2.105	0.373 0.505	0.266	0.872	2.017	0.383 0.395	0.156	0.940	1.635	0.632 0.343	0.251	0.354	1.362	0.376 0.323	0.511	0.512

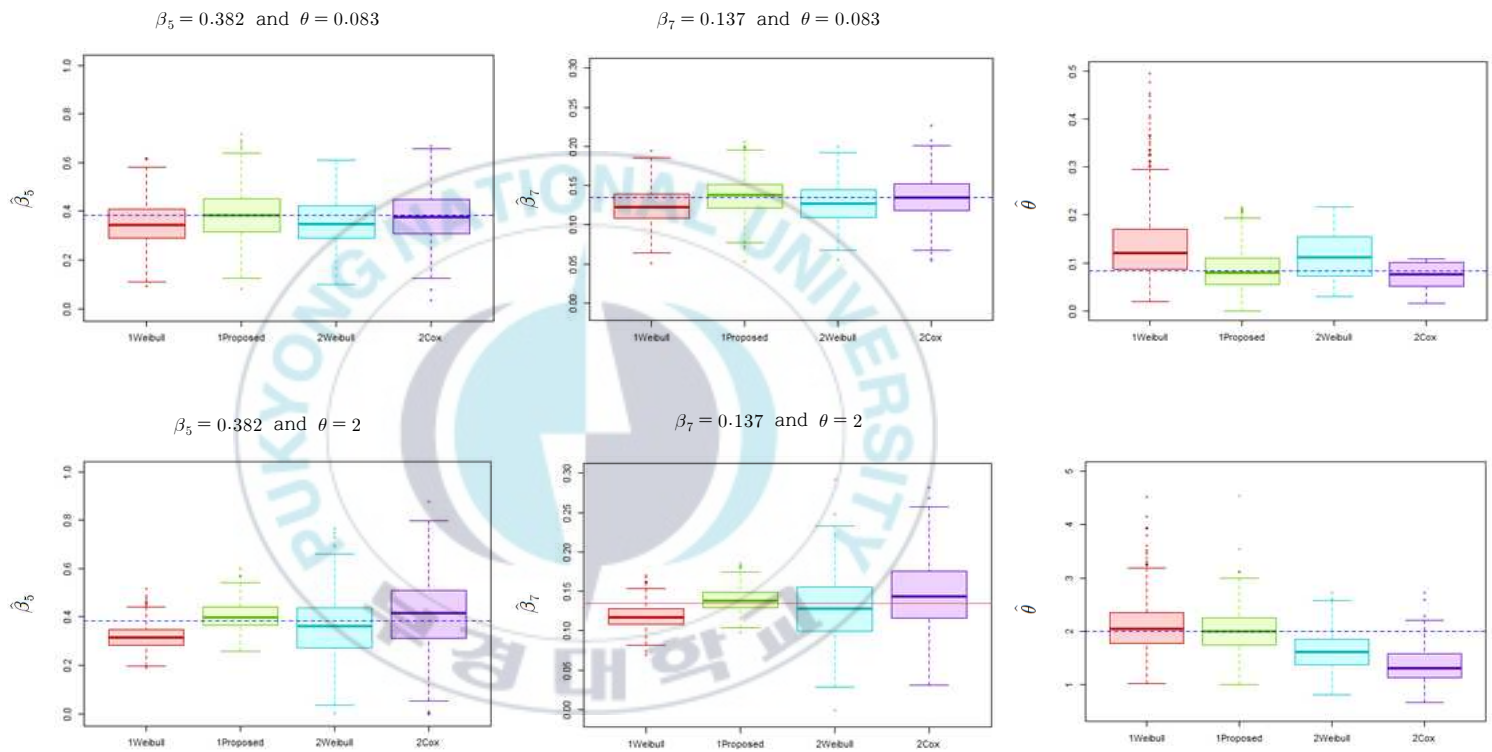


Figure 4.1.5. Comparison of simulation results on one-stage and two-stage estimation methods with different cluster size over 500 replications; dotted line, true values of β_5, β_7, and , respectively

4.2. Misspecified copula models

With a small numerical study, we now investigate the performance of the proposed method when the assumed Clayton copula model is misspecified. For this, we consider the Gumbel-Hougaard (GH) copula as the true copula function. Thus, the event times are simulated from a GH copula survival model (2.1.3) with association parameter $\theta = 0.6$ and 0.3 which give corresponding Kendall's tau (i.e. $\tau = 1-\theta$) 0.4 and 0.7, and Weibull distribution as the true marginal hazard function

$$\lambda(t|x) = \exp(\phi \log t + \beta_0 + \beta x),$$

where we set a shape parameter $\phi = 1.5$ (i.e. an increasing hazard), a log-scale parameter $\beta_0 = 0$ and a regression parameter $\beta = 0.5$. Similarly to (4.1.3), the survival times T_{ij}'s are generated from

$$T_{ij} = \{\log M_{ij} / \exp(\beta_0 + \beta x)\}^{1/\phi},$$

where $M_{ij} = \psi(-\log U_{ij}/Z_i)$ with $\psi(s) = \exp(-s^{\theta})$ for $0 < \theta < 1$, and $U_{ij} \sim iid\, U(0,1)$ and $Z_i \sim iid$ positive stable distribution with shape parameter (i.e. association parameter) θ. The sample size is set as $n = (50, 4)$ and $n = (200, 4)$, and censoring rate is 20%. The remaining schemes are the same as that of Section 4.1.

The simulated data are fitted using the two Clayton copula methods, i.e. the proposed one-stage M-spline method and the two-stage Cox method. For the presentation of the degree of association, we here report the estimation results of Kendall's tau (i.e. $\tau = \theta/(\theta+2)$), rather than θ itself, by the two fitted Clayton copula methods. Here, we

investigate the behaviors for the estimates of parameters of interest (β, θ). The simulation results are summarized in Table 4.2.1. Table 4.2.1 shows that for estimation of β, the proposed method can lead to a wrong estimate if the assumed Clayton copula model is misspecified, whereas the two-stage Cox method seems to give consistent and robust estimates. However, in the proposed method $\hat{\tau}$ is less biased with smaller MSEs, whereas in the Cox method it is seriously biased downward. Care is necessary for the inference of β by the proposed method when a copula function, not a marginal hazard distribution, is misspecified.

Table 4.2.1. Simulation results on 500 replications of fitting the proposed one-stage M-spline and two-stage Cox methods under Gumbel-Hougaard (GH) copula models with Weibull marginal hazard; shape parameter $\phi = 1.5$; 20% censoring rate; $\beta = 0.5$

Baseline hazard function			Proposed					Two-stage Cox				
(q, n_i)	τ	EST	Mean	SE	SD	MSE	CP	Mean	SE	SD	MSE	CP
(50,4)	0.4	$\hat{\beta}$	0.420	0.066	0.087	0.014	0.664	0.509	0.089	0.094	0.009	0.940
		$\hat{\tau}$	0.379	0.063	0.081	0.007	0.852	0.299	0.054	0.050	0.013	0.530
	0.7	$\hat{\beta}$	0.295	0.038	0.080	0.048	0.132	0.508	0.093	0.102	0.010	0.912
		$\hat{\tau}$	0.728	0.041	0.072	0.006	0.536	0.541	0.060	0.056	0.029	0.258
(200,4)	0.4	$\hat{\beta}$	0.410	0.032	0.040	0.010	0.250	0.499	0.045	0.049	0.002	0.926
		$\hat{\tau}$	0.375	0.031	0.035	0.002	0.834	0.298	0.027	0.027	0.011	0.048
	0.7	$\hat{\beta}$	0.272	0.019	0.058	0.055	0.002	0.503	0.047	0.049	0.002	0.934
		$\hat{\tau}$	0.742	0.021	0.054	0.005	0.162	0.560	0.029	0.029	0.021	0

V. ILLUSTRATION FOR COPULA SURVIVAL MODELS

For the illustration of the proposed method in Section 4.1, we consider three data sets of correlated survival data. The first data set is on the bivariate kidney infection survival times (McGilchrist and Aisbett, 1991). The second one is on the CGD (chronic granulomatous disease) recurrent infection survival times with different cluster sizes. The third one is data from a multicenter bladder cancer trial (Oddens et al. 2013; Park and Ha, 2019). We fit the Clayton copula survival models with unknown marginal baseline hazard using the proposed one-stage procedure using the M-spline method.

5.1. Kidney infection data

The event times from the same patient can be correlated due to a shared patient effect. We consider two covariates in the kidney data: Age and Sex(1=F(female), 0=M(male)). The fitted results (i.e. the estimated regression coefficients and their SEs) of the copula models via the proposed method are summarized in Table 5.1.1.

The estimates of association parameter θ in the six methods are all similar. Note that as mentioned in Section 4.1, the two-stage Cox estimates and SEs for regression parameters (i.e. Age and Sex effects) are the same as the marginal Cox estimates based on the GEE approach (Spiekerman and Lin, 1998). Following the Wald test statistic, the Age effect is not significant at the 5% significance level for all six methods, but the Sex effect gives different significance.

The Sex effect is significant according to the one-stage method, but it is not in the two-stage method due to larger SE of Sex effect which is also confirmed from the simulation results in Tables 4.1.1-4.1.2. It is well-known that the Sex effect is significant according to the results of many literatures (Hougaard, 2000; Ha et al., 2017; McGilchrist and Aisbett, 1991). Care is necessary in conducting the inference using the two-stage method in clustered survival data.

Table 5.1.1. Kidney infection data: estimation results of Clayton copula models using the proposed and existing five methods

Baseline hazard function	One-stage			Two-stage		
	Weibull	PE	Proposed	Weibull	PE	Cox
Parameter	Est (SE)	Est (SE)	Est (SE)	Est (SE)	Est (SE)	Est (SE)
Age	0.003 (0.010)	0.001 (0.010)	0.002 (0.010)	0.004 (0.009)	0.002 (0.006)	0.002 (0.008)
Sex :Female	**-0.937 (0.301)**	**-0.924 (0.310)**	**-0.890 (0.312)**	-0.875 (0.510)	-0.871 (0.576)	-0.829 (0.483)
θ	0.207 (0.196)	0.202 (0.211)	0.213 (0.212)	0.211 (0.473)	0.196 (0.065)	0.209 (0.110)

5.2. Recurrent CGD data

The event times for a given patient can be correlated as in the above kidney infection data. We model the recurrent infection survival times, with the two covariates: Treatment x_1 (0=placebo, 1=γ-IFN) and Sex x_8 (0=M(male), 1=F(female)). Notes that x_1 is the main covariate in the clinical trial. The fitted results are given in Table

5.2.1. Here, the two-stage Cox estimates and SEs for Treatment and Sex effects are again the same as the marginal Cox estimates by the GEE approach (Spiekerman and Lin, 1998). All six methods give similar estimation results for fixed effects. According to the Wald test statistic, the Treatment effect at the 5% significance level is significant, but the Sex effect is not. The estimates of θ are different; the one-stage estimates are larger than the two-stage estimates due to underestimation of θ; this fact is confirmed from the simulation results of Tables 4.1.1-4.1.6. The one-stage proposed and PE methods give similar estimation results; this is also confirmed from the simulation results of Tables 4.1.1-4.1.4.

Table 5.2.1. Recurrent CGD data: estimation results of Clayton copula models using the proposed and existing five methods

Baseline hazard function	One-stage			Two-stage		
	Weibull	PE	Proposed	Weibull	PE	Cox
Parameter	Est (SE)	Est (SE)	Est (SE)	Est (SE)	Est (SE)	Est (SE)
Sex :Female	-0.189 (0.351)	-0.162 (0.353)	-0.162 (0.352)	-0.272 (0.372)	-0.255 (0.384)	-0.257 (0.372)
Treatment : γ-IFN	-0.828 (0.281)	-0.860 (0.283)	-0.883 (0.285)	-1.025 (0.302)	-1.058 (0.384)	-1.080 (0.372)
θ	1.288 (0.586)	1.492 (0.663)	1.458 (0.647)	0.710 (0.385)	0.786 (0.311)	0.770 (0.336)

5.3. Multicenter bladder cancer data

We illustrate the proposed method via data from a multicenter bladder cancer clinical trial 30962 conducted by the EORTC (Oddens et al., 2013). The data set used in this study was the duration of

disease-free interval (DFI): the time (days) to the first recurrence after surgery (transurethral resection) in 1,066 patients having the bladder cancer from $q=46$ centers in 13 European countries. Here, the number of patients per center n_i varied from 1 to 63, with mean 23.2 and median 7. Bacillus Calmette-Guerin (BCG) was given after surgery to try for reducing the risk of recurrence. In order to reduce its toxicity which is a disadvantage of BCG, two different doses (1/3 dose, and full dose) and durations of maintenance BCG therapy (1 year and 3 years) were assessed. Out of the 1,066 patients, 594 patients (55.7 per cent) without recurrence were censored at the date of last follow up. In this paper, we aim to find the significant risk factors affecting the time to recurrence among 9 ninepence potential prognostic factor. That is, Trtdose, Trtduration, Age, Gender, TypeBC, Tumsize, Nbtum, Tstage, and Ggrade (G1, G2, G3) which were considered in Table 5.3.1. In previous analysis (Park and Ha, 2019) of the bladder cancer data using AFT (accelerated failure time) random-effect model, the four variables (i.e. Trtduration, TypeBC, Nbtum, and G1) were found to be significant variables.

Table 5.3.1 summarizes the estimation results using the one-stage and two-stage methods. We observe that our method gives very similar results to the two-stage Cox models for estimated regression parameters and θ. We also find that by the Wald test statistic, the same four variables above are significant at the 5% significance level for all 6 methods on Table 5.3.1, which confirms the previous results by Park and Ha (2019).

Table 5.3.1. Bladder cancer data: estimation results of Clayton copula models using the proposed and existing five methods

Baseline hazard function	One-stage			Two-stage		
	Weibull	PE	Proposed	Weibull	PE	Cox
Parameter	Est (SE)	Est (SE)	Est (SE)	Est (SE)	Est (SE)	Est (SE)
x_1 : Trtdose	-0.144 (0.091)	-0.145 (0.090)	-0.146 (0.090)	-0.148 (0.064)	-0.150 (0.037)	-0.153 (0.059)
x_2 : Trtduration	-0.225 (0.092)	-0.193 (0.091)	-0.203 (0.091)	-0.229 (0.083)	-0.203 (0.050)	-0.204 (0.079)
x_3 : Age	-0.001 (0.005)	-0.002 (0.004)	-0.002 (0.004)	-0.001 (0.005)	-0.003 (0.003)	-0.003 (0.004)
x_4 : Gender	0.142 (0.114)	0.167 (0.114)	0.162 (0.114)	0.147 (0.152)	0.161 (0.090)	0.162 (0.146)
x_5 : TypeBC	0.389 (0.102)	0.375 (0.102)	0.382 (0.102)	0.386 (0.080)	0.386 (0.048)	0.386 (0.076)
x_6 : Timsize	0.001 (0.004)	0.002 (0.004)	0.002 (0.004)	-0.001 (0.004)	0.000 (0.002)	0.000 (0.004)
x_7 : Nbtum	0.134 (0.025)	0.129 (0.025)	0.135 (0.025)	0.136 (0.032)	0.134 (0.020)	0.135 (0.030)
x_8 : Tstage	0.060 (0.125)	0.069 (0.125)	0.056 (0.124)	0.021 (0.146)	0.042 (0.082)	0.038 (0.137)
x_9 : G1	-0.334 (0.160)	-0.319 (0.159)	-0.344 (0.159)	-0.344 (0.167)	-0.301 (0.095)	-0.302 (0.152)
x_{10} : G2	-0.240 (0.138)	-0.216 (0.137)	-0.233 (0.137)	-0.259 (0.134)	-0.214 (0.077)	-0.213 (0.125)
θ	0.068 (0.035)	0.094 (0.052)	0.083 (0.045)	0.063 (0.033)	0.082 (0.030)	0.086 (0.053)

Ⅵ. PENALIZED VARIABLE SELECTION IN COPULA SURVIVAL MODELS

In this chapter, we propose a one-stage method for variable selection in the copula survival models based on penalized likelihood. For this purpose, we study four penalty functions.

6.1 Construction of penalized likelihood

For variable selection of the regression parameters under the copula survival models, we consider the Clayton copula survival model having a parametric marginal hazard. For simplicity, we consider only the Weibull marginal hazard having a scale parameter θ_0 and a shape parameter ϕ in Clayton copula model with (2.1.5) and (2.1.6), even if our variable selection method can be easily extended to other marginal parametric hazard functions. Here, the Weibull marginal hazard function is given by

$$\lambda_{ij}(t|x_{ij}) = \lambda_0(t)\exp\left(x_{ij}^T\beta\right), \tag{6.1.1}$$

where $\lambda_0(t) = \theta_0\phi t^{\phi-1} = \phi t^{\phi-1}\exp(\beta_0)$ with $\beta_0 = \log(\theta_0)$ is Weibull baseline hazard. Then, in (6.1.1) the regression parameters $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ and the covariates $x_{ij} = (1, x_{ij1}, \cdots, x_{ijp})^T$ are expressed as $(p+1)\times 1$ vectors.

For the existence of MLEs, we add the following assumption as in Assumption (A3).

Assumption (A4): The penalized log−likelihood $\ell_p(\beta, \phi, \theta)$ is continuous on Ω.

Here $\Omega = \{(\beta^T, \phi^T, \theta)^T | \beta \in \mathbb{R}^{p+1}, \phi > 0, \theta > 0\}$ is the parameter space having a finite dimension.

Proposition 6.1. For the variable selection of regression coefficients β in the copula survival models, we propose a one−stage estimation method using the following penalized log−likelihood ℓ_p

$$\ell_p(\beta, \phi, \theta) = \ell_c(\beta, \phi, \theta) - n\sum_{k=0}^{p} J_\gamma(|\beta_k|), \tag{6.1.2}$$

where ℓ_c is the log−likelihood in (3.1.3) and $J_\gamma(|\cdot|)$ is a penalty function having tuning parameter γ.

Under (A1), (A2) and (A4), the penalized MLEs $(\hat{\beta}, \hat{\phi}, \hat{\theta})$ of (β, ϕ, θ) are obtained by maximizing ℓ_p, i.e.,

$$(\hat{\beta}, \hat{\phi}, \hat{\theta}) = \arg\max_{(\beta, \phi, \theta) \in \Omega} \ell_p. \tag{6.1.3}$$

Proof. Because the parameter space Ω is a finite dimension and $\ell_p(\beta, \phi, \theta)$ is continuous on Ω, the penalized MLEs $(\hat{\beta}, \hat{\phi}, \hat{\theta})$ can be easily obtained by maximizing ℓ_p of (6.1.2) (Green, 1987; Ha et al., 2014). Thus, we can find the MLEs $(\hat{\beta}, \hat{\phi}, \hat{\theta})$ by solving the following three estimating equations simultaneously:

$$\begin{cases} \dfrac{\partial \ell_p}{\partial \beta} = \dfrac{\partial \ell_c}{\partial \beta} - n\dfrac{\partial}{\partial \beta}\left\{\sum_{k=0}^{p} J_\gamma(|\beta_k|)\right\} = 0, \\ \dfrac{\partial \ell_p}{\partial \phi} = \dfrac{\partial \ell_c}{\partial \phi} = 0, \\ \dfrac{\partial \ell_p}{\partial \theta} = \dfrac{\partial \ell_c}{\partial \theta} = 0. \end{cases} \tag{6.1.4}$$

Here we use the local quadratic approximation (LQA; Fan and Li, 2001) for the derivative of the penalty function $J_r(|\beta_k|)$ in (6.1.4). □

For the penalty function $J_\gamma(\cdot)$, we use the four functions, LASSO, ALASSO (adaptive LASSO), SCAD and HL; the forms are shown in Table 6.1.1.

Table 6.1.1. Description of the four penalty function

Penalty function	Description
LASSO (Tibshiran, 1996)	$J_\gamma(\|\beta\|) = \gamma\|\beta\|$
ALASSO (Zou, 2006)	$J_\gamma(\|\beta\|) = \gamma\|\beta\|\,\omega_0,\ \omega_0 = 1/\|\hat{\beta}\|$ ω_0 denotes a known weights vector
SCAD (Fan and Li, 2001)	$J_\gamma'(\|\beta\|) = \gamma I(\|\beta\| \le \gamma) + \dfrac{(a\gamma - \|\beta\|)_+}{a-1} I(\|\beta\| > \gamma),\ a = 3.7$ z_+: the positive part of z
HL (Lee and Oh, 2014)	$J_\gamma(\|\beta\|) \equiv J_{(a,\omega)}(\|\beta\|) = \dfrac{\beta^2}{2au(\|\beta\|)} + \dfrac{(\omega-2)\log u(\|\beta\|)}{2\omega} + \dfrac{u(\|\beta\|)}{\omega}$ $u(\beta) = \left[\{8b\beta^2/a + (2-\omega)^2\}^{1/2} + (2-\omega)\right]/4$

A good penalty function must produce estimates satisfying the three oracle properties, which are unbiasedness, sparsity and continuity (Fan and Li, 2001). The LASSO function is the most usual penalty

function, but does not satisfy the oracle properties. However, Fan and Li (2001, 2002) and Zou (2006) have shown that SCAD and ALASSO perform well with oracle properties, respectively.

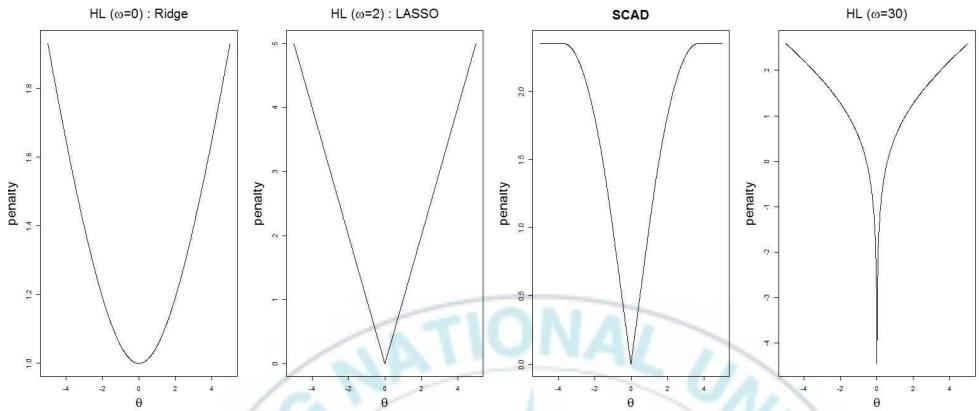


Figure 6.1.1. The four penalty functions

The four penalty functions are shown in Figure 6.1.1. In particular, the HL function changes its shape for the value of ω in the HL penalty function in Table 5.1.1, it becomes a ridge penalty when $\omega \approx 0$ and becomes a LASSO penalty when $\omega = 2$. When $\omega > 2$, it becomes an unbounded form at the origin (Lee and Oh, 2014). The HL (Lee and Oh, 2014) also satisfies the oracle property and provides shrinkage estimators when $\omega > 2$. In this thesis, we use $\omega = 30$ in the HL of Table 6.1.1 from the suggestion by Lee and Oh (2014).

6.2 Penalized variable selection procedure

For penalized variable selection of β, we should estimate parameters (β, ϕ, θ). Below, we show how to estimate these

parameters using the penalized log-likelihood ℓ_p in (6.1.2). In this thesis, we present an efficient profile likelihood procedure which gives a fast convergence. Given association parameter θ, the penalized MLEs $(\hat{\beta}, \hat{\phi})$ of parameters (β, ϕ) in the marginal hazard are obtained by solving the following estimating equations:

$$\frac{\partial \ell_p}{\partial \beta_k} = \frac{\partial \ell_c}{\partial \beta_k} - n\frac{\partial}{\partial \beta_k}\left\{\sum_{k=0}^{p} J_\gamma(|\beta_k|)\right\} = 0, \quad (k = 0, 1, \cdots, p) \tag{6.2.1}$$

and

$$\frac{\partial \ell_p}{\partial \phi} = \frac{\partial \ell_c}{\partial \phi} = 0. \tag{6.2.2}$$

Note that (6.2.1) is adjusted estimating equations derived by adding the penalty terms, while (6.2.2) is the same as standard estimating equation of marginal PH model without penalty. However, with four penalty functions considered in Table 6.1.1, $J_\gamma(\cdot)$ in the estimating equations of β of (6.2.1) becomes non-differentiable at the zero, and does not have continuous second-order derivatives. These problems lead to difficulties in solving (6.2.1). Therefore, we use the LQA (Fan and Li, 2001) for such penalty functions. That is, given an initial value $\beta^{(0)}$ close to the true value of β, the penalty function $J_\gamma(\cdot)$ can be locally approximated by a quadratic function as

$$[J_\gamma(|\beta_k|)]' = J_\gamma'(|\beta_k|)\mathrm{sgn}(\beta_k) \approx \{J_\gamma'(|\beta_k^{(0)}|)/|\beta_k^{(0)}|\}\beta_k \text{ for } \beta_k \approx \beta_k^{(0)}. \tag{6.2.3}$$

According to Ha et al. (2014), the negative Hessian matrix of (β, ϕ)

using ℓ_p is given by

$$H_p = -\frac{\partial^2 \ell_p}{\partial(\beta,\,\phi)^2} = \begin{pmatrix} -\dfrac{\partial^2 \ell_c}{\partial\beta\partial\beta^T} + n\Sigma_\gamma & -\dfrac{\partial^2 \ell_c}{\partial\beta\partial\phi^T} \\ -\dfrac{\partial^2 \ell_c}{\partial\phi\partial\beta^T} & -\dfrac{\partial^2 \ell_c}{\partial\phi^2} \end{pmatrix}, \tag{6.2.4}$$

where $\Sigma_\gamma = \mathrm{diag}\{J_\gamma'(|\beta_j|)/|\beta_j|\}$ is a $(p+1)\times(p+1)$ diagonal matrix. The estimating equations of (6.2.1) and (6.2.2) are solved using the Newton-Raphson method with H_p as in Ha et al.(2014); for the derivation of H_p see Appendix B.

Next, the association parameter θ is estimated by maximizing a profile likelihood based on ℓ_c in (2.1.5). That is, since $\partial\ell_p/\partial\theta = \partial\ell_c/\partial\theta = 0$, we use a copula-based profile likelihood of θ (denoted by $\ell_{cp}(\theta)$)

$$\ell_{cp}(\theta) = \ell_c\big(\theta, \hat\beta(\theta).\hat\phi(\theta)\big) = \ell_c(\theta, \beta, \phi)|_{\beta=\hat\beta(\theta),\,\phi=\hat\phi(\theta)}. \tag{6.2.5}$$

Thus, the profile MLE of θ is defined by

$$\hat\theta = \arg\max_\theta \ell_{cp}(\theta), \tag{6.2.6}$$

and it is obtained by solving the estimating equation (6.2.7)

$$\frac{\partial \ell_p}{\partial\theta}\Big|_{\beta=\hat\beta(\theta),\,\phi=\hat\phi(\theta)} = 0, \tag{6.2.7}$$

where $\hat{\beta}(\theta)$ and $\hat{\phi}(\theta)$ are updated in each iteration. The equation (6.2.7) is also solved by using the Newton-Raphson method with $-\partial^2 \ell_{cp}/\partial\theta^2$.

In order to select tuning parameter γ, we use a type of Bayesian information criterion (BIC) criterion (Ha et al., 2014),

$$\mathrm{BIC}(\gamma) = -2\ell(\beta, \phi) + e(\gamma)\log(n), \tag{6.2.8}$$

where $\ell(\beta, \phi) = \sum_{ij}(\delta_{ij}\log\lambda_{ij} - \Lambda_{ij})$ is the ordinary log-likelihood for the marginal hazard model (2.1.6). Here, $e(\gamma) = tr\left[(H_\beta + n\Sigma_\gamma)^{-1} H_\beta\right]$ is the effective number of parameters (Lee and Nelder, 1996; Ha et al., 2014). Here, $H_\beta = -\partial^2\ell/\partial\beta\partial\beta^T = X^T W X$, where X is model matrix for β and $W = \mathrm{diag}(\Lambda_{ij})$ is weight matrix with $\Lambda_{ij} = \Lambda_0(y_{ij})\exp(x_{ij}^T\beta)$. Notice that

$$\hat{\gamma} = \arg\min_\gamma \mathrm{BIC}(\gamma)$$

is computed via a simple grid search method.

6.3 Fitting algorithm for the variable selection

The variable-selection algorithm for the copula models (2.1.5) having a Weibull marginal hazard is summarized as follows.

- Step 0: Find the initial values of (β, ϕ, θ).

(i) The initial estimates (β, ϕ, θ) of LASSO: use of no-penalty solutions

(ii) The initial estimates of ALASSO, SCAD and HL: use of LASSO

solutions

- Step 1: In the inner loop, we estimate (β, ϕ, θ).
  (i) The estimation of (β, ϕ): It is obtained by solving the (6.2.1) and (6.2.2)
  (ii) The estimation of θ: It is obtained by solving equation (6.2.7)
- Step 2: In the outer loop, we select tuning parameter γ that minimizes BIC(γ) using a grid search method.

After convergence, the estimated SEs of $\hat{\beta}$ is calculated as follows. Because this penalized procedure gives the parameter estimation and variable selection simultaneously, the SEs can be directly obtained via the Newton-Raphson method. Following Fan and Li(2001) and Ha et al. (2014), the SEs for $\hat{\beta}$ are obtained from a sandwich formula:

$$\mathrm{cov}(\hat{\beta}) = (H_\beta + n\Sigma_\gamma)^{-1} H_\beta (H_\beta + n\Sigma_\gamma)^{-1}, \tag{6.3.1}$$

where H_β is given by

$$H_\beta = \left\{ \left(-\frac{\partial^2 \ell_c}{\partial\beta\partial\beta^T} \right) - \left(-\frac{\partial^2 \ell_c}{\partial\beta\partial\phi} \right) \left(-\frac{\partial^2 \ell_c}{\partial\phi^2} \right)^{-1} \left(-\frac{\partial^2 \ell_c}{\partial\phi\partial\beta^T} \right) \right\} \Big|_{\phi=\hat{\phi}}. \tag{6.3.2}$$

6.4 Simulation study for penalized variable selection

The simulation study is demonstrated conducted to evaluate the performance of the proposed variable selection method in a Clayton copula model with Weibull marginal hazard using 200 simulation data. Here, we compare the performances of four variable selection methods using LASSO, ALASSO, SCAD and HL penalties.

According to Kwon et al. (2020), the simulation scheme is designed as follows:

• Event times are simulated from a Clayton copula survival model (2.1.8) having association parameter $\theta = 0.5$ and Weibull marginal function $S(t|x) = \exp\{-t^{\phi}\exp(x^T\beta)\}$ with $\phi = 0.8$ which is decreasing hazard.

• As in Prenen et al. (2017a) and Ha et al. (2019), data are generated via the sampling algorithm of Marshall and Olkin (1988) as described in Chapter Ⅳ.

• Following the simulation setting by Fan and Li (2002), the regression parameters are set to

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)^T = (1, 0.8, 0, 0, 1, 0, 0, 0.6, 0)^T. \qquad (6.4.1)$$

The corresponding covariates are $x = \left(1, x^{*^T}\right)^T$.

• For multicollinearity among the covariates, $x^* = (x_1, \cdots, x_8)^T$ are generated from an AR(1) structure having the correlation coefficient $\rho = 0.5$ (Ha et al., 2014; Park and Ha, 2019). Note that x_1, x_4 and x_7 are important covariates.

• We use three types of sample sizes: $n = \sum_{i=1}^{q} n_i = 200, 400, 600$ with $(q, n_i) = (100, 2), (100, 4), (300, 2)$ for all i, where q is the number of clusters and n_i is the cluster size.

• The censoring times C_{ij} are generated from an exponential distribution having a parameter that is empirically determined to achieve around 20% and 40% censoring rates.

6.5 Simulation result for penalized variable selection

As the measures for variable selection, we consider the following quantities:

• C (Here, the best is 5): The average number of regression coefficients, of the five true zeros that were correctly found to zero.

• IC (Here, the best is 0): The average number of the four true nonzeros incorrectly set to zero.

• PT: The probability of choosing the true model.

• MSE: The mean squared error; it is defined by (Zhang and Lu, 2007)

$$\mathrm{MSE}(\hat{\beta}) = (\hat{\beta} - \beta)^{\mathrm{T}} \Sigma (\hat{\beta} - \beta),$$

where Σ is the population covariance matrix of covariates.

(i) The simulation results are presented in Table 6.3.1. The MSE is increased as the censoring rate is increased and it is decreased as the sample size is increased. The ALASSO, SCAD, and HL methods with oracle properties overall perform better, and are all superior to LASSO in terms of 'PT', 'C' and 'MSE'. The SCAD and HL methods are also improved as the sample size q or n_i increases, even if the censoring rate is as high as 40%. Particularly, SCAD offers the smallest MSE among all the settings, but HL consistently outperforms ALASSO and SCAD in terms of 'PT' and 'C'.

Table 6.3.1. Simulation results using 200 replications under copula survival models

		Censoring 20%				Censoring 40%			
(q, n_i)	Method	C(5)	IC(0)	PT	MSE	C(5)	IC(0)	PT	MSE
$(100, 2)$	LASSO	1.76	0	0	0.068	1.66	0	0.01	0.084
	ALASSO	3.61	0	0.1	0.042	3.41	0	0.12	0.053
	SCAD	4.49	0	0.65	0.030	4.53	0	0.68	0.044
	HL	4.67	0	0.71	0.041	4.74	0	0.75	0.050
$(100, 4)$	LASSO	1.70	0	0	0.042	1.81	0	0	0.052
	ALASSO	3.92	0	0.16	0.018	4.00	0	0.14	0.021
	SCAD	4.55	0	0.73	0.017	4.46	0	0.67	0.023
	HL	4.77	0	0.79	0.017	4.76	0	0.78	0.025
$(300, 2)$	LASSO	1.75	0	0.01	0.027	1.86	0	0	0.032
	ALASSO	3.92	0	0.12	0.013	3.95	0	0.15	0.015
	SCAD	4.63	0	0.77	0.010	4.59	0	0.77	0.013
	HL	4.70	0	0.73	0.012	4.76	0	0.78	0.017

(ii) In Table 6.3.2, we also summarize the frequency which each variable was selected among 200 replications. Four all methods (LASSO, ALASSO, SCAD, and HL) identify almost correctly the three important variables (x_1, x_4, and x_7) including the intercept x_0. However, LASSO selects unimportant variables (x_2, x_3, x_5, x_6, and x_8) much more often than the other three methods (ALASSO, SCAD, and HL) in all the settings, as evident in the simulation results of the frailty models (Fan and Li, 2002; Ha et al., 2014) and AFT random-effect survival models (Park and Ha, 2019).

Table 6.3.2 Simulation results using 200 replications: frequency of variable selection under copula survival models

		Censoring 20%								
(q, n_i)	Method	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$(100, 2)$	LASSO	200	200	129	135	200	124	129	200	131
	ALASSO	200	200	59	51	200	53	51	200	64
	SCAD	200	200	23	23	200	22	17	200	18
	HL	200	200	10	9	200	9	9	200	16
$(100, 4)$	LASSO	200	200	133	131	200	120	134	200	143
	ALASSO	200	200	43	34	200	38	48	200	53
	SCAD	200	200	19	20	200	16	15	200	20
	HL	200	200	15	5	200	9	9	200	9
$(300, 2)$	LASSO	200	200	126	112	200	137	123	200	153
	LASSO	200	200	41	46	200	52	44	200	34
	SCAD	200	200	11	15	200	13	20	200	15
	HL	200	200	12	10	200	14	10	200	14
		Censoring 40%								
(q, n_i)	Method	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$(100, 2)$	LASSO	200	200	131	140	200	135	132	200	131
	ALASSO	200	200	69	51	200	68	62	200	68
	SCAD	200	200	19	23	200	17	19	200	16
	HL	200	200	10	11	200	10	11	200	11
$(100, 4)$	LASSO	200	200	128	117	200	132	125	200	136
	ALASSO	200	200	38	39	200	37	40	200	47
	SCAD	200	200	21	17	200	28	21	200	21
	HL	200	200	15	11	200	10	8	200	11
$(300, 2)$	LASSO	200	200	119	117	200	131	128	200	134
	ALASSO	200	200	35	47	200	46	44	200	39
	SCAD	200	200	12	16	200	14	19	200	21
	HL	200	200	10	14	200	11	11	200	10

(iii) In Table 6.3.3, we summarize the mean, SE, SD, MSE and CP on $\hat{\beta}_1$, $\hat{\beta}_4$ and $\hat{\beta}_7$ estimated from 200 simulations, respectively, for 20% censoring rates in the four variable selection methods. The biases of the SCAD and the HL estimates of β are the smallest compared to those of LASSO and ALASSO. The proposed SE is consistently underestimated as compared to SD in a smaller sample $n=200$ (Hunter and Li, 2005; Ha et al., 2014). However, the SEs in

ALASSO, SCAD and HL are improved because such mismatch between SE and SD is decreased as q or n_i increases to $n=400$ or $n=600$.

Table 6.3.3. Simulation results for coefficients of β_1, $\beta_4$and β_7 among non-zero coefficients of β under copula survival models with Censoring rate 20%

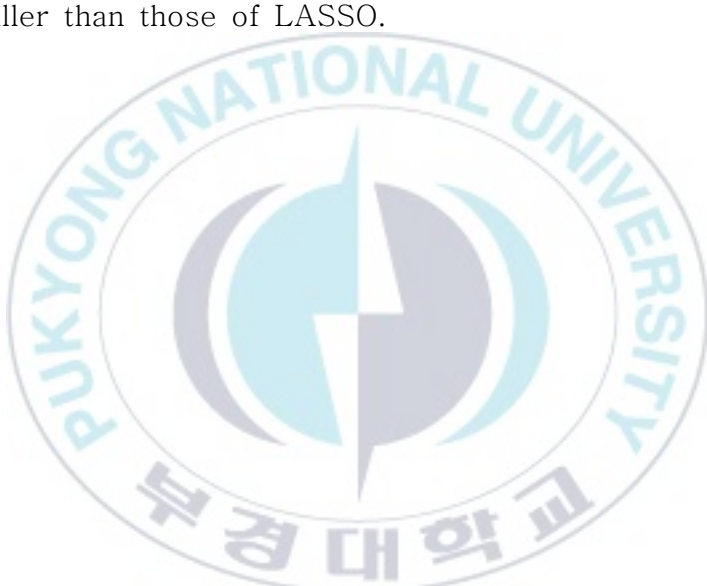
		$\hat{\beta}_1$					$\hat{\beta}_4$					$\hat{\beta}_7$				
(q, n_i)	Method	Mean	SE	SD	MSE	CP	Mean	SE	SD	MSE	CP	Mean	SE	SD	MSE	CP
True value		$\beta_1=0.8$					$\beta_4=1$					$\beta_7=0.6$				
(100,2)	LASSO	0.713	0.067	0.073	0.015	0.715	0.897	0.075	0.080	0.019	0.715	0.531	0.068	0.080	0.009	0.805
	ALASSO	0.783	0.076	0.090	0.008	0.910	0.984	0.006	0.085	0.010	0.885	0.546	0.075	0.093	0.007	0.890
	SCAD	0.806	0.088	0.079	0.008	0.925	1.003	0.086	0.091	0.008	0.920	0.603	0.072	0.071	0.005	0.940
	HL	0.805	0.082	0.086	0.008	0.935	1.005	0.090	0.104	0.011	0.915	0.596	0.076	0.076	0.006	0.950
(100,4)	LASSO	0.735	0.052	0.053	0.007	0.715	0.924	0.059	0.060	0.009	0.680	0.546	0.045	0.051	0.006	0.710
	ALASSO	0.796	0.051	0.054	0.003	0.950	0.916	0.057	0.060	0.004	0.925	0.589	0.046	0.060	0.003	0.900
	SCAD	0.805	0.053	0.052	0.003	0.955	1.010	0.062	0.057	0.004	0.910	0.597	0.047	0.050	0.003	0.840
	HL	0.799	0.052	0.057	0.003	0.925	1.002	0.057	0.064	0.003	0.925	0.595	0.047	0.050	0.002	0.920
(300,2)	LASSO	0.739	0.043	0.045	0.006	0.665	0.935	0.048	0.055	0.007	0.675	0.548	0.040	0.045	0.005	0.695
	ALASSO	0.795	0.045	0.050	0.003	0.925	0.995	0.049	0.053	0.003	0.950	0.596	0.041	0.046	0.002	0.925
	SCAD	0.802	0.045	0.046	0.002	0.955	1.010	0.050	0.054	0.003	0.925	0.602	0.042	0.044	0.002	0.930
	HL	0.800	0.045	0.050	0.002	0.925	0.997	0.049	0.052	0.003	0.940	0.601	0.042	0.046	0.002	0.920

For the convenience of identification of the estimation results of $(\beta_1, \beta_4, \beta_7)$ in Table 6.3.3, the estimation results of $n=200$, $n=400$, and $n=600$ in Figure 6.3.1, 6.3.2 and 6.3.3, were visualized as a box plot, respectively. The results of 200 simulations, under the 40% censoring

rate, of the four variable selection methods, respectively, presented in Table D5 and Figures D1, D2, and D3 of the Appendix D. The results are similar to those using 20% censoring rate.

In summary, we recommend the use of ALASSO, SCAD or HL method to conduct variable selection of regression parameters in the copula survival models (2.1.7) since the three methods identify well both zero and non-zero coefficients.

As shown in the box plots in Figures 6.3.1-6.3.3, the biases of the estimated regression parameters of ALASSO, SCAD and HL are generally smaller than those of LASSO.



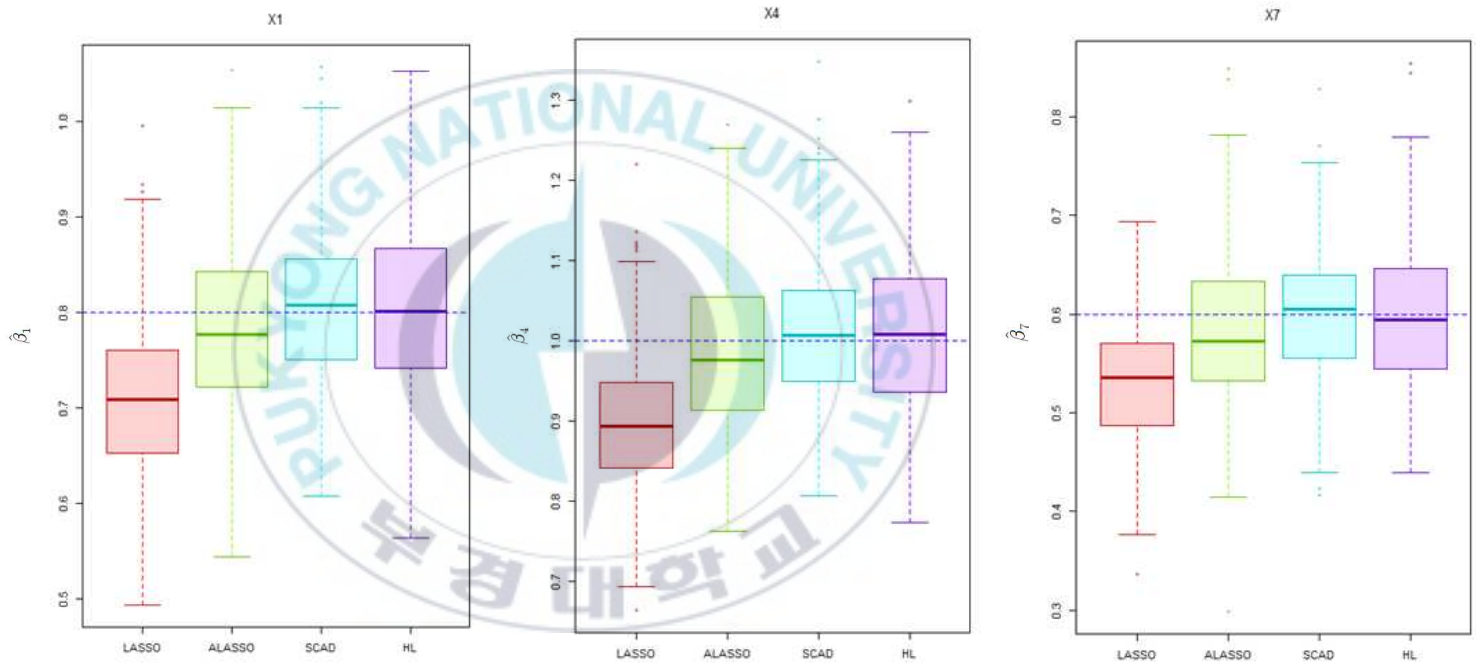


Figure 6.3.1. $(q, n_i) = (100, 2)$: Simulation result of copula variable selection using 200 replications; 20% censoring rate; dotted line, true values of β_1, $\beta_4$and β_7, respectively

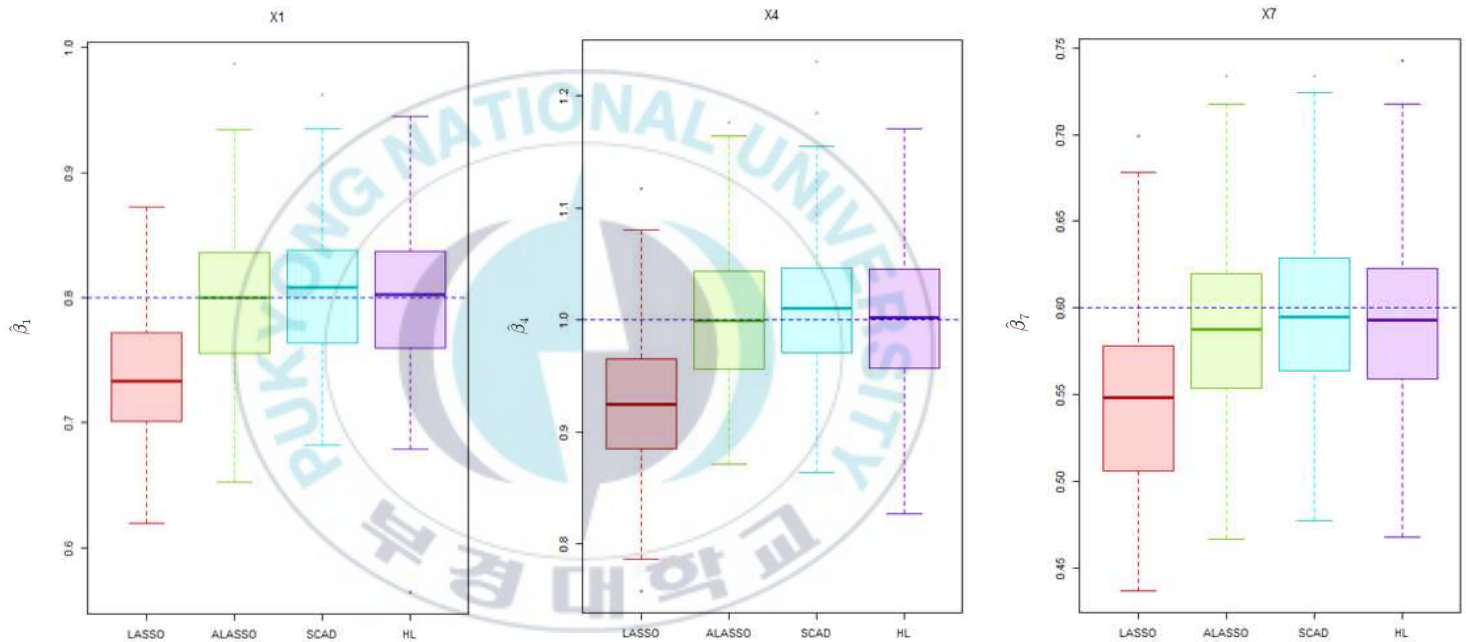


Figure 6.3.2. $(q, n_i) = (100, 4)$: Simulation result of copula variable selection using 200 replications; 20% censoring rate; dotted line, true values of β_1, $\beta_4$and β_7, respectively

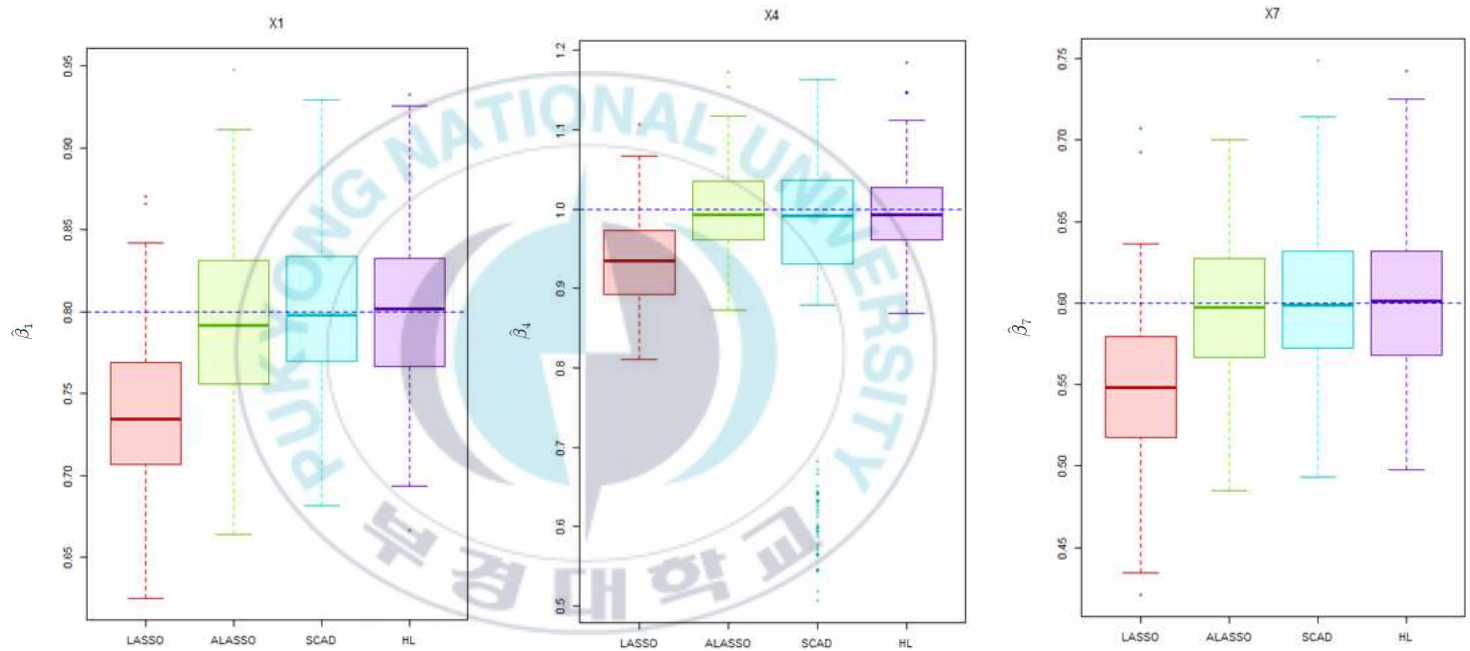


Figure 6.3.3. $(q, n_i) = (300, 2)$: Simulation result of copula variable selection using 200 replications; 20% censoring rate; dotted line, true values of β_1, $\beta_4$and β_7, respectively

6.6. Illustration for penalized variable selection

(1) Kidney infection data

We consider five covariates in the kidney infection data (in R package frailtyHL):

- x_1: Age (in years);
- x_2: Sex (0=male, 1=female);
- x_3: GN (disease type=0);
- x_4: AN (disease type=1);
- x_5: PKD (disease type=2).

Here, Age only is standardized because other covariates are all binary. It is well known that the Sex covariate in the kidney data is of great importance by various survival modeling approaches (Ha et al., 2014, 2017). The fitted estimation results (i.e. the estimated coefficients and their SEs) of the copula survival model via the proposed penalized method are summarized in Table 6.4.1.

The tuning parameters values selected by BIC in (5.2.8) were 0.035, 0.028, 0.218 and 0.066 for the LASSO, ALASSO, SCAD, and HL, respectively. The estimates of the Weibull shape parameter ϕ and association parameter θ are given by $(\hat{\phi}, \hat{\theta})$=(1.034, 0.000), (0.877, 0.143), (0.937, 0.112), (0.995, 0.054) and (0.923, 0.122) for the no−penalty, LASSO, ALASSO, SCAD and HL, respectively. Four all variable selection methods (LASSO, ALASSO, SCAD, and HL) select the intercept term, x_0. The covariate Sex and PKD are significant in all five methods.

In Table 6.4.1, the LASSO and ALASSO, respectively, choose four covariates (x_1, x_2, x_4, and x_5) and three covariates (x_2, x_4, and x_5).

Notice here that the LASSO selects two more covariates (x_1 and x_4) which are not significant under no-penalty. This confirms the simulation results in Table 6.2.1 because the LASSO chooses unimportant variables more frequently than the other two methods, as evident in lower 'C' values of the LASSO in Table 6.3.1. These findings indicate that the LASSO might not properly identify important covariates in the copula survival models, as shown in the frailty survival models (Ha et al., 2014). The SCAD and HL select two covariates (x_2 and x_5) which are significant under no-penalty ($\gamma=0$). Note that the SCAD and HL give high shrinkage estimators that are beneficial in prediction, even though the SCAD shrinks less than the HL. It is also known that the LASSO chooses many covariates with excessive shrinkage in the non-zero regression coefficients (Ha et al., 2017, Lee et al., 2017).

Table 6.4.1. Kidney infection data: estimated coefficients and standard errors using copula survival models

Variable		No-penalty (SE)	LASSO (SE)	ALASSO (SE)	SCAD (SE)	HL (SE)
x_0:	Intercept	-2.093 (0.721)	-1.632 (0.352)	-1.916 (0.488)	-1.910 (0.624)	-1.803 (0.456)
x_1:	Age	0.029 (0.165)	0.001 (0.003)	0 (0)	0 (0)	0 (0)
x_2:	Sex	-1.663 (0.367)	-1.429 (0.217)	-1.425 (0.268)	-1.565 (0.345)	-1.431 (0.259)
x_3:	GN	0.051 (0.408)	0 (0)	0 (0)	0 (0)	0 (0)
x_4:	AN	0.538 (0.396)	0.189 (0.131)	0.109 (0.068)	0 (0)	0 (0)
x_5:	PKD	-1.388 (0.601)	-0.718 (0.246)	-0.952 (0.335)	-1.413 (0.513)	-0.962 (0.329)
$\hat{\phi}$		1.034	0.877	0.937	0.995	0.923
$\hat{\theta}$		0.000	0.143	0.112	0.054	0.122
BIC		691.042	680.744	679.826	681.192	679.895
tuning $\hat{\gamma}$		0	0.035	0.028	0.218	0.066

In addition, it is interested to select a proper variable selection model using the BIC in (6.2.8). Note that the smaller value of BIC indicates a better model. Thus, the BIC in Table 6.4.1 chooses the ALASSO and HL models among the four variable selection models.

(2) Recurrent CGD data

We consider eight covariates in the CGD data (in R package frailtyHL):

- x_1: treatment (0=placebo, 1=γ-IFN);
- x_2: pattern of inheritance (0=autosomal recessive, 1=X-linked);
- x_3: age (in years);
- x_4: height (in cm);
- x_5: weight (in kg);
- x_6: the use of corticosteroids at the time of study entry (0=no, 1=yes);
- x_7: the use of prophylactic antibiotics at the time of study entry (0=no, 1=yes);
- x_8: sex (0=male, 1=female).

Notice that x_1 is the main covariate in this clinical trial. Here, the three covariates (age x_3, height x_4, and weight x_5) are standardized because other covariates are all binary. The fitted estimation results of the Clayton copula models using the proposed penalized method are presented in Table 6.4.2.

In Table 6.4.2, the selected values of the tuning parameters γ by

BIC were, respectively, 0.016, 0.019, 0.189 and 0.191 for LASSO, ALASSO, SCAD and HL. The estimates of ϕ and θ are given by $(\hat{\phi}, \hat{\theta}) = (1.003, 0.980)$, $(0.830, 1.026)$, $(0.937, 1.344)$, $(0.978, 1.376)$ and $(0.937, 1.323)$ for no-penalty, the LASSO, ALASSO, SCAD and HL, respectively. All four variable selection methods (i.e. LASSO, ALASSO, SCAD and HL) also choose the intercept term (x_0), as shown in Table 6.4.2. The LASSO, ALASSO and SCAD select three covariates (x_1, x_3 and x_7), three covariates (x_1, x_3 and x_6), and two covariates (x_1, x_6), respectively. Particularly, the LASSO chooses x_7 which is non-significant under no-penalty, whereas the ALASSO selects three covariates (x_1, x_3 and x_6) which are significant under no-penalty. However, HL selects only the main covariate (x_1) which is also confirmed in the variable selection of the frailty survival model (Ha et al., 2014). We again confirm that the HL shrinks more than the SCAD does. From Table 6.4.2, we also find that selections of covariates of the proposed method are similar to those of the Ha et al.'s (2014) method, as shown in Table 6.4.1.

Furthermore, the BIC in Table 6.4.2 selects the HL model as a proper model for the CGD data, which confirms good performances of the HL in the simulation results of Table 6.3.1.

Table 6.4.2. CGD infection data: estimated coefficients and standard errors using copula survival models

Variable		No-penalty (SE)	LASSO (SE)	ALASSO (SE)	SCAD (SE)	HL (SE)
x_0:	Intercept	-5.922	-4.856	-5.900	-6.065	-5.821
		(0.587)	(0.402)	(0.485)	(0.519)	(0.483)
x_1:	Gamma-IFN	-0.870	-0.666	-0.520	-0.816	-0.726
		(0.266)	(0.189)	(0.161)	(0.258)	(0.222)
x_2:	Inheritance	0.542	0	0	0	0
		(0.263)	(0)	(0)	(0)	(0)
x_3:	Age	-0.795	-0.183	-0.124	0	0
		(0.336)	(0.093)	(0.076)	(0)	(0)
x_4:	Height	0.173	0	0	0	0
		(0.319)	(0)	(0)	(0)	(0)
x_5:	Weight	0.339	0	0	0	0
		(0.348)	(0)	(0)	(0)	(0)
x_6:	Steroids	1.567	0	0	0.896	0
		(0.590)	(0)	(0)	(0.451)	(0)
x_7:	Prophylac	-0.434	-0.384	0	0	0
		(0.305)	(0.166)	(0)	(0)	(0)
x_8:	Sex	-0.578	0	0	0	0
		(0.385)	(0)	(0)	(0)	(0)
	$\hat{\phi}$	1.003	0.830	0.937	0.978	0.937
	$\hat{\theta}$	0.980	1.026	1.344	1.376	1.323
	BIC	1099.521	1080.299	1076.776	1079.694	1075.524
	tuning $\hat{\gamma}$	0	0.016	0.019	0.189	0.191

Ⅶ. DISCUSSION

It is well known that optimization for a copula-based full likelihood involving an unknown baseline hazard function in an infinite-dimensional parameter space is very difficult. In order to overcome this problem, we reduced the infinite dimension to a finite dimension by approximating the baseline hazard to the M-spline basis function with the number $L=5$ of bases, regardless of sample size or censoring rate. In this consideration, we proposed a one-stage M-spline copula modeling approach which effectively reflects on the dependence among survival times.

In copula models, the two-stage likelihood approach estimates separately the marginal parameters and the association parameter, which leads to an inefficient inference result. However, the use of one-stage likelihood approach gives an efficient inference result by jointly estimating both parameters (Marra and Radice, 2020; Cheng et al,, 2014; Romeo et al., 2018).

We have first shown that the proposed one-stage M-spline method performs well via simulation study and three real data sets. In particular, we have found through simulation study that the proposed method gives similar estimation results with the one-stage PE method when the strength of association is not high. However, our method provides better estimation results when the strength of association is high because the one-stage PE method gives larger variations (i.e. SD and MSE) for estimated regression parameters, leading to lower CPs. The remaining methods (one-stage Weibull, and two-stage Weibull, PE and Cox) have shown inferior performances in the estimation of β and/or θ. The implementation of proposed method is simple and gives a fast fitting algorithm for clustered copula

regression models by using the five basis functions in the cubic M-spline.

As shown in the simulation study, the proposed one-stage spline method is robust against misspecification of baseline hazard due to the flexibility of the M-spline to the underlying hazard function. However, we have also found via simulation study that when the assumed Clayton copula model is incorrectly specified as Gumbel-Hougaard copula model, the estimated regression parameters β by the proposed method are biased. Care is necessary for the inference of β by the proposed method when a copula function is misspecified.

For the variable selection procedure in copula models, we also proposed a one-stage copula estimation method based on the penalized likelihood. We have demonstrated via simulation studies and two real data sets that the proposed procedure with SCAD or HL penalty works well. In particular, we have found that the HL method gives better performance in terms of measures of variable selection. The advantage of our variable selection method is that it can be easily implemented by a slight modification to the existing likelihood estimation procedures (Ha et al., 2019).

In this thesis, we have proposed one-stage M-spline and variable selection methods under Clayton copula models only. For further extensive study, it would be necessary to extend the proposed method to other parametric copula (e.g. Gumbel-Hougaard) or robust copula function (Gribkova and Lopez, 2015). Another extensions to clustered competing risks (Emura et al., 2020) or interval censoring (Sun and Ding, 2019) would be also an interesting future work. In addition, developing a penalized variable selection using a M-spline copula modeling approach would be also an interested topic.

References

[1] Andersen E. W. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis*, 11, 333-350.

[2] Balan, T. A. & Putter, H. (2017) FrailtyEM: An R package for estimating semiparametric shared frailty models. CRAN.

[3] Breslow N. E. (1972). Discussion on Professor Cox's paper. *Journal of the Royal Statatistical Society: Series B (Methodol),* 34, 216-217.

[4] Chen X., Fan Y. & Tsyrennikov V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association,* 101, 1228-1240.

[5] Chen X., Hu T. & Sun J. (2017). Sieve maximum likelihood estimation for the proportional hazards model under informative censoring. *Computational Statistics & Data Analysis,* 112, 224-234.

[6] Cheng G., Zhou L., Chen X. & Huang J. Z. (2014). Efficient estimation of semiparametric copula models for bivariate survival data. *Journal of Multivariate Analysis,* 123, 330-344.

[7] Commenges D. & Jacqmin-Gadda H. (2016). *Dynamical biostatistical models.* Chapman and Hall: London.

[8] Cox D. R. (1972). Regression models and life-tables. *Journal of the Royal Statatistical Society: Series B (Methodol),* 34, 187-202.

[9] Cox D. R. & Hinkley D. V. (1974). *Theoretical Statistics.* Chapman and Hall: London.

[10] Duchateau L. & Janssen P. (2008). *The frailty model.* Springer: Berlin.

[11] Emura T. (2019). joint.Cox: joint frailty-copula models for

tumour progression and death in meta-analysis. CRAN.

[12] Emura T., Matsui S. & Rondeau V. (2019). *Survival Analysis with Correlated Endpoints; Joint Frailty-Copula Models.* JSS Research Series in Statistics, Springer.

[13] Emura T., Nakatochi M, Murotani K. & Rondeau V. (2017). A joint frailtycopula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research,* 26, 2649-2666.

[14] Emura T., Shih J. H., Ha I. D. & Wilke R. A. (2019). Comparison of the marginal hazard model and the sub-distribution hazard model for competing risks under an assumed copula, *Stat Methods Med Res.* 29, 2307-2327.

[15] Fan J. & Li R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association,* 96, 1348-1360.

[16] Fan J. & Li R. (2001). Variable Selection for Cox's proportional Hazards Model and Frailty Model. *Ann. Statist.* 30, 74-99.

[17] Fleming T. R. & Harrington D. P. (1991). *Counting processes and survival analysis.* New York: Wiley.

[18] Geman S. & Hwang C. R. (1982). Nonparmetric maximum likelihood estimation by the method of sieves. *Annals of Statistics,* 10, 401-414.

[19] Goethals T., Janssen P. & Duchateau L. (2008). Frailty models and copulas: similarities and differences. *Journal Applied Statistics,* 35, 1071-1079.

[20] Good I. J. & Gaskins R. A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika,* 58, 255-257.

[21] Green P. J. (1987). Penalized Likelihood for General Semi-parametric Regression Models. *International Statistical*

Review, 55, 245-259.

[22] Grenander U. (1981). *Abstract Inference*. Wiley, New York.

[23] Gribkova S. & Lopez O. (2015). Non-parametric copula estimation under bivariate censoring. *Scandinavian Journal of Statistics,* 42, 925-946.

[24] Gumbel E. J. (1960). Distribution des valeurs extremes en plusieurs dimensions. *Publications de l' Institut de Statistique de L' Universite de Paris,* 9, 171-173.

[25] Ha I. D., Pan J., Oh S. & Lee Y. (2014). Variable selection in general frailty models using penalized h-likelihood. *Journal of Computational and Graphical Statistics,* 23, 1044-1060.

[26] Ha I. D., Jeong J. H. & Lee Y. (2017). *Statistical modelling of survival data with random effects: h-likelihood approach*. Singapore: Springer.

[27] Ha, I. D., Noh M., Kim J. & Lee Y. (2018). frailtyHL: frailty models using h-likelihood. CRAN.

[28] Ha I. D., Kim J. M. & Emura T. (2019). Profile likelihood approaches for semiparametric copula and frailty models for clustered survival data. *Journal of Applied Statistics,* 46, 2553-2571.

[29] Hougaard P. (1986). Survival Models for Heterogeneous Populations Derived from Stable Distributions. *Biometrika,* 73, 387-396.

[30] Hougaard P. (2000). *Analysis of multivariate survival data.* New York: Springer.

[31] Hunter D. & Li R. (2005). Variable selection using M algorithms. *The Annals of Statistics,* 33, 1617-1642.

[32] Joe H. (1997). *Multivariate models and dependence concepts.* London: Chapman and Hall.

[33] Kwon S. & Ha I. D. (2019). Comparison of Copula and Frailty Models for Correlated Survival Data. *Journal of the Korean Data & Information Science Society,* 30, 551-562.

[34] Kwon S., Ha I. D. & Kim J. M. (2020). Penalized variable selection in copula survival models for clustered time-to-event data. *Journal of Statistical Computation and Simulation,* 90, 657-675.

[35] Lee Y. & Nelder J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society B,* 58, 619-678.

[36] Lee Y. & Oh H. S. (2014). A new sparse variable selection via random-effect model. *Journal of Multivariate Analysis,* 125, 89-99.

[37] Lee Y., Nelder J. A. & Pawitan Y. (2017). *Generalised Linear Models with Random Eects: unied analysis via h-likelihood. 2nd edn*. Chapman and Hall: Baca Raton.

[38] Liang, K. Y. & Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika,* 73, 13-22.

[39] Liu D., Kalbfleish J. D. & Douglas S. E. (2011) A Positive stable Frailty Model for Clustered Failure Time Data with Covariate-Dependent Frailty. *Biometrics,* 67, 8-17

[40] Lu M., Zhang Y. & Huang J. (2007) Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika,* 94, 705-718.

[41] Ma L., Hu T. & Sun J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika,* 102, 731-738

[42] Marra G. & Radice R. (2020). Copula link-based additive models for rightcensored event time data. *Journal of the American*

Statistical Association, 115, 886-895

[43] Marshall A. W. & Olkin I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association,* 83, 834-841.

[44] McGilchrist C. A. & Aisbett C. W. (1991). Regression with frailty in survival analysis. *Biometrics,* 47, 461-466.

[45] Nielsen, G. G., Gill, R. D., Andersen, P. K. & Sorensen. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statististics,* 19, 25-44.

[46] Oakes, D. (1989). Bivariate survival models induced by frailty. *Journal of the American Statistical Association,* 84, 487-493.

[47] Park E. & Ha I. D. (2019). Penalized variable selection for accelerated failure time models with random effects. *Statistics in Medicine,* 38, 878-892.

[48] Prenen L., Braekers R. & Duchateau L. (2017 a). Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of the Royal Statistical Society B,* 79, 483-505.

[49] Prenen L., Braekers R., Duchateau L. & Troyer E. D. (2017 b). Sunclarco: Survival analysis using copulas. CRAN.

[50] Ramsay J. O. (1988). Monotone regression splines in action. *Statistical Science,* 3, 425-441.

[51] Romeo J., Meyer R. & Gallardo D. (2018). Bayesian bivariate survival analysis using the power variance function copula. *Lifetime Data Analysis,* 24, 355-383.

[52] Shih J. H. & Emura T. (2020). Penalized Cox regression with a five-parameter spline model. *Communications in Statistics-Theory and Methods*.

[53] Shih J. H. & Louis T. A. (1995). Inferences on the association

parameter in copula models for bivariate survival data. *Biometrics,* 51, 1384-1399.

[54] Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Publications de l' Institut Statistique de l' Universite de Paris*, 8, 229-231.

[55] Spiekerman C. F. & Lin D. Y. (1998). Marginal Regression Models for Multivariate Failure Time Data. *Journal of the American Statistical Association,* 93, 1164-1175.

[56] Sun T. & Ding Y. (2021). Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*, 22, 315-330.

[57] Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B,* 58, 267-288.

[58] Vaupei J. W., Manton K. G. & Stallard E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography,* 16, 439-454.

[59] Zou H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association,* 101, 1418-1429.

군집된 다변량 생존 자료에 대한 1단계 코플라 모형 접근법

권숙희

부 경 대 학 교　대 학 원　통 계 학 과

요　　약

군집된 다변량 생존 시간 자료의 분석을 위해 코플라 모형과 프레일티 모형이 폭 넓게 사용되어 왔다. 코플라 모형은 주변 분포와 코플라 함수로 구성된다. 코플라 모형은 주변 모형인 반면에 프레일티 모형은 조건부 모형이다. 특히 대표적 코플라 족인 아르키메데스 코플라 함수는 이러한 자료 간 연관성을 모형화하는 데 유용하다. 일반적으로 코플라 모형에서 가능도 기반한 추론은 1단계 및 2단계 추정 방법이 사용되어왔다. 2단계 추정 절차는 코플라의 주변 모수와 의존성 모수를 독립적으로 추정하기 때문에 비효율적인 추정 결과를 제공할 수 있다. 하지만 효율적인 1단계 추정 절차는 미지의 주변 기저 위험 함수를 갖는 가능도의 복잡성으로 인해 주변 분포의 제한된 모수적 가정하에서 주로 개발되어왔다.

본 논문에서는 1단계 가능도 절차에 기반한 융통성있는 M-스플라인 아르키메데스 코플라 모형 접근법을 제안한다. 즉, 가능도의 복잡성을 줄이기 위해 미지의 주변 기저 위험은 M-스플라인 기저 함수를 기반으로 모형화한다. 제안된 방법의 추정 절차를 유도하고, 이론적 성질을 또한 연구한다. 모의실험에 의하면 제안된 1단계 추정 방법이 기존의 1단계 및 2단계 방법보다 합리적인 편의 추정 및 보다 효율적인 추론 결과를 제공함을 보여준다. 세 가지 실제 자료의 분석을 통해 제안된 방법을 예증한다. 또한 본 논문에서는 벌점화 가능도를 기반으로 한 1단계 추정 방법을 사용하여 코플라 생존 모형에서 변수선택 절차를 제안한다. 제안된 변수선택 방법의 성능은 모의실험 연구를 통해 입증하고 새로운 방법의 유용성은 두 가지 임상 자료의 분석을 통해 예증한다.

Appendix A. M-Spline Basis Functions

This appendix defines the M-spline basis functions used in $\lambda_0(t;h)=\sum_{l=1}^{5} h_l M_l(t)=h^T M(t)$. For a knot sequence $\xi_1<\xi_2<\xi_3$ with an equally spaced mesh $\Delta=\xi_2-\xi_1=\xi_3-\xi_2$, let $z_1(t)=(t-\xi_1)/\Delta$, $z_2(t)=(t-\xi_2)/\Delta$ and $z_3(t)=(t-\xi_3)/\Delta$. Define M-spline basis functions as

$$
\begin{aligned}
M_1(t)&=\frac{I(\xi_1\le t<\xi_2)}{\Delta}\{-4z_2(t)^3\},\\
M_2(t)&=\frac{I(\xi_1\le t<\xi_2)}{\Delta}\left\{\frac{7}{2}z_1(t)^3-9z_1(t)^2+6z_1(t)\right\}+\frac{I(\xi_2\le t<\xi_3)}{\Delta}\left\{-\frac{1}{2}z_3(t)^3\right\},\\
M_3(t)&=\frac{I(\xi_1\le t<\xi_2)}{\Delta}\{-2z_1(t)^3+3z_1(t)^2\}+\frac{I(\xi_2\le t<\xi_3)}{\Delta}\{2z_2(t)^3-3z_2(t)^2+1\},\\
M_4(t)&=\frac{I(\xi_1\le t<\xi_2)}{\Delta}\left\{\frac{1}{2}z_1(t)^3\right\}\\
&\quad+\frac{I(\xi_2\le t<\xi_3)}{\Delta}\left\{-\frac{7}{2}z_2(t)^3+\frac{3}{2}z_2(t)^2+\frac{3}{2}z_2(t)+\frac{1}{2}\right\},\\
M_5(t)&=\frac{I(\xi_2\le t<\xi_3)}{\Delta}\{4z_2(t)^3\}
\end{aligned}
$$

Define the I-spline basis function, $I_l(t)=\int_{\xi_1}^{t} M_l(\omega)d\omega$, which can be written as

$$
\begin{aligned}
I_1(t)&=I(\xi_1\le t<\xi_2)\{-z_2(t)^4\}+1,\\
I_2(t)&=I(\xi_1\le t<\xi_2)\left\{\frac{7}{8}z_1(t)^4-3z_1(t)^3+3z_1(t)^2\right\}+I(\xi_2\le t<\xi_3)\left\{-\frac{1}{8}z_4(t)^3+1\right\},\\
I_3(t)&=I(\xi_1\le t<\xi_2)\left\{-\frac{1}{2}z_1(t)^4+z_1(t)^3\right\}
\end{aligned}
$$

$$+ I(\xi_2 \le t < \xi_3)\left\{\frac{1}{2}z_2(t)^4 - z_2(t)^3 + z_2(t) + \frac{1}{2}\right\},$$

$$I_4(t) = I(\xi_1 \le t < \xi_2)\left\{\frac{1}{8}z_1(t)^4\right\}$$

$$+ I(\xi_2 \le t < \xi_3)\left\{-\frac{7}{8}z_2(t)^4 + \frac{1}{2}z_2(t)^3 + \frac{3}{4}z_2(t)^2 + \frac{1}{2}z_2(t) + \frac{1}{8}\right\},$$

$$I_5(t) = I(\xi_2 \le t < \xi_3)\left\{z_2(t)^4\right\}$$

Appendix B. Derivations

Derivations of the second derivatives in H_p of (6.2.4) and (6.2.7) under the Clayton copula model with Weibull marginal hazard

The log-likelihood in (3.3.1) under the copula model with Weibull marginal hazard is given by

$$\ell_c = \sum_{ij} \delta_{ij}\{\log\lambda_{ij} + \theta\Lambda_{ij}\} - \sum_{i=1}^{q}\left[\left(d_i + \theta^{-1}\right)\log\left(1 + S_{i+}^{*}\right) - \sum_{a=0}^{d_i-1}(1+a\theta)\right],$$

where $\lambda_{ij} = \Lambda'_{ij}$, $\Lambda_{ij} = \Lambda_0(y_{ij})\exp\left(x_{ij}^T\beta\right) = y_{ij}^{\phi}\exp\left(x_{ij}^T\beta\right)$ and $S_{i+}^{*} = \sum_{j=1}^{n_i}(S_{ij}^{-\theta} - 1)$ with $S_{ij} = S_j(y_{ij}|x_{ij}) = \exp(-\Lambda_{ij})$. Given θ, the first derivatives of $(\beta,\ \phi)$ are as follows:

$$\frac{\partial\ell_c}{\partial\beta_k} = \sum_{ij}\delta_{ij}(1+\theta\Lambda_{ij})x_{ijk} - \sum_{i}\left(d_i + \theta^{-1}\right)\left\{\sum_{j}\theta\Lambda_{ij}e^{\theta\Lambda_{ij}}x_{ijk}/\left(S_{i+}^{*}+1\right)\right\},\ (k=0,1,\cdots,p).$$

$$\frac{\partial\ell_c}{\partial\phi} = \sum_{ij}\delta_{ij}\left\{\phi^{-1} + (1+\theta\Lambda_{ij})\log(y_{ij})\right\} - \sum_{i}\left(d_i + \theta^{-1}\right)\left\{\sum_{j}\theta\Lambda_{ij}e^{\theta\Lambda_{ij}}\log(y_{ij})/\left(S_{i+}^{*}+1\right)\right\},\ (k=0,1,\cdots,p).$$

For the variable selection of β, we use the penalized likelihood ℓ_p in (5.1.1) with the copula-based likelihood ℓ_c in (3.3.1). For this

purpose, we need to compute the following estimating equations of β using ℓ_p:

$$\frac{\partial \ell_p}{\partial \beta_k} = \frac{\partial \ell_c}{\partial \beta_k} - n \sum_k J_\gamma'(|\beta_k|)\mathrm{sgn}(|\beta_k|),\ (k = 0,1,\cdots,p),$$

where $\mathrm{sgn}(\cdot)$ is the sign function. Note that $\partial \ell_p / \partial \phi = \partial \ell_c / \partial \phi$. The negative Hessian matrix H_p in (5.2.4) with the second derivatives of ℓ_p with respect to (β, ϕ) is given by

$$H_p = -\frac{\partial^2 \ell_p}{\partial(\beta, \phi)^2} = \begin{pmatrix} -\frac{\partial^2 \ell_p}{\partial \beta \partial \beta^T} + n\Sigma_\gamma & -\frac{\partial^2 \ell_c}{\partial \beta \partial \phi^T} \\ -\frac{\partial^2 \ell_c}{\partial \phi \partial \beta^T} & -\frac{\partial^2 \ell_c}{\partial \phi^2} \end{pmatrix},$$

where

$$\begin{aligned} -\frac{\partial^2 \ell_c}{\partial \beta_k \partial \beta_s} &= -\sum_{ij} x_{ijk}\delta_{ij}\theta\Lambda_{ij}x_{ijs} - \sum_{ij} x_{ijk}\left\{(d_i + \theta^{-1})\theta\Lambda_{ij}e^{\theta\Lambda_{ij}}/(S_{i+}^* + 1)\right\}x_{ijs} \\ &\quad + \sum_{ij} x_{ijk}(d_i + \theta^{-1})\left[(\theta\Lambda_{ij})^2 e^{\theta\Lambda_{ij}} - (\theta\Lambda_{ij}e^{\theta\Lambda_{ij}})^2/(S_{i+}^* + 1)\right]/(S_{i+}^* + 1)x_{ijs} \end{aligned}$$

$$\begin{aligned} -\frac{\partial^2 \ell_c}{\partial \beta_k \partial \phi} &= -\sum_{ij} x_{ijk}\delta_{ij}\theta\Lambda_{ij}\log(y_{ij}) - \sum_{ij} x_{ijk}\left\{(d_i + \theta^{-1})\theta\Lambda_{ij}e^{\theta\Lambda_{ij}}/(S_{i+}^* + 1)\right\}\log(y_{ij}) \\ &\quad + \sum_{ij} x_{ijk}(d_i + \theta^{-1})\left[(\theta\Lambda_{ij})^2 e^{\theta\Lambda_{ij}} - (\theta\Lambda_{ij}e^{\theta\Lambda_{ij}})^2/(S_{i+}^* + 1)\right]/(S_{i+}^* + 1)\log(y_{ij}) \end{aligned}$$

and

$$\begin{aligned} -\frac{\partial^2 \ell_c}{\partial \beta_k \partial \phi} &= -\sum_{ij} \log(y_{ij})\delta_{ij}\theta\Lambda_{ij}\log(y_{ij}) - \sum_{ij} \log(y_{ij})\left\{(d_i + \theta^{-1})\theta\Lambda_{ij}e^{\theta\Lambda_{ij}}/(S_{i+}^* + 1)\right\}\log(y_{ij}) \\ &\quad + \sum_{ij} \log(y_{ij})(d_i + \theta^{-1})\left[(\theta\Lambda_{ij})^2 e^{\theta\Lambda_{ij}} - (\theta\Lambda_{ij}e^{\theta\Lambda_{ij}})^2/(S_{i+}^* + 1)\right]/(S_{i+}^* + 1)\log(y_{ij}) \\ &\quad + \phi^{-2}\sum_{ij}\delta_{ij} \end{aligned}$$

Next, for the estimation of association parameter θ, we use ℓ_c since $\partial \ell_p / \partial \theta = \partial \ell_c / \partial \theta$. The first derivative of θ is given by

$$\frac{\partial \ell_c}{\partial \theta} = \sum_{ij} \delta_{ij} \Lambda_{ij} + \theta^2 \sum_i \log\left(S_{i+}^* + 1\right) - \sum_i \left(d_i + \theta^{-1}\right) \left\{ \sum_j \Lambda_{ij} e^{\theta \Lambda_{ij}} / \left(S_{i+}^* + 1\right) \right\} + \sum_i \sum_{a=0}^{d_i - 1} \frac{a}{1 + a\theta}.$$

This leads to the negative second derivative, given by

$$-\frac{\partial^2 \ell_c}{\partial \theta^2} = -2 \sum_{ij} \left(S_{i+}^* + 1\right) / \theta^3 + 2 \sum_{ij} x_{ijk} \Lambda_{ij} e^{\theta \Lambda_{ij}} / \left\{ \theta^2 \left(S_{i+}^* + 1\right) \right\} + \sum_i \left(d_i + \theta^{-1}\right) \left[\sum_j (\Lambda_{ij})^2 e^{\theta \Lambda_{ij}} / \left(S_{i+}^* + 1\right) + \left\{ \sum_j \Lambda_{ij} e^{\theta \Lambda_{ij}} / \left(S_{i+}^* + 1\right) \right\}^2 \right] - \sum_i \sum_{a=0}^{d_i - 1} \left\{ \frac{a}{(1 + a\theta)} \right\}^2.$$

Thus, the estimating equation (5.2.7) of θ is easily solved using the Newton Raphson method, with the first and second derivatives above.

Appendix C. R Codes

(C.1) one-stage M-spline copula estimation for kidney, CGD and Bladder cancer data

```
rm(list=ls())
library(survival)
library(joint.Cox)
##### Define log-likelihood function #####
logL= function(para){
beta = para[1:p]
theta=para[p+1]
g1=exp(para[(p+2):(p+k+1)])#g1=exp(h):baseline-hazard parameters in
      M-spline
eta <- exp(X%*%beta)
tmin = min(t_event);tmax = max(t_event)
lam<- M.spline(t_event,tmin,tmax)%*%g1*eta #M-spline for hazard
Lam<- I.spline(t_event,tmin,tmax)%*%g1*eta #I-spline for cumulative hazard
Sur <- exp(-Lam)
Sur_s <- t(Z)%*%(Sur^(-theta))-ni
di <- t(Z)%*%event
sum1 <- 0
for(i in 1:q){
su1 <-0
for(a in 0:(di[i]-1)){
su1<- su1+log(1+a*theta)
ifelse(di[i]<1,su1<-0,su1) }
sum1<-sum1 + su1 }
loglike
      <-sum(event*(log(lam)+theta*Lam))-sum((di+1/theta)*log(1+Sur_s))+
      sum1
return(loglike) #log-likelihood
```

```
}
##### Model fitting for kidney data #####
data(kidney)
t_event = kidney$time; event = kidney$status
sex=as.integer(kidney$sex)-1
X<-model.matrix(~0+kidney$age+sex)
p<-ncol(X) # No. of covariates
q<-length(unique(kidney$id)) # No. of clusters
n<-nrow(X) # n: total sample size
Z=model.matrix(~0+factor(kidney$id))
ni <- t(Z)%*%as.matrix(rep(1,n)) # ni: cluster size
k<-5 # No. of knots
para_est = c(0,0,0.5,rep(0,k)) # initial values of (beta,theta,h)
kid_fit = optim(para_est,logL,method = "BFGS",
control = list(fnscale = -1),hessian = TRUE)
V <- solve(-kid_fit$hessian) # inverse of negative Hessian matrix
Estimate<-kid_fit$par[1:(p+1)]
SE<- sqrt(diag(V))[1:(p+1)]
kidney_result<-rbind(Estimate,SE)
colnames(kidney_result)<-c("Age", "Sex","theta")
print(kidney_result)
##### Model fitting for CGD data #####
data(cgd)
time=cgd$tstop-cgd$tstart
t_event = time; event = cgd$status
treat=as.integer(cgd$treat)-1
sex=as.integer(cgd$sex)-1
X<-model.matrix(~0+ treat +sex)
p<-ncol(X) # No. of covariates
q<-length(unique(cgd$id)) # No. of clusters
n<-nrow(X) # n: total sample size
Z=model.matrix(~0+factor(cgd$id))
```

```
ni <- t(Z)%*%as.matrix(rep(1,n)) # ni: cluster size
k<-5 # No. of knots
para_est = c(0,0,0.5,rep(0,k)) # initial values of (beta,theta,h)
cgd_fit = optim(para_est,logL,method = "BFGS",
control = list(fnscale = -1),hessian = TRUE)
V <- solve(-cgd_fit$hessian) # inverse of negative Hessian matrix
Estimate<-cgd_fit$par[1:(p+1)]
SE<- sqrt(diag(V))[1:(p+1)]
cgd_result<-rbind(Estimate,SE)
colnames(cgd_result)<-c("Treat","Sex","theta")
print(cgd_result)
##### Model fitting for Bladder cancer data #####
eortc<-read.csv('eortcdata_BCG.csv',sep="," ,header=T)
data(eortc)
eortc$g1=ifelse(eortc$ggrade==1,1,0)
eortc$g2=ifelse(eortc$ggrade==2,1,0)
time=eortc$timeDFI
t_event = time; event = eortc$statusDFIc
treat=as.integer(cgd$treat)-1
sex=as.integer(cgd$sex)-1
X<-model.matrix(~0+trtdose+trtduration+age+gender+typeB
  +tumsize+nbtum+tstage+g1+g2, data=p<-ncol(X) # No. of covariates
q<-length(unique(eortc$institution)) # No. of clusters
n<-nrow(X) # n: total sample size
Z=model.matrix(~0+factor(eortc$institution))
ni <- t(Z)%*%as.matrix(rep(1,n)) # ni: cluster size
k<-5 # No. of knots
para_est = c(rep(1,10),0.5,rep(0,k)) # initial values of (beta,theta,h)
eortc_fit = optim(para_est,logL,method = "BFGS",
control = list(fnscale = -1),hessian = TRUE)
V <- solve(-eortc_fit$hessian) # inverse of negative Hessian matrix
Estimate<-eortc_fit$par[1:(p+1)]
```

```
SE<- sqrt(diag(V))[1:(p+1)]
eortc_result<-rbind(Estimate,SE)
colnames(eortc_result)<-c("trtdose","trtduration","age","gender","typeBC",
"tumsize","nbtum","tstage","g1","g2", "theta")
print(eortc_result)
```

(C.2) Penalized variable selection of copula regression model

```
#Prior to running, set working directory to file location
rm(list=ls())
setwd("G:/Copula")
source("Copula_VS_NR.txt")
#== kidney infection data(5 covariates)==#
library(frailtyHL)
data(kidney)
kidney$age<-(kidney$age-mean(kidney$age))/sd(kidney$age)
kidney$GN<-as.numeric(kidney$disease=="GN")
kidney$AN<-as.numeric(kidney$disease=="AN")
kidney$PKD<-as.numeric(kidney$disease=="PKD")
kidney$sex<-kidney$sex
kidney$id<-kidney$id
attach(kidney)
kidney.formula<- Surv(time,status)~age+sex+GN+AN+PKD +id
beta00<-c(0,0,0,0,0,0) #initial values
phi0=1
theta0=0.01
# NO_PENALTY
kid_res<-copula.vs(kidney.formula,penalty="LASSO",
      tun_range=c(0), beta=beta00,phi=phi0,
       theta=theta0,data="kidney",maxiter=2000)
kid_res
beta0<-kid_res$Est_beta[,1] #initial values using No_penalty
phi0<-kid_res$Est_phi[1] #initial values using No_penalty
```

```
theta0<-kid_res$Est_theta[1] #initial values using No_penalty
# LASSO
kid_res.LASSO<-copula.vs(kidney.formula,penalty="LASSO",
tun_range=seq(0,0.1, 0.001),beta=beta0,phi=phi0,
theta=theta0, data="kidney",maxiter=2000)
kid_res.LASSO
beta0L<-kid_res.LASSO$Est_beta[,1] #initial values using LASSO
phi0L<-kid_res.LASSO$Est_phi[1] #initial values using LASSO
theta0L<-kid_res.LASSO$Est_theta[1] #initial values using LASSO
# ALASSO
kid_res.ALASSO<-copula.vs(kidney.formula,penalty="ALASSO",
tun_range=seq(0,0.1,0.001),beta=beta0L,phi=phi0L,
theta=theta0L, weight0=abs(1/beta0), data="kidney",
maxiter=2000)
kid_res.ALASSO
# SCAD
kid_res.SCAD<-copula.vs(kidney.formula,penalty="SCAD",
tun_range=seq(0,0.3,0.001),beta=beta0L,phi=phi0L,
theta=theta0L, data="kidney",maxiter=2000)
kid_res.SCAD
# HL
kid_res.HL<-copula.vs(kidney.formula,penalty="HL",
tun_range=seq(0.001,0.2,0.001),beta=beta0L,phi=phi0L,
theta=theta0L, data="kidney",maxiter=2000)
kid_res.HL
#=== kidney infection data(8 covariates) ==#
library(survival)
data(kidney)
kidney$age<-(kidney$age-mean(kidney$age))/sd(kidney$age)
kidney$GN<-as.numeric(kidney$disease=="GN")
kidney$AN<-as.numeric(kidney$disease=="AN")
kidney$PKD<-as.numeric(kidney$disease=="PKD")
```

```
kidney$sex<-kidney$sex
kidney$id<-kidney$id
S.GN<-kidney$sex*kidney$GN
S.AN<-kidney$sex*kidney$AN
S.PKD<-kidney$sex*kidney$PKD
data_kid <- kidney
attach(data_kid)
kidney.formula1<- Surv(time,status)~age+sex+GN+AN+PKD
      +S.GN+S.AN+S.PKD+id
beta00<-c(0,0,0,0,0,0,0,0,0)
phi0=1
theta0=0.01
# NO_PENALTY
kid_res1<-copula.vs(kidney.formula1,penalty="LASSO",tun_range=c(0),beta=b
      eta00, phi=phi0, theta=theta0,data="data_kid",maxiter=2000)
kid_res1
beta0<-kid_res1$Est_beta[,1] #No_penalty
phi0<-kid_res1$Est_phi[1] #No_penalty
theta0<-kid_res1$Est_theta[1] #No_penalty
theta0<- ifelse(theta0 <= 0.00001 , theta0<-0.001,
          theta0 <- theta0)
# LASSO
kid_res.LASSO1<-copula.vs(kidney.formula1,penalty="LASSO",
tun_range=seq(0,0.1, 0.001),beta=beta0,phi=phi0,
theta=theta0, data="data_kid",maxiter=2000)
kid_res.LASSO1
beta0L<-kid_res.LASSO1$Est_beta[,1] #LASSO
phi0L<-kid_res.LASSO1$Est_phi[1] #LASSO
theta0L<-kid_res.LASSO1$Est_theta[1] #LASSO
theta0L<- ifelse(theta0L <= 0.00001 , theta0L<-0.001, theta0L <- theta0L)
# ALASSO
kid_res.ALASSO1<-copula.vs(kidney.formula1,penalty="ALASSO",
```

```
tun_range=seq(0,0.1,0.001),beta=beta0L,phi=phi0L,
theta=theta0L, weight0=abs(1/beta0), data="data_kid",
maxiter=2000)
kid_res.ALASSO1
# SCAD
kid_res.SCAD1<-copula.vs(kidney.formula1,penalty="SCAD",
tun_range=seq(0,0.3,0.001),beta=beta0L,phi=phi0L,
theta=theta0L, data="data_kid",maxiter=2000)
kid_res.SCAD1
# HL
kid_res.HL1<-copula.vs(kidney.formula1,penalty="HL",
tun_range=seq(0.001,0.2,0.001),beta=beta0L,phi=phi0L,
theta=theta0L, data="data_kid", maxiter=2000)
kid_res.HL1
```

Appendix D. Further Simulation Results

Table D1. (q, n_i) = (50, 2): Simulation results on one-stage and two-stage estimation methods over 500 replications under Clayton copula models with Gompertz marginal hazard; 20% censoring rate; $\beta = 1$; $\theta = 2$ (Kendal's tau: $\tau = 0.5$); PE, piecewise exponential

ϕ	Est	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
One-stage			Weibull				PE				Proposed		
0.2	$\hat{\beta}$	0.986	0.118 0.128	0.016	0.924	1.055	0.131 0.137	0.022	0.932	1.053	0.130 0.138	0.022	0.922
	$\hat{\theta}$	2.042	0.632 0.657	0.433	0.944	2.232	0.703 0.769	0.643	0.946	2.152	0.669 0.730	0.554	0.930
1	$\hat{\beta}$	0.896	0.113 0.129	0.028	0.778	1.052	0.130 0.137	0.021	0.932	1.050	0.129 0.138	0.021	0.926
	$\hat{\theta}$	1.947	0.600 0.586	0.345	0.932	2.236	0.698 0.780	0.663	0.936	2.171	0.668 0.723	0.550	0.940
3	$\hat{\beta}$	0.828	0.111 0.131	0.047	0.594	1.050	0.131 0.138	0.022	0.936	1.047	0.130 0.136	0.021	0.930
	$\hat{\theta}$	1.937	0.602 0.567	0.325	0.948	2.254	0.712 0.798	0.701	0.934	2.180	0.679 0.727	0.561	0.942
Two-stage			Weibull				PE				Cox		
0.2	$\hat{\beta}$	1.002	0.141 0.151	0.023	0.934	1.056	0.133 0.168	0.031	0.858	1.033	0.152 0.158	0.026	0.924
	$\hat{\theta}$	1.847	0.600 0.521	0.295	0.900	1.764	0.450 0.532	0.338	0.752	1.790	0.591 0.528	0.322	0.854
1	$\hat{\beta}$	0.934	0.135 0.142	0.024	0.894	1.053	0.132 0.165	0.030	0.864	1.034	0.151 0.159	0.026	0.932
	$\hat{\theta}$	1.765	0.561 0.509	0.314	0.868	1.727	0.447 0.531	0.356	0.746	1.757	0.593 0.524	0.333	0.858
3	$\hat{\beta}$	0.881	0.131 0.137	0.033	0.830	1.052	0.133 0.167	0.030	0.872	1.033	0.153 0.162	0.027	0.928
	$\hat{\theta}$	1.740	0.572 0.506	0.323	0.876	1.712	0.453 0.527	0.360	0.752	1.752	0.603 0.533	0.345	0.848

Table D2. (q, n_i) = (200, 2): Simulation results on one-stage and two-stage estimation methods over 500 replications under Clayton copula models with Gompertz marginal hazard; 20% censoring rate; $\beta = 1$; $\theta = 2$ (Kendal's tau: $\tau = 0.5$); PE, piecewise exponential

ϕ	Est	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
One-stage			Weibull				PE				Proposed		
0.2	$\hat{\beta}$	0.960	0.057 0.060	0.005	0.844	1.005	0.061 0.063	0.004	0.954	1.011	0.061 0.062	0.004	0.942
	$\hat{\theta}$	1.982	0.307 0.308	0.095	0.940	2.046	0.320 0.337	0.016	0.934	2.035	0.316 0.319	0.103	0.944
1	$\hat{\beta}$	0.911	0.069 0.069	0.013	0.738	0.997	0.061 0.062	0.004	0.944	1.011	0.061 0.061	0.004	0.950
	$\hat{\theta}$	1.831	0.285 0.292	0.114	0.878	2.043	0.317 0.322	0.105	0.942	2.041	0.315 0.316	0.101	0.946
3	$\hat{\beta}$	0.859	0.067 0.067	0.024	0.438	0.996	0.061 0.063	0.004	0.950	1.011	0.061 0.062	0.004	0.950
	$\hat{\theta}$	1.798	0.287 0.295	0.127	0.870	2.041	0.321 0.320	0.104	0.950	2.039	0.319 0.319	0.103	0.946
Two-stage			Weibull				PE				Cox		
0.2	$\hat{\beta}$	0.976	0.072 0.072	0.006	0.924	1.008	0.075 0.075	0.006	0.948	1.007	0.076 0.075	0.006	0.952
	$\hat{\theta}$	1.937	0.305 0.299	0.093	0.928	1.926	0.302 0.299	0.095	0.914	1.920	0.319 0.295	0.093	0.936
1	$\hat{\beta}$	0.911	0.069 0.069	0.013	0.738	1.004	0.074 0.075	0.006	0.958	1.007	0.076 0.075	0.006	0.956
	$\hat{\theta}$	1.831	0.285 0.292	0.114	0.878	1.907	0.296 0.294	0.095	0.910	1.904	0.316 0.286	0.091	0.930
3	$\hat{\beta}$	0.859	0.067 0.067	0.024	0.438	1.003	0.075 0.076	0.006	0.954	1.007	0.077 0.076	0.006	0.952
	$\hat{\theta}$	1.798	0.287 0.295	0.127	0.870	1.898	0.302 0.297	0.098	0.918	1.892	0.319 0.289	0.095	0.932

Table D3. (q, n_i) = (50, 4): Simulation results on one-stage and two-stage estimation methods over 500 replications under Clayton copula models with Gompertz marginal hazard; 20% censoring rate; $\beta = 1$; $\theta = 2$ (Kendal's tau: $\tau = 0.5$); PE, piecewise exponential

ϕ	Est	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
One-stage			Weibull				PE				Proposed		
0.2	$\hat{\beta}$	0.967	0.082 0.083	0.008	0.914	1.027	0.101 0.101	0.011	0.952	1.032	0.091 0.091	0.009	0.932
	$\hat{\theta}$	1.986	0.423 0.406	0.165	0.936	2.082	0.493 0.504	0.261	0.946	2.031	0.433 0.428	0.184	0.952
1	$\hat{\beta}$	0.865	0.077 0.083	0.025	0.562	1.016	0.092 0.092	0.009	0.952	1.031	0.092 0.091	0.009	0.954
	$\hat{\theta}$	1.920	0.409 0.379	0.150	0.928	2.066	0.448 0.450	0.207	0.952	2.039	0.436 0.442	0.197	0.942
3	$\hat{\beta}$	0.785	0.073 0.081	0.053	0.218	1.011	0.092 0.092	0.009	0.952	1.028	0.092 0.090	0.009	0.958
	$\hat{\theta}$	1.944	0.412 0.374	0.143	0.942	2.070	0.450 0.450	0.207	0.950	2.038	0.439 0.437	0.192	0.946
Two-stage			Weibull				PE				Cox		
0.2	$\hat{\beta}$	0.992	0.109 0.111	0.012	0.948	1.029	0.103 0.131	0.018	0.850	1.023	0.116 0.123	0.016	0.938
	$\hat{\theta}$	1.897	0.406 0.392	0.164	0.920	1.857	0.350 0.437	0.211	0.794	1.807	0.417 0.388	0.188	0.870
1	$\hat{\beta}$	0.926	0.104 0.103	0.016	0.876	1.028	0.098 0.122	0.016	0.890	1.022	0.116 0.122	0.015	0.930
	$\hat{\theta}$	1.793	0.372 0.370	0.179	0.864	1.785	0.317 0.397	0.203	0.758	1.779	0.414 0.385	0.197	0.850
3	$\hat{\beta}$	0.872	0.100 0.099	0.026	0.746	1.026	0.099 0.124	0.016	0.876	1.022	0.117 0.124	0.016	0.924
	$\hat{\theta}$	1.762	0.369 0.360	0.186	0.860	1.780	0.320 0.392	0.202	0.768	1.771	0.416 0.385	0.200	0.844

Table D4. (q, n_i) = (200, 4): Simulation results on one-stage and two-stage estimation methods over 500 replications under Clayton copula models with Gompertz marginal hazard; 20% censoring rate; $\beta = 1$; $\theta = 2$ Kendal's tau: $\tau = 0.5$); PE, piecewise exponential

ϕ	Est	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP	Mean	SE SD	MSE	CP
One-stage			Weibull				PE				Proposed		
0.2	$\hat{\beta}$	0.947	0.040 0.041	0.004	0.702	0.995	0.044 0.042	0.002	0.966	1.007	0.044 0.044	0.002	0.954
	$\hat{\theta}$	1.976	0.211 0.196	0.039	0.950	2.022	0.219 0.213	0.046	0.956	2.011	0.214 0.214	0.046	0.940
1	$\hat{\beta}$	0.848	0.038 0.041	0.025	0.050	0.985	0.044 0.044	0.002	0.930	1.008	0.045 0.045	0.002	0.950
	$\hat{\theta}$	1.914	0.204 0.176	0.038	0.944	2.023	0.219 0.211	0.045	0.954	2.010	0.216 0.214	0.046	0.944
3	$\hat{\beta}$	0.769	0.036 0.040	0.055	0	0.980	0.044 0.044	0.002	0.916	1.007	0.045 0.045	0.002	0.948
	$\hat{\theta}$	1.943	0.206 0.171	0.032	0.970	2.024	0.220 0.211	0.045	0.958	2.010	0.217 0.215	0.046	0.942
Two-stage			Weibull				PE				Cox		
0.2	$\hat{\beta}$	0.973	0.056 0.058	0.004	0.900	1.004	0.059 0.063	0.004	0.930	1.006	0.060 0.063	0.004	0.942
	$\hat{\theta}$	1.934	0.211 0.202	0.045	0.926	1.938	0.212 0.210	0.048	0.912	1.923	0.223 0.209	0.050	0.914
1	$\hat{\beta}$	0.909	0.053 0.055	0.011	0.584	1.000	0.059 0.063	0.004	0.912	1.006	0.061 0.063	0.004	0.938
	$\hat{\theta}$	1.822	0.193 0.183	0.065	0.848	1.918	0.209 0.208	0.050	0.900	1.913	0.222 0.208	0.051	0.904
3	$\hat{\beta}$	0.856	0.051 0.053	0.024	0.194	0.998	0.208 0.063	0.004	0.934	1.006	0.061 0.064	0.004	0.932
	$\hat{\theta}$	1.793	0.193 0.179	0.075	0.808	1.906	0.208 0.206	0.051	0.896	1.904	0.222 0.204	0.051	0.904

Table D5. Simulation results for coefficients of β_1, $\beta_4$and β_7 among non-zero coefficients of β under copula survival models with Censoring rate 40%

		$\hat{\beta}_1$					$\hat{\beta}_4$					$\hat{\beta}_7$				
(q, n_i)	Method	Mean	SE	SD	MSE	CP	Mean	SE	SD	MSE	CP	Mean	SE	SD	MSE	CP
True value		$\beta_1 = 0.8$					$\beta_4 = 1$					$\beta_7 = 0.6$				
(100,2)	LASSO	0.712	0.080	0.092	0.016	0.730	0.900	0.088	0.07	0.020	0.740	0.529	0.075	0.080	0.011	0.810
	ALASSO	0.781	0.084	0.104	0.011	0.890	0.982	0.092	0.110	0.012	0.890	0.577	0.093	0.104	0.010	0.895
	SCAD	0.811	0.087	0.101	0.010	0.915	1.013	0.102	0.102	0.010	0.945	0.605	0.075	0.085	0.007	0.915
	HL	0.791	0.080	0.091	0.008	0.930	0.994	0.095	0.093	0.012	0.895	0.589	0.080	0.080	0.008	0.930
(100,4)	LASSO	0.736	0.054	0.059	0.008	0.745	0.926	0.060	0.071	0.011	0.715	0.539	0.051	0.059	0.007	0.720
	ALASSO	0.792	0.057	0.060	0.004	0.930	0.998	0.063	0.067	0.004	0.950	0.589	0.052	0.056	0.003	0.930
	SCAD	0.806	0.058	0.060	0.003	0.955	1.011	0.064	0.072	0.005	0.920	0.594	0.053	0.057	0.003	0.935
	HL	0.797	0.057	0.064	0.004	0.925	1.002	0.063	0.069	0.005	0.920	0.593	0.053	0.060	0.004	0.910
(300,2)	LASSO	0.743	0.048	0.048	0.005	0.745	0.936	0.052	0.059	0.008	0.695	0.549	0.045	0.052	0.005	0.765
	ALASSO	0.792	0.049	0.055	0.003	0.930	0.996	0.054	0.056	0.003	0.945	0.592	0.044	0.051	0.003	0.910
	SCAD	0.806	0.050	0.050	0.003	0.950	1.010	0.054	0.058	0.003	0.960	0.605	0.046	0.051	0.003	0.940
	HL	0.798	0.049	0.051	0.003	0.955	0.999	0.054	0.055	0.003	0.945	0.593	0.046	0.050	0.003	0.945

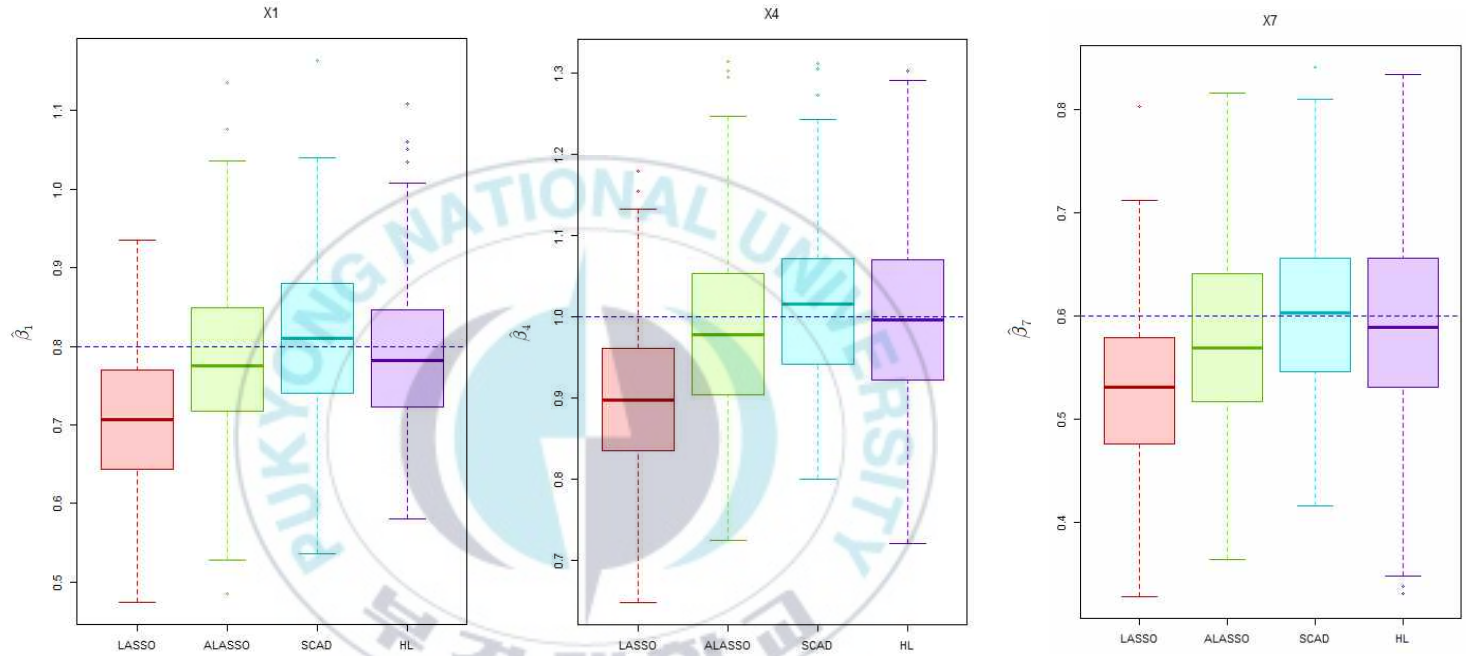


Figure D1. $(q, n_i) = (100, 2)$: Simulation result of copula variable selection using 200 replications; 40% censoring rate; dotted line, true values of β_1, $\beta_4$and β_7, respectively

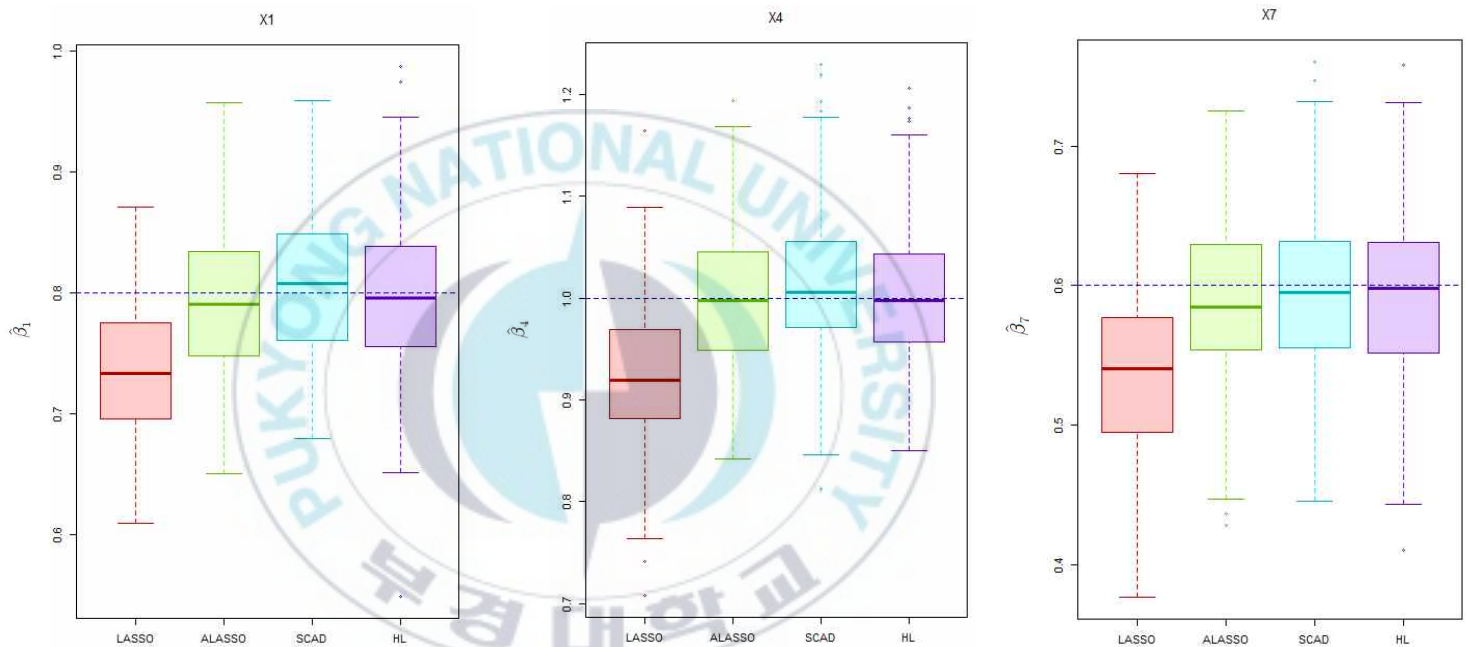


Figure D2. $(q, n_i) = (100, 4)$: Simulation result of copula variable selection using 200 replications; 40% censoring rate; dotted line, true values of β_0, β_1, $\beta_4$and β_7, respectively

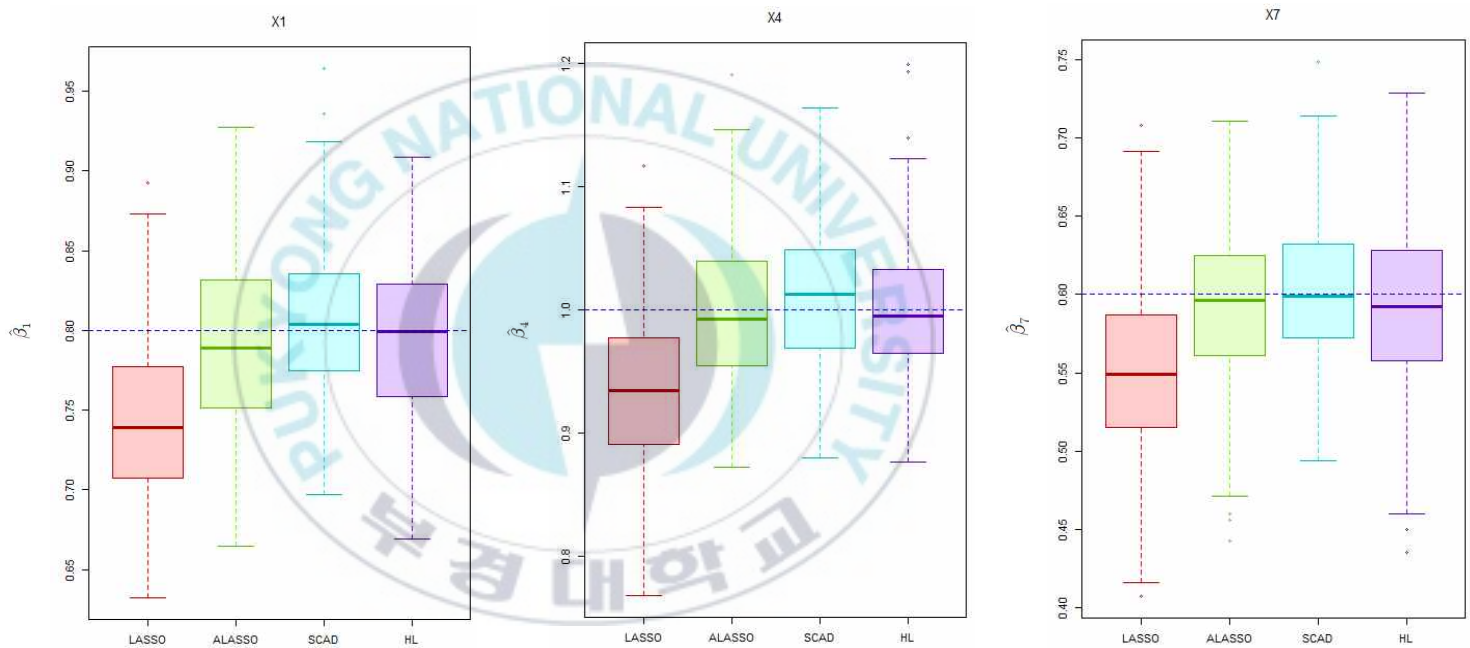


Figure D3. $(q, n_i) = (300, 2)$: Simulation result of copula variable selection using 200 replications; 40% censoring rate; dotted line, true values of β_0, β_1, $\beta_4$and β_7, respectively